

*Cambridge Handbook of
Experimental Political Science*

Edited by
James N. Druckman
Donald P. Green
James H. Kuklinski
Arthur Lupia

Table of Contents

Contributors

List of Tables

List of Figures

1. Experimentation in Political Science

James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

I. Designing Experiments

2. Experiments: An Introduction to Core Concepts

James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

3. Internal and External Validity

Rose McDermott

4. Students as Experimental Participants: A Defense of the “Narrow Data Base”

James N. Druckman and Cindy D. Kam

5. Economics vs. Psychology Experiments: Stylization, Incentives, and Deception

Eric S. Dickson

II. The Development of Experiments in Political Science

6. Laboratory Experiments in Political Science

Shanto Iyengar

7. Experiments and Game Theory’s Value to Political Science

John H. Aldrich and Arthur Lupia

8. The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation

Paul M. Sniderman

9. Field Experiments in Political Science

Alan S. Gerber

III. Decision Making

10. Attitude Change Experiments in Political Science

Allyson L. Holbrook

11. Conscious and Unconscious Information Processing with Implications for Experimental Political Science

Milton Lodge, Charles Taber, and Brad Verhulst

12. Political Knowledge
Cheryl Boudreau and Arthur Lupia

IV. Vote Choice, Candidate Evaluations, and Turnout

13. Candidate Impressions and Evaluations
Kathleen M. McGraw

14. Media and Politics
Thomas E. Nelson, Sarah M. Bryner, and Dustin M. Carnahan

15. Candidate Advertisements
Shana Kushner Gadarian and Richard R. Lau

16. Voter Mobilization
Melissa R. Michelson and David W. Nickerson

V. Interpersonal Relations

17. Trust and Social Exchange
Rick K. Wilson and Catherine C. Eckel

18. An Experimental Approach to Citizen Deliberation
Christopher F. Karpowitz and Tali Mendelberg

19. Social Networks and Political Context
David W. Nickerson

VI. Identity, Ethnicity, and Politics

20. Candidate Gender and Experimental Political Science
Kathleen Dolan and Kira Sanbonmatsu

21. Racial Identity and Experimental Methodology
Darren Davis

22. The Determinants and Political Consequences of Prejudice
Vincent L. Hutchings and Spencer Piston

23. Politics from the Perspective of Minority Populations
Dennis Chong and Jane Junn

VII. Institutions and Behavior

24. Experimental Contributions to Collective-Action Theory
Eric Coleman and Elinor Ostrom

25. Legislative Voting and Cycling
Gary Miller

26. Electoral Systems and Strategic Voting (Laboratory Election Experiments)
Rebecca B. Morton and Kenneth C. Williams

27. Experimental Research on Democracy and Development
Ana L. De La O and Leonard Wantchekon

VIII. Elite Bargaining

28. Coalition Experiments
Daniel Diermeier

29. Negotiation and Mediation
Daniel Druckman

30. The Experiment and Foreign Policy Decision Making
Margaret G. Hermann and Binnur Ozkececi-Taner

IX. Advanced Experimental Methods

31. Treatment Effects
Brian J. Gaines and James H. Kuklinski

32. Making Effects Manifest in Randomized Experiments
Jake Bowers

33. Design and Analysis of Experiments in Multilevel Populations
Betsy Sinclair

34. Analyzing the Downstream Effects of Randomized Experiments
Rachel Milstein Sondheimer

35. Mediation Analysis Is Harder than It Looks
John G. Bullock and Shang E. Ha

Afterword

36. Campbell's Ghost
Donald R. Kinder

Index

Contributors

John H. Aldrich, Duke University
Cheryl Boudreau, University of California, Davis
Jake Bowers, University of Illinois at Urbana-Champaign
Sarah M. Bryner, The Ohio State University
John G. Bullock, Yale University
Dustin M. Carnahan, The Ohio State University
Dennis Chong, Northwestern University
Eric Coleman, Florida State University
Darren Davis, University of Notre Dame
Ana L. De La O, Yale University
Eric S. Dickson, New York University
Daniel Diermeier, Northwestern University
Kathleen Dolan, University of Wisconsin, Milwaukee
Daniel Druckman, George Mason University
James N. Druckman, Northwestern University
Catherine C. Eckel, University of Texas at Dallas
Shana Kushner Gadarian, University of California, Berkeley
Brian J. Gaines, University of Illinois at Urbana-Champaign
Alan S. Gerber, Yale University
Donald P. Green, Yale University
Shang E. Ha, Brooklyn College, City University of New York
Margaret G. Hermann, Syracuse University
Allyson L. Holbrook, University of Illinois at Chicago
Vincent L. Hutchings, University of Michigan
Shanto Iyengar, Stanford University
Jane Junn, University of Southern California
Cindy D. Kam, Vanderbilt University
Christopher F. Karpowitz, Brigham Young University
Donald R. Kinder, University of Michigan
James H. Kuklinski, University of Illinois at Urbana-Champaign
Richard R. Lau, Rutgers University
Milton Lodge, Stony Brook University
Arthur Lupia, University of Michigan
Rose McDermott, Brown University
Kathleen M. McGraw, The Ohio State University
Tali Mendelberg, Princeton University
Melissa R. Michelson, California State University, East Bay
Gary Miller, Washington University in St. Louis
Rebecca B. Morton, New York University
Thomas E. Nelson, The Ohio State University
David W. Nickerson, University of Notre Dame
Elinor Ostrom, Indiana University
Binnur Ozkececi-Taner, Hamline University
Spencer Piston, University of Michigan

Kira Sanbonmatsu, Rutgers University
Betsy Sinclair, University of Chicago
Paul M. Sniderman, Stanford University
Rachel Milstein Sondheimer, United States Military Academy
Charles Taber, Stony Brook University
Brad Verhulst, Stony Brook University
Leonard Wantchekon, New York University
Kenneth C. Williams, Michigan State University
Rick K. Wilson, Rice University

List of Tables

- 4-1. Sampling distribution of b_T , single treatment effect
- 9-1. Approximate Cost of Adding One Vote to Candidate Vote Margin
- 9-2. Voter Mobilization Experiments Prior to 1998 New Haven Experiment
- 11-1. A Schematic Figure of the Racial IAT using pleasant and unpleasant words and Euro-American and Afro-American Stereotype words
- 25-1. Testing the Uncovered Set with Previous Majority Rule Experiments - Bianco et al. (2006)
- 26-1. Forsythe et al (1993) Payoff Schedule
- 26-2. Dasgupta et al. (2008) Payoff Schedule
- 32-1. Balance tests for covariates adjusted for blocking in the blocked thirty two city study
- 32-2. Balance tests for covariates in the blocked 8 city study
- 32-3. Balance tests for covariates adjusted for covariates by post-stratification in the blocked thirty two city study
- 33-1. ITT Effects
- 34-1. Classification of Target Population in Downstream Analysis of Educational Intervention

List of Figures

- 1-1. Experimental Articles in the *APSR*
- 4-1. Sampling distribution of b_T , single treatment effect
- 4-2. Sampling distribution of b_T , heterogeneous treatment effects
- 4-3. Sampling distributions of b_T and b_{TZ} , heterogeneous treatment effects
- 6-1. Race of Suspect Manipulation
- 6-2. The Facial Similarity Manipulation
- 9-1. Graphical Representation of Treatment Effects with Noncompliance
- 10-1. Pretest-Posttest Control Group Design (Campbell and Stanley 1963, 13)
- 10-2. Pretest-Posttest Multiple Experimental Condition Design
- 10-3. Posttest-Only Control Group Design
- 10-4. Posttest-Only Multiple Experimental Group Design
- 11-1. Spreading Activation in a Sequential Priming Paradigm for Short and Long SOA
- 15-1. Methodological Consequences of Differences Between Observational and Experimental Studies of Candidate Advertisements
- 21-1. Example of Experimental Design for Racial Identity
- 24-1. A Prisoner's Dilemma Game
- 24-2. Contributions in a Public-Goods Game
- 24-3. A Common-Pool Resource Game
- 25-1. Outcomes of Majority Rule Experiments without a Core - Fiorina and Plott (1978)
- 25-2. Majority Rule with Issue-by-Issue Voting - McKelvey and Ordeshook (1984)
- 25-3. The Effect of Backward and Forward Agendas - Wilson (2008b)
- 25-4. The Effect of Monopoly Agenda Setting - Wilson (2008b)
- 25-5a. A Sample Majority-Rule Trajectory for Configuration 1 - Bianco et al. (2008)
- 25-5b. A Sample Majority-Rule Trajectory for Configuration 2 - Bianco et al. (2008)
- 25-6a. The Uncovered Set and Outcomes for Configuration 1 - Bianco et al. (2008)
- 25-6b. The Uncovered Set and Outcomes for Configuration 2 - Bianco et al. (2008)
- 25-7. Senatorial Ideal Points and Proposed Amendments for the Civil Rights Act of 1964 - Jeong et al. (2009)
- 26-1. Median Voter Theorem
- 28-1. Potential Coalitions and their Respective Payoffs
- 32-1. The efficiency of paired and unpaired designs in simulated turnout data
- 32-2. Graphical assessment of balance on distributions of baseline turnout for the thirty two city experiment data
- 32-3. Post-stratification adjusted confidence intervals for the difference in turnout between treated and control cities in the thirty two-city turnout experiment
- 32-4. Covariance-adjustment in a simple random experiment
- 32-5. Covariance-adjustment in a blocked random experiment
- 32-6. Covariance-adjusted confidence intervals for the difference in turnout between treated and control cities in the thirty-two-city turnout experiment data
- 33-1. Multilevel Experiment Design
- 36-1. Number of Articles Featuring Experiments Published in *The American Political Science Review*, 1906-2009

Acknowledgements

This volume has its origins in the *American Political Science Review*'s special 2006 centennial issue celebrating the evolution of the study of politics. For that issue, we proposed a paper that traced the history of experiments within political science. The journal's editor, Lee Sigelman, responded to our proposal for the issue with a mix of skepticism – for example, asking about the prominence of experiments in the discipline – and encouragement. We moved forward and eventually published a paper in the special issue, and there is no doubt it was much better than it would have been absent Lee's constant constructive guidance. Indeed, Lee, who himself conducted some remarkably innovative experiments, pushed us to think about what makes political science experiments unique relative to the other psychological and social sciences. It was this type of prodding that led us to conceive of this *Handbook*. Sadly, Lee did not live to see the completion of the *Handbook*, but we hope it approaches the high standards he always set. We know we are not alone in saying that Lee is greatly missed.

Our first task in developing the *Handbook* was to generate a list of topics and possible authors; we were overwhelmed by the positive responses to our invitations to contribute. While we leave it to the reader to assess the value of the book, we can say the experience of assembling this volume could not have been more enjoyable and instructive, thanks to authors.

Nearly all of the authors attended a conference held at Northwestern University (in Evanston, IL, USA) on May 28th-29th, 2009. We were extremely fortunate to have an exceptionally able group of discussants take the lead in presenting and commenting on the chapters; we deeply appreciate the time and insights they provided. The discussants included: Kevin Arceneaux, Ted Brader, Ray Duch, Kevin Esterling, Diana Mutz, Mike Neblo, Eric Oliver, Randy Stevenson, Nick Valentino, and Lynn Vavreck. Don Kinder played a special role at the conference offering his overall assessment at the end of the proceedings. A version of these thoughts appears as the volume's Afterword.

We also owe thanks to the more than thirty graduate students who attended the conference, met with faculty, and offered their perspectives. These students (many of whom became Professors before the publication of the volume) included Lene Aarøe, Emily Alvarez, Christy Aroopala, Bernd Beber, Toby Bolsen, Kim Dionne, Katie Donovan, Ryan Enos, Brian Falb, Mark Fredrickson, Fernando Garcia, Ben Gaskins, Seth Goldman, Daniel Hidalgo, Samara Klar, Yanna Krupnikov, Thomas Leeper, Adam Levine, Peter Loewen, Kristin Michelitch, Daniel Myers, Jennifer Ogg Anderson, Spencer Piston, Josh Robison, Jon Rogowski, Mark Schneider, Geoff Sheagley, Alex Theodoridis, Catarina Thomson, Dustin Tingley, Brad Verhulst, and Abby Wood.

We thank a number of others who attended the conference and offered important comments, including, but not limited to David Austen-Smith, Traci Burch, Fay Cook, Jeremy Freese, Jerry Goldman, Peter Miller, Eugenia Mitchelstein, Ben Page, Jenn Richeson, Anne Sartori, Victor Shih, and Salvador Vazquez del Mercado.

The conference would not have been possible without the exceptional contributions of a number of individuals. Of particular note are the many staff members of Northwestern's Institute for

Policy Research. We thank the Institute's director, Fay Cook, for supporting the conference, and we are indebted to Patricia Reese for overseeing countless logistics. We also thank Eric Betzold, Arlene Dattels, Sarah Levy, Michael Weis, and Bev Zack. A number of Northwestern's political science Ph.D. students also donated their time to ensure a successful event – including Emily Alvarez, Toby Bolsen, Brian Falb, Samara Klar, Thomas Leeper, and Josh Robison. We also thank Nicole, Jake, and Sam Druckman for their patience and help in ensuring everything at the conference was in place.

Of course the conference and the production of the volume could not have been possible without generous financial support and we gratefully acknowledge National Science Foundation (SES-0851285), Northwestern University's Weinberg College of Arts and Sciences, and the Institute for Policy Research.

Following the conference, authors engaged in substantial revisions and, along the way, a number of others provided instructive comments – including Cengiz Erisen, Jeff Guse, David Llanos, and the anonymous press reviewers. We also thank the participants in Druckman's graduate experimental class who read a draft of the volume and commented on each chapter; these impressive students included: Emily Alvarez, Toby Bolsen, Brian Falb, Samara Klar, Thomas Leeper, Rachel Moskowitz, Taryn Nelson, Christoph Nguyen, Josh Robison, and Xin Sun. We have no doubt that countless others offered advice (of which we, as the editors, are not directly aware), and we thank them for their contributions. A special acknowledgement is due to Samara Klar and Thomas Leeper who probably have read the chapters more than anyone else, and without fail, have offered helpful advice and skillful coordination. Finally, it was a pleasure working with Eric Crahan and Jason Przybylski at Cambridge University Press.

We view this *Handbook* as a testament to the work of many scholars (a number of whom are authors in this volume) who set the stage for experimental approaches in political science. While we cannot be sure what many of them will think of the volume, we do hope it successfully addresses a question raised by an editor's (Druckman's) son who was 7 when he asked “why is political ‘science’ a ‘science’ since it doesn't do things that science does, like run experiments?”

--James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

Introduction

1. Experimentation in Political Science

James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia¹

In his 1909 American Political Science Association presidential address, A. Lawrence Lowell advised the fledgling discipline against following the model of the natural sciences: “We are limited by the impossibility of experiment. Politics is an observational, not an experimental science...” (Lowell 1910, 7). The lopsided ratio of observational to experimental studies in political science, over the one hundred years since Lowell’s statement, arguably affirms his assessment. The next hundred years are likely to be different. The number and influence of experimental studies are growing rapidly as political scientists discover ways of using experimental techniques to illuminate political phenomena.

The growing interest in experimentation reflects the increasing value that the discipline places on causal inference and empirically-guided theoretical refinement. Experiments facilitate causal inference through the transparency and content of their procedures, most notably the random assignment of observations (a.k.a., subjects or experimental participants) to treatment and control groups. Experiments also guide theoretical development by providing a means for pinpointing the effects of institutional rules, preference configurations, and other contextual factors that might be difficult to assess using other forms of inference. Most of all, experiments guide theory by providing stubborn facts – that is to say, reliable information about cause and effect that inspires and constrains theory.

Experiments bring new opportunities for inference along with new methodological challenges. The goal of the *Cambridge Handbook of Experimental Political Science* is to help scholars more effectively pursue experimental opportunities while better understanding the

challenges. To accomplish this goal, the *Handbook* offers a review of basic definitions and concepts, compares experiments with other forms of inference in political science, reviews the contributions of experimental research, and presents important methodological issues. It is our hope that discussing these topics in a single volume will help facilitate the growth and development of experimentation in political science.

1. The Evolution and Influence of Experiments in Political Science

Social scientists answer questions about social phenomena by constructing theories, deriving hypotheses, and evaluating these hypotheses by empirical or conceptual means. One way to evaluate hypotheses is to intervene deliberately in the social process under investigation. An important class of interventions is known as experiments. An experiment is a deliberate test of a causal proposition, typically with random assignment to conditions.ⁱⁱ Investigators design experiments to evaluate the causal impacts of potentially informative explanatory variables.

While scientists have conducted experiments for hundreds of years, modern experimentation made its debut in the 1920s and 1930s. It was then that, for the first time, social scientists began to use random assignment in order to allocate subjects to control and treatment groups.ⁱⁱⁱ One can find examples of experiments in political science as early as the 1940s and 1950s. The first experimental paper in the *American Political Science Review (APSR)* appeared in 1956 (Eldersveld 1956).^{iv} In that study, the author randomly assigned potential voters to a control group that received no messages, or to treatment groups that received messages encouraging them to vote via personal contact (which included phone calls or personal visits) or via a mailing. The study showed that more voters in the personal contact treatment groups turned out to vote than those in either the control group or the mailing group; personal contact caused a relative increase in turnout. A short time after Eldersveld's study, an active research program

using experiments to study international conflict resolution began (e.g., Mahoney and Druckman 1975; Guetzkow and Valadez 1981), and, later, a periodic but now extinct journal, *The Experimental Study of Politics*, began publication (also see Brody and Brownstein 1975).

These examples are best seen as exceptions, however. For much of the discipline's history, experiments remained on the periphery. In his widely-cited methodological paper from 1971, Lijphart (1971) states, "The experimental method is the most nearly ideal method for scientific explanation, but unfortunately it can only rarely be used in political science because of practical and ethical impediments" (684). In their oft-used methods text, King, Keohane, and Verba (1994) provide virtually no discussion of experimentation, stating only that experiments are helpful in so far as they "provide a useful model for understanding certain aspects of non-experimental design" (125).

A major change in the status of experiments in political science occurred during the last decades of the twentieth century. Evidence of the change is visible in Figure 1-1. This figure comes from a content analysis of the discipline's widely-regarded flagship journal, the *APSR*. The figure shows a sharp increase, in recent years, in the number of articles using a random-assignment experiment. In fact, more than half of the 71 experimental articles that appeared in the *APSR* during its first 103 years were published after 1992. Other signs of the rise of experiments include the many graduate programs now offering courses on experimentation, National Science Foundation support for experimental infrastructure, and the proliferation of survey experiments in both private and publicly supported studies.^v

[Figure 1-1 here]

Experimental approaches have not been confined to single subfields or approaches. Instead, political scientists have employed experiments across fields, and have drawn on and

developed a notable range of experimental methods. These sources of diversity make a unifying *Handbook* particularly appealing for the purpose of facilitating coordination and communication across varied projects.

2. Diversity of Applications

Political scientists have implemented experiments for various purposes to address a variety of issues. Roth (1995) identifies three non-exclusive roles that experiments can play, and a cursory review makes clear that political scientists employ them in all three ways. First, Roth describes “searching for facts,” where the goal is to “isolate the cause of some observed regularity, by varying details of the way the experiments were conducted. Such experiments are part of the dialogue that experimenters carry on with one another” (22). These types of experiments often complement observational research (e.g., work not employing random assignment) by arbitrating between conflicting results derived from observational data. “Searching for facts” describes many experimental studies that attempt to estimate the magnitudes of causal parameters, such as the influence of racial attitudes on policy preferences (Gilens 1996) or the price-elasticity of demand for public and private goods (Green 1992).

A second role entails “speaking to theorists,” where the goal is “to test the predictions [or the assumptions] of well articulated formal theories [or other types of theories]... Such experiments are intended to feed back into the theoretical literature – i.e., they are part of a dialogue between experimenters and theorists” (Roth 1995, 22). The many political science experiments that assess the validity of claims made by formal modelers epitomize this type of correspondence (e.g., Ostrom, Walker, and Gardner 1992; Morton 1993; Fréchette, Kagel, and Lehrer 2003).^{vi} The third usage is “whispering in the ears of princes,” which facilitates “the dialogue between experimenters and policy-makers... [The] experimental environment is

designed to resemble closely, in certain respects, the naturally occurring environment that is the focus of interest for the policy purposes at hand” (Roth 1995, 22). Cover and Brumberg’s (1982) field experiment examining the effects of mail from members of the U.S. Congress on their constituents’ opinions exemplifies an experiment that whispers in the ears of legislative “princes.”

Although political scientists might share rationales for experimentation with other scientists, their attention to focal aspects of politically relevant contexts distinguishes their efforts. This distinction parallels the use of other modes of inference by political scientists. As Druckman and Lupia (2006) argue, “[c]ontext, not methodology, is what unites our discipline... Political science is united by the desire to understand, explain, and predict important aspects of contexts where individual and collective actions are intimately and continuously bound” (109). The environment in which an experiment takes place is thus of particular importance to political scientists.

And, while it might surprise some, political scientists have implemented experiments in a wide range of contexts. Examples can be found in every subfield. Applications to American politics include not only topics such as media effects (e.g., Iyengar and Kinder 1987), mobilization (e.g., Gerber and Green 2000), and voting (e.g., Lodge, McGraw, and Stroh 1989), but also studies of congressional and bureaucratic rules (e.g., Eavey and Miller 1984; Miller, Hammond, and Kile 1996). The field of international relations, in some ways, lays claim to one of the longest ongoing experimental traditions with its many studies of foreign policy decision-making (e.g., Geva and Mintz 1997) and international negotiations (e.g., D. Druckman 1994). Related work in comparative politics explores coalition bargaining (e.g., Riker 1967; Fréchet et al. 2003) and electoral systems (e.g., Morton and Williams 1999); and recently, scholars have

turned to experiments to study democratization and development (Wantchekon 2003), culture (Henrich et al. 2004) and identity (e.g., Sniderman, Hagendoorn, and Prior 2004; Habyarimana et al. 2007). Political theory studies include explorations into justice (Frohlich and Oppenheimer 1992) and deliberation (Simon and Sulkin 2001).

Political scientists employ experiments across subfields and for a range of purposes. At the same time, many scholars remain unaware of this range of activity, which limits the extent to which experimental political scientists have learned from one another. For example, scholars studying coalition formation and international negotiations experimentally can benefit from talking to one another, yet there is little sign of engagement between the respective contributors to these literatures. Similarly, there are few signs of collaboration amongst experimental scholars who study different kinds of decision-making (e.g., foreign policy decision-making and voting decisions). Of equal importance, scholars within specific fields who have not used experiments may be unaware of when and how experiments can be effective. A goal of this *Handbook* is to provide interested scholars with an efficient and effective way to learn about a broad range of experimental applications, how these applications complement and supplement non-experimental work, and the opportunities and challenges inherent in each type of application.

3. Diversity of Experimental Methods

The most apparent source of variation in political science experiments is where they are conducted. To date, most experiments have been implemented in one of three contexts: laboratories, surveys, and the field. These types of experiments differ in terms of where participants receive the stimuli (e.g., messages encouraging them to vote), with that exposure taking place, respectively, in a controlled setting, in the course of a phone, in-person, or web-

based survey, or in a naturally occurring setting such as the voter's home (e.g., in the course of everyday life, and often without the participants' knowledge).^{vii}

Each type of experiment presents methodological challenges. For example, scholars have long bemoaned the artificial settings of campus-based laboratory experiments and the widespread use of student-aged subjects. While experimentalists from other disciplines have examined implications of running experiments "on campus," this literature is not often cited by political scientists (e.g., Dipboye and Flanagan 1979; Kardes 1996; Kühberger 1998; Levitt and List 2007). Some political scientists claim that the problems of campus-based experiments can be overcome by conducting experiments on representative samples. This may be true. However, the conditions under which such changes produce more valid results have not been broadly examined (see, e.g., Greenberg 1987).^{viii}

Survey experiments, while not relying on campus-based "convenience samples," also raise questions about external validity. Many survey experiments, for example, expose subjects to phenomena they might have also encountered prior to participating in an experiment, which can complicate causal inference (Gaines, Kuklinski, and Quirk 2007).

Field experiments are seen as a way to overcome the artificiality of other types of experiments. In the field, however, there can be less control over what experimental stimuli subjects observe. It may also be more difficult to get people to participate due to an inability to recruit subjects or to subjects' unwillingness to participate as instructed once they are recruited.

Besides where they are conducted, another source of diversity in political science experiments is the extent to which they follow experimental norms in neighboring disciplines, such as psychology and economics. This diversity is notable because psychological and economic approaches to experimentation differ from each other. For example, where

psychological experiments often include some form of deception, economists consider it taboo. Psychologists rarely pay subjects for specific actions they undertake during an experiment. Economists, on the other hand, often require such payments (Smith 1976). Indeed, the inaugural issue of *Experimental Economics* stated that submissions that used deception or did not pay participants for their actions would not be accepted for publication.^{ix}

For psychologists and economists, differences in experimental traditions reflect differences in their dominant paradigms. Since most political scientists seek first and foremost to inform political science debates, norms about what constitutes a valid experiment in economics or psychology are not always applicable. So, for any kind of experiment, an important question to ask is: which experimental method is appropriate?

The current debate about this question focuses on more than the validity of the inferences that different experimental approaches can produce. Cost is also an issue. Survey and field experiments, for example, can be expensive. Some scholars question whether the added cost of such endeavors (compared to, say, campus-based laboratory experiments) is justifiable. Such debates are leading more scholars to evaluate the conditions under which particular types of experiments are cost-effective. With the evolution of these debates has come the question of whether the immediate costs of fielding an experiment are offset by what Green and Gerber (2002) call the “downstream benefits of experimentation.” Downstream benefits refer to subsequent outcomes that are set in motion by the original experimental intervention, such as the transmission of effects from one person to another or the formation of habits. In some cases, the downstream benefits of an experiment only become apparent decades afterward.

In sum, the rise of an experimental political science brings both new opportunities for discovery and new questions about the price of experimental knowledge. This *Handbook* is

organized to make the broad range of research opportunities more apparent and to help scholars manage the challenges with greater effectiveness and efficiency.

4. The Volume

In concluding his book on the ten most fascinating experiments in the history of science, Johnson (2008) explains that “I’ve barely finished the book and already I’m second-guessing myself” (158). We find ourselves in an analogous situation. There are many exciting kinds of experimental political science on which we can focus. While the *Handbook*’s content does not cover all possible topics, we made every effort to represent the broad range of activities that contemporary experimental political science entails. The content of the *Handbook* is as follows.

We begin with a series of chapters that provide an introduction to experimental methods and concepts. These chapters provide detailed discussion of what constitutes an experiment, as well as the key considerations underlying experimental designs (i.e., internal and external validity, student subjects, payment, and deception). While these chapters do not delve into the details of precise designs and statistical analyses (see, e.g., Keppel and Wickens 2004; Morton and Williams 2010), their purpose is to provide a sufficient base for reading the rest of the *Handbook*. We asked the authors of these chapters not only to review extant knowledge, but also to present arguments that help place the challenges of, and opportunities in, experimental political science in a broader perspective. For example, our chapters regard questions about external validity (i.e., the extent to which one can generalize experimental findings) as encompassing much more than whether a study employs a representative (or, at least, non-student) sample. This approach to the chapters yields important lessons about when student-based samples, and other common aspects of experimental designs, are and are not problematic.^x

The next set of chapters contains four essays written by prominent scholars who each played an important role in the development of experimental political science.^{xi} These essays provide important historical perspectives and relevant biographical information on the development of experimental research agendas. The authors describe the questions they hoped to resolve with experiments and why they think that their efforts succeeded and failed as they did. These essays also document the role experiments played in the evolution of much broader fields of inquiry.

The next six sections of the *Handbook* explore the role of political science experiments on a range of scholarly endeavors. The chapters in these sections clarify how experiments contribute to scientific and social knowledge of many important kinds of political phenomena. They describe cases in which experiments complement non-experimental work, as well as cases where experiments advance knowledge in ways that non-experimental work cannot. Each chapter describes how to think about experimentation on a particular topic and provides advice about how to overcome practical (and, when relevant, ethical) hurdles to design and implementation.

In developing this part of the *Handbook*, we attempted to include topics where experiments have already played a notable role. We devoted less space to “emerging” topics in experimental political science that have great potential to answer important questions but that are still in early stages of development. Examples of such work include genetic and neurobiological approaches (e.g., Fowler and Schreiber 2008), non-verbal communication (e.g., Bailenson et al. 2008), emotions (e.g., Druckman and McDermott 2008), cultural norms (e.g. Henrich et al. 2004), corruption (e.g., Ferraz and Finan 2008; Malesky and Samphantharak 2008), ethnic identity (e.g., Humphreys, Posner, and Weinstein 2002), and elite responsiveness (e.g., Esterling,

Lazer, and Neblo 2009; Richardson and John 2009). Note that the *Handbook* is written in such a way that any of the included chapters can be read and used without having read the chapters that precede them.

The final section of the book covers a number of advanced methodological debates. The chapters in this section address the challenges of making causal inferences in complex settings and over time. As with the earlier methodological chapters, these chapters do more than review basic issues, they also develop arguments on how to recognize and adapt to such challenges in future research.

The future of experimental political science offers many new opportunities for creative scholars. It also presents important challenges. We hope that this *Handbook* makes the challenges more manageable for you and the opportunities easier to seize.

5. Conclusion

In many scientific disciplines, experimental research is the focal form of scholarly activity. In these fields of study, disciplinary norms and great discoveries are indescribable without reference to experimental methods. For the most part, political science is not such a science. Its norms and great discoveries often come from scholars who integrate and blend multiple methods. In a growing number of topical areas, experiments are becoming an increasingly common and important element of a political scientist's methodological tool kit (see also Falk and Heckman 2009). Particularly in recent years, there has been a massive expansion in the number of political scientists who see experiments as useful and, in some cases, transformative.

Experiments appeal to our discipline because of their potential to generate stark and powerful empirical claims. Experiments can expand our abilities to change how critical target

audiences think about important phenomena. The experimental method produces new inferential power by inducing researchers to exercise control over the subjects of study, to randomly assign subjects to various conditions, and to carefully record observations. Political scientists who learn how to design and conduct experiments carefully are often rewarded with a clearer view of cause and effect.

While political scientists disagree about a great many methodological matters, perhaps there is a consensus that political science best serves the public when its findings give citizens and policymakers a better understanding of their shared environs. When such understandings require stark and powerful claims about cause and effect, the discipline should encourage experimental methods. When designed in a way that target audiences find relevant, experiments can enlighten, inform, and transform critical aspects of societal organization.

References

- Bailenson, Jeremy N., Shanto Iyengar, Nick Yee, and Nathan A. Collins. 2008. "Facial Similarity between Voters and Candidates Causes Influence." *Public Opinion Quarterly* 72: 935-61.
- Brody, Richard A., and Charles N. Brownstein. 1975. "Experimentation and Simulation." In *Handbook of Political Science 7*, eds. Fred I. Greenstein, and Nelson Polsby. Reading, MA: Addison-Wesley.
- Brown, Stephen R., and Lawrence E. Melamed. 1990. *Experimental Design and Analysis*. Newbury Park, London: Sage Publications.
- Campbell, Donald T. 1969. "Prospective: Artifact and Control." In *Artifact in Behavioral Research*, eds. Robert Rosenthal, and Robert Rosnow. Academic Press.
- Cover, Albert D., and Bruce S. Brumberg. 1982. "Baby Books and Ballots: The Impact of Congressional Mail on Constituent Opinion." *American Political Science Review* 76:347-59.
- Dipboye, Robert L., and Michael F. Flanagan. 1979. "Research Settings in Industrial and Organizational Psychology: Are Findings in the Field More Generalizable Than in the Laboratory?" *American Psychologist* 34:141-50.

- Druckman, Daniel. 1994. "Determinants of Compromising Behavior in Negotiation: A Meta-Analysis." *Journal of Conflict Resolution* 38: 507-56.
- Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research Political Science." *American Political Science Review* 100: 627-36.
- Druckman, James N. and Arthur Lupia. 2006. "Mind, Will, and Choice." In *The Oxford Handbook on Contextual Political Analysis*, eds. Charles Tilly, and Robert E. Goodin. Oxford: Oxford University Press.
- Druckman, James N., and Rose McDermott. 2008. "Emotion and the Framing of Risky Choice." *Political Behavior* 30: 297-321.
- Eavey, Cheryl L., and Gary J. Miller. 1984. "Bureaucratic Agenda Control: Imposition or Bargaining?" *American Political Science Review* 78: 719-33.
- Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50: 154-65.
- Esterling, Kevin Michael, David Lazer, and Michael Neblo. 2009. "Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment Involving Members of Congress and their Constituents." Unpublished paper, University of California, Riverside.
- Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are A Major Source of Knowledge in the Social Sciences." *Science* 326: 535-38.
- Ferraz, Claudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123: 703-45.
- Fowler, James H., and Darren Schreiber. 2008. "Biology, Politics, and the Emerging Science of Human Nature." *Science* 322(November 7): 912-14.
- Fréchette, Guillaume, John H. Kagel, and Steven F. Lehrer. 2003. "Bargaining in Legislatures." *American Political Science Review* 97: 221-32.
- Frohlich, Norman, and Joe A. Oppenheimer. 1992. *Choosing Justice: An Experimental Approach to Ethical Theory*. Berkeley: University of California Press.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout." *American Political Science Review* 94: 653-63.

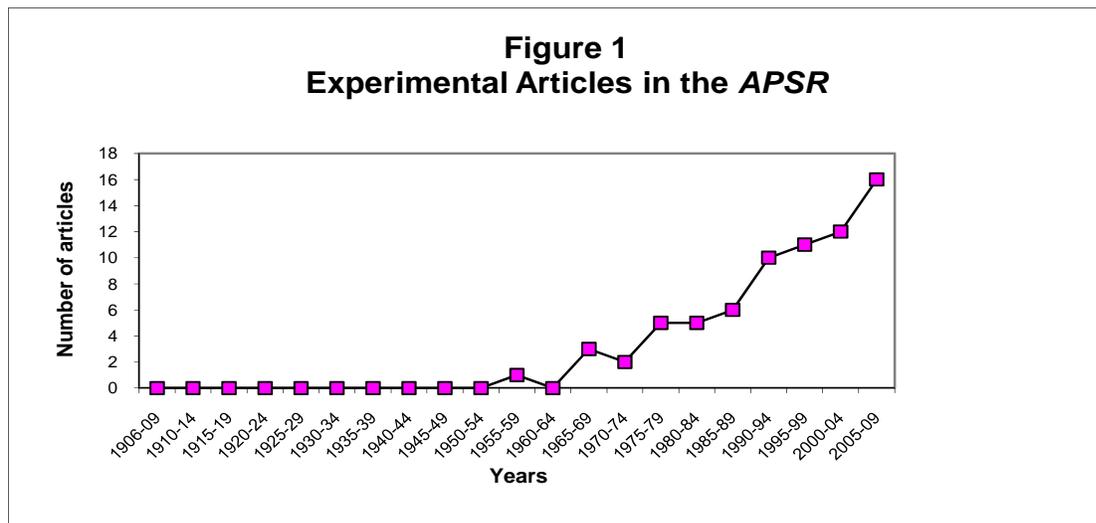
- Geva, Nehemia, and Alex Mintz. 1997. *Decision-making on War and Peace: The Cognitive-Rational Debate*. Boulder, CO: Lynne Rienner Publishers.
- Gilens, Martin. 1996. "'Race Coding' and White Opposition to Welfare." *American Political Science Review* 90: 593-604.
- Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review* 20: 869-74.
- Green, Donald P. 1992. "The Price Elasticity of Mass Preferences." *American Political Science Review* 86: 128-48.
- Green, Donald P., and Alan S. Gerber. 2002. "The Downstream Benefits of Experimentation." *Political Analysis* 10: 394-402.
- Greenberg, Jerald. 1987. "The College Sophomore as Guinea Pig: Setting the Record Straight." *Academy of Management Review* 12: 157-9.
- Guetzkow, Harold, and Joseph J. Valadez, eds. 1981. *Simulated International Processes*. Beverly Hills, CA: Sage.
- Habyarimana, James, Macartan Humphreys, Daniel Posner, and Jeremy M. Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101: 709-25.
- Halpern, Sydney A. 2004. *Lesser Harms: The Morality of Risk in Medical Research*. Chicago: University of Chicago Press.
- Hauck, Robert J-P. 2008. "Protecting Human Research Participants, IRBs, and Political Science Redux: Editor's Introduction." *PS: Political Science & Politics* 41: 475-6.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds. 2004. *Foundations of Human Society*. Oxford: Oxford University Press.
- Humphreys, Macartan, Daniel N. Posner, and Jeremy M. Weinstein. 2002. "Ethnic Identity, Collective Action, and Conflict: An Experimental Approach." Paper presented at the annual meeting of the American Political Science Association, Boston, MA.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: The University of Chicago Press.
- Johnson, George. 2008. *The Ten Most Beautiful Experiments*. New York: Alfred A. Knopf.
- Kardes, Frank R. 1996. "In Defense of Experimental Consumer Psychology." *Journal of Consumer Psychology* 5: 279-96.
- Keppel, Geoffrey, and Thomas D. Wickens. 2004. *Design and Analysis: A Researcher's Handbook*. 4th Ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
- Kühberger, Anton. 1998. "The Influence of Framing on Risky Decisions." *Organizational Behavior and Human Decision Processes* 75: 23-55.
- Levitt, Steven D., and John A. List. 2007. "What do Laboratory Experiments Measuring Social Preferences tell us about the Real World?" *Journal of Economic Perspectives* 21: 153-74.
- Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65: 682-93.
- Lodge, Milton, Kathleen M. McGraw, and Patrick Stroh. 1989. "An Impression-driven Model of Candidate Evaluation." *American Political Science Review* 83: 399-419.
- Lodge, Milton, Marco R. Steenbergen, and Shawn Brau. 1995. "The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89: 309-26.
- Lowell, A. Lawrence. 1910. "The Physiology of Politics." *American Political Science Review* 4: 1-15.
- Mahoney, Robert, and Daniel Druckman. 1975. "Simulation, Experimentation, and Context." *Simulation & Games* 6: 235-70.
- Malesky, Edmund J., and Krislert Samphantharak. 2008. "Predictable Corruption and Firm Investment: Evidence from a Natural Experiment and Survey of Cambodian Entrepreneurs." *Quarterly Journal of Political Science* 3: 227-67.
- Miller, Gary J., Thomas H. Hammond, and Charles Kile. 1996. "Bicameralism and the Core: An Experimental Test." *Legislative Studies Quarterly* 21: 83-103.
- Morton, Rebecca B. 1993. "Incomplete Information and Ideological Explanations of Platform Divergence." *American Political Science Review* 87: 382-92.
- Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.
- Morton, Rebecca B., and Kenneth Williams. 1999. "Information Asymmetries and Simultaneous versus Sequential Voting." *American Political Science Review* 93: 51-67.
- Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and Without a Sword." *American Political Science Review* 86: 404-17.
- Richardson, Liz, and Peter John. 2009. "Is Lobbying Really Effective? A Field Experiment of Local Interest Group Tactics to Influence Elected Representatives in the UK." Paper

presented at the European Consortium for Political Research Joint Sessions, Lisbon, Portugal.

- Riker, William H. 1967. "Bargaining in a Three-Person Game." *American Political Science Review* 61: 642-56.
- Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, eds. John H. Kagel, and Alvin E. Roth. Princeton, NJ: Princeton University Press.
- Simon, Adam, and Tracy Sulkin. 2001. "Habermas in the Lab: An Experimental Investigation of the Effects of Deliberation." *Political Psychology* 22: 809-826.
- Singer, Eleanor, and Felice J. Levine. 2003. "Protection of Human Subjects of Research: Recent Developments and Future Prospects for the Social Sciences." *Public Opinion Quarterly* 67: 148-64.
- Smith, Vernon L. 1976. "Experimental Economics: Induced Value Theory." *American Economic Review* 66: 274-79.
- Sniderman, Paul M, Look Hagendoorn, and Markus Prior. 2004. "Predispositional Factors and Situational Triggers." *American Political Science Review* 98: 35-50.
- Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior." *World Politics* 55: 399-422.

Figure 1-1. Experimental Articles in the *APSR*



ⁱ Parts of this chapter come from Druckman, Green, Kuklinski, and Lupia (2006).

ⁱⁱ This definition implicitly excludes so-called natural experiments, where nature initiates a random process. We discuss natural experiments in the next chapter.

ⁱⁱⁱ Brown and Melamed (1990) explain that “[r]andomization procedures mark the dividing line between classical and modern experimentation and are of great practical benefit to the experimenter” (3).

^{iv} Gosnell’s (1926) well known voter mobilization field study was not strictly an experiment as it did not employ random assignment.

^v The number of experiments has not only grown, but experiments appear to be particularly influential in shaping research agendas. Druckman, Green, Kuklinski, and Lupia (2006) compared the citation rates for experimental articles published in the *APSR* (through 2005) with the rates for (a) a random sample of approximately six non-experimental articles in every *APSR* volume where at least one experimental article appeared, (b) that same random sample narrowed to include only quantitative articles, and (c) the same sample narrowed to two articles on the same substantive topic that appeared in the same year as the experimental article or in the year before it appeared. They report that experimental articles are cited significantly more often than each of the comparison groups of articles (e.g., respectively, 47%, 74% and 26% more often).

^{vi} The theories need not be formal; for example, Lodge and his colleagues have implemented a series of experiments to test psychological theories of information processing (e.g., Lodge, McGraw, and Stroh 1989; Lodge, Steenbergen, and Brau 1995).

^{vii} In some cases, whether an experiment is one type or another is ambiguous (e.g., a web-survey administered in a classroom); the distinctions can be amorphous.

^{viii} As Campbell (1969) states, “...had we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Wistar staring white rats” (361).

^{ix} Of the laboratory experiments identified as appearing in the *APSR* through 2005, half employed induced value theory, such that participants received financial rewards contingent on their performance in the experiment. Thirty-one percent of laboratory experiments used deception; no experiments used both induced value and deception.

^x Perhaps the most notable topic absent from our introductory chapters is ethics and institutional review boards. We do not include a chapter on ethics because it is our sense that, to date, it has not surfaced as a major issue in political science experimentation. Additionally, more general relevant discussions are readily available (e.g., Singer and Levine 2003; Hauck 2008). Also see Halpern (2004) on ethics in clinical trials. Other methodological topics for which we do not have chapters include internet methodology and quasi-experimental designs.

^{xi} Of course, many others played critical roles in the development of experimental political science, and we take some comfort that most of these others have contributed to other volume chapters.

I. Designing Experiments

2. Experiments: An Introduction to Core Concepts

James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupiaⁱ

The experimental study of politics has exploded in the past two decades. Part of that explosion takes the form of a dramatic increase in the number of published articles that use experiments. Perhaps less evident, and arguably more important, experimentalists are exploring topics that would have been unimaginable only a few years ago. Laboratory researchers have studied topics ranging from the effects of media exposure (Iyengar and Kinder 1987) to the conditions under which groups solve collective action problems (Ostrom et al. 1992), and, at times, have identified empirical anomalies that produced new theoretical insights (McKelvey and Palfrey 1992). Some survey experimenters have developed experimental techniques to measure prejudice (Kuklinski et al. 1997) and its effects on support for policies such as welfare or affirmative action (Sniderman and Piazza 1995), while others have explored the ways in which framing, information, and decision cues influence voters' policy preferences and support for public officials (Druckman 2004; Tomz 2007). And while the initial wave of field experiments focused on the effects of campaign communications on turnout and voters' preferences (Eldersveld 1956; Gerber and Green 2000; Wantchekon 2003), researchers increasingly use field experiments to study phenomena as varied as election fraud (Hyde 2009), representation (Butler and Nickerson 2009), counterinsurgency (Lyll 2009), and interpersonal communication (Nickerson 2008).

With the rapid growth and development of experimental methods in political science come a set of terms and concepts that political scientists must know and understand. In this

chapter, we review concepts and definitions that often appear in the *Handbook* chapters. We also highlight features of experiments that are unique to political science.

1. What Is An Experiment?

In contrast to modes of research that address descriptive or interpretive questions, researchers design experiments to address causal questions. A causal question invites a comparison between two states of the world: one in which some sort of intervention is administered (a treated state; i.e., exposing a subject to a stimulus) and another in which it is not (an untreated state). The *fundamental problem of causal inference* arises because we cannot simultaneously observe a person or entity in its treated and untreated states (Holland 1986).

Consider, for example, the causal effect of viewing a presidential debate. Rarely are the elections of 1960, 1980, 1984, or 2000 recounted without mentioning the critical role that debates played in shaping voter opinion. What is the basis for thinking that viewing a presidential debate influences the public's support for the candidates? We do not observe how viewers of the debate would have voted had they not seen the debate. We do not observe how non-viewers would have voted had they watched (Druckman 2003). Nature does not provide us with the observations we would need to make the precise causal comparisons that we seek.

Social scientists have pursued two empirical strategies to overcome this conundrum: observational research and experimental research. *Observational research* involves a comparison between people or entities subjected to different treatments (at least, in part, of their own choosing). In the example referenced above, suppose that some people watched the debates, while others did not. To what extent can we determine the effect of debate-watching by comparing the post-debate behaviors of viewers and non-viewers? The answer depends on the extent to which viewers and non-viewers are truly comparable. It might be that most debate

watchers already supported one candidate, while most non-watchers favored the other. In such cases, observed differences between the post-debate opinions of watchers and non-watchers could stem largely from differences in the opinions they held before the debate even started. Hence, to observe that viewers and non-viewers express different views about a candidate after a debate does not say unequivocally that watching the debate caused these differences.

In an effort to address such concerns, observational researchers often attempt to compare treated and untreated people only when they share certain attributes, such as age or ideology. Researchers implement this general approach in many ways (e.g., multiple regression analysis, case-based matching, case control methodology), but all employ a similar underlying logic: find a group of seemingly comparable observations that have received different treatments, then base the causal evaluation primarily or exclusively on these observations.

Such approaches often fail to eliminate comparability problems. There might be no way to know whether individuals who look similar in terms of a (usually limited) set of observed attributes would in fact have responded identically to a particular treatment. Two groups of individuals who look the same to researchers could differ in multiple and unmeasured ways (e.g., openness to persuasion). This problem is particularly acute when people self-select into or out of a treatment. Whether people decide to watch or not watch a debate, for example, might depend on unmeasured attributes that predict which candidate they support (e.g., people who favor the front-running candidate before the debate might be more likely to watch the debate than those who expect their candidate to lose).

Experimental research differs from observational research in that the entities under study are randomly assigned to different treatments. Here, *treatments* refer to potentially causal interventions. For example, an experimenter might assign some people to watch a debate (one

treatment) and assign others to watch a completely different program (a second treatment). In some, but not all designs, there also is a *control group* that does not receive a treatment (e.g., they are neither told to watch nor discouraged from watching the debate) and/or multiple treatment groups (e.g., each group is told to watch a different part of the debate). *Random assignment* means that each entity being studied has an equal chance to be in a particular treatment condition.ⁱⁱ

Albertson and Lawrence (2009) and Mullainathan et al. (2010), for example, discuss experiments with *encouragement designs* in which the researcher randomly encourages some survey respondents to view an upcoming candidate debate (treatment group) and neither encourages or discourages others (control group). After the debate, the researcher conducts a second interview with both groups in order to ascertain whether they watched the debate and to measure their candidate preferences.

How does random assignment overcome the fundamental problem of causal inference? Suppose for the time being that everyone who was encouraged to view the debate did so and that no one watched unless encouraged. Although we cannot observe a given individual in both his/her treated and untreated states, random assignment enables the researcher to estimate the *average treatment effect*. Prior to the intervention, the randomly assigned treatment and control groups have the same expected responses to viewing the debate. Apart from random sampling variability, in other words, random assignment provides a basis for assuming that the control group behaves as the treatment group would have behaved had it not received the treatment (and vice versa). By comparing the average outcome in the treatment group to the average outcome in the control group, the experimental researcher estimates the average treatment effect. Moreover, the researcher can perform statistical tests to clarify whether the differences between groups

happened simply by chance (sampling variability) rather than as a result of experimental treatments.

When we speak of an experiment in this *Handbook*, we mean a study in which the units of observation (typically, subjects or human participants in an experiment) are randomly assigned to different treatment or control groups (although see note 2). Experimental studies can take many forms. It is customary to classify randomized studies according to the settings in which they take place: a *lab experiment* involves an intervention in a setting created and controlled by the researcher; a *field experiment* takes place in a naturally occurring setting; and a *survey experiment* involves an intervention in the course of an opinion survey (which might be conducted in-person, over the phone, or via the web). This classification scheme is not entirely adequate, however, as studies often blend different aspects of lab, field, and survey experiments. For example, some experiments take place in lab-like settings, such as a classroom, but require the completion of a survey that contains the experimental treatments (e.g., the treatments might entail providing individuals with different types of information about an issue).

2. Random Assignment or Random Sampling?

When evaluating whether a study qualifies as an experiment, by our definition, random *assignment* should not be confused with random *sampling*. Random sampling refers to a procedure by which participants are selected for certain kinds of studies. A common random sampling goal is to choose participants from a broader population in a way that gives every potential participant the same probability of being selected into the study. Random assignment differs. It does not require that participants be drawn randomly from some larger population. Experimental participants might come from undergraduate courses or from particular towns. The key requirement is that a random procedure, such as a coin flip, determines whether they receive

a particular treatment. Just as an experiment does not require a random sample, a study of a random sample need not be an experiment. A survey that merely asks a random sample of adults whether they watched a presidential debate might be a fine study, but it is not an experimental study of the effects of debate-viewing because watching or not watching the debate was not randomly assigned.

The typical social science experiment uses a *between-subjects design*, insofar as the researcher randomly assigns participants to distinct treatment groups. An alternative approach is the *within-subjects design* in which a given participant is observed before and after receiving a treatment (e.g., there is no random assignment between subjects). Intuitively, the within-subjects design seems to overcome the fundamental problem of causal inference; in practice, it is often vulnerable to confounds – meaning, unintended and uncontrolled factors that influence the results. For example, suppose that a researcher measures subjects’ attitudes toward a candidate before they watch a debate and then again after they have watched it, to determine whether the debate changed their attitudes. If subjects should hear attitude-changing news about the candidate after the first measurement and prior to the second, or if simply filling out the pre-debate questionnaire induces them to watch the debate differently than they otherwise would have watched, a comparison of pre- and post-attitudes will produce misleading conclusions about the effect of the debate.ⁱⁱⁱ

3. Internal and External Validity

Random assignment enables the researcher to formulate the appropriate comparisons, but random assignment alone does not ensure that the comparison will speak convincingly to the original causal question. The theoretical interpretation of an experimental result is a matter of *internal validity* – “did in fact the experimental stimulus [e.g., the debate] make some significant

difference [e.g., in attitude toward the candidate] in this specific instance” (Campbell 1957, 297).^{iv} In the preceding example, the researcher seeks to gauge the causal effect of viewing a televised debate, but if viewers of the debate are inadvertently exposed to attitude-changing news, the estimated effect of viewing the debate will be conflated with the effect of hearing the news.

The interpretation of the estimated causal effect also depends on what the control group receives as a treatment. If, in the above example, the control group watches another TV program that airs campaign commercials, the researcher must understand the treatment effect as the relative influence of viewing debates compared to viewing commercials.^v This comparison differs from a comparison of those who watch a debate with those who, experimentally, watch nothing.

More generally, every experimental treatment entails subtle nuances that the researcher must know, understand, and explicate. Hence, in the example above, he or she must judge whether the causative agent was viewing a debate per se, viewing any 90 minute political program, or viewing any political program of any length. Researchers can, and should, conduct multiple experiments or experiments with a wide array of different conditions in an effort to isolate the precise causative agent but, at the end of the day, they must rely on theoretical stipulations to decide which idiosyncratic aspects of the treatment are relevant and explain why they, and not others, are relevant.

Two aspects of experimental implementation that bear directly on internal validity are *noncompliance* and *attrition*. Noncompliance occurs when those assigned to the treatment group do not receive the treatment, or when those assigned to the control group inadvertently receive the treatment (e.g., those encouraged to watch do not watch or those not encouraged do watch).

In this case, the randomly assigned groups remain comparable, but the difference in their average outcomes measures the effect of the experimental assignment rather than actually receiving the treatment. The appendix to this chapter describes how to draw causal inferences in such circumstances.

Attrition involves the failure to measure outcomes for certain subjects (e.g., some do not report their vote preference in the follow-up). Attrition is particularly problematic when it afflicts some experimental groups more than others. For example, if debate viewers become more willing than non-viewers to participate in a post-debate interview, comparisons between treatment and control group could be biased. Sometimes researchers unwittingly contribute to the problem of differential attrition by exerting more effort to gather outcome data from one of the experimental groups or by expelling participants from the study if they fail to follow directions when receiving the treatment.

A related concern for experimental researchers is *external validity*. Researchers typically conduct experiments with an eye toward questions that are bigger than ‘What is the causal effect of the treatment on this particular group of people?’ For example, they may want to provide insight about voters generally, despite having data on relatively few voters. How far one can generalize from the results of a particular experiment is a question of *external validity*: the extent to which the “causal relationship holds over variations in persons, settings, treatments, and outcomes” (Shadish et al. 2002, 83).^{vi}

As suggested in the Shadish et al. quote, external validity covers at least four aspects of experimental design: whether the participants resemble the actors who are ordinarily confronted with these stimuli, whether the context (including the time) within which actors operate resembles the context (and time) of interest, whether the stimulus used in the study resembles the

stimulus of interest in the world, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest. The fact that several criteria come into play means that experiments are difficult to grade in terms of external validity, particularly since the external validity of a given study depends on what kinds of generalizations one seeks to make.

Consider the external validity of our example of the debate-watching encouragement experiment. The subjects in encouragement studies come from random samples of the populations of adults or registered voters. Random sampling bolsters the external validity of the study insofar as the people in the survey better reflect the target population. However, if certain types of people comply with encouragement instructions more than others, then our post-treatment inferences will reflect differences between compliers and non-compliers that are unrelated to any effects of watching the debate.

A related concern in such experiments is whether the context and time at which participants watch the debate resembles settings to which the researcher hopes to generalize. Are the viewers allowed to ignore the debate and read a magazine if they wish (as they could outside of the study)? Are they watching with the same types of people they would watch with outside of the study? There also are questions about the particular debate program used in the study (e.g., the stimulus): does it typify debates in general? To the extent that it does not, it will be harder to make general claims about debate-viewing that are regarded as externally valid. Before generalizing from the results of such an experiment, we would need to know more about the tone, content, and context of the debate.^{vii}

Finally, suppose our main interest is in how debate-viewing affects Election Day behaviors. If we wish to understand how exposure to debates affects voting, a questionnaire given on Election Day might be regarded as a better measurement than one taken immediately

after the debate and well before the election, since behavioral intentions may change after the debate but before the election.

Whether any of these concerns make a material difference to the external validity of an experimental finding can be addressed as part of an extended research program in which scholars vary relevant attributes of the research design, such as the subjects targeted for participation, the alternative viewing (or reading) choices available (to address the generalizability of effects from watching a particular debate in a particular circumstance), the types of debates watched, and the timing of post-debate interviews. A series of such experiments could address external validity concerns by gradually assessing how treatment effects vary depending on different attributes of experimental design.

4. Documenting and Reporting Relationships

When researchers detect a statistical relationship between a randomly assigned treatment and an outcome variable, they often want to probe further to understand the mechanisms by which the effect is transmitted. For example, having found that watching a televised debate increased the likelihood of voting, they ask why watching the debate has this effect. Is this because viewers become more interested in the race? Do they feel more confident about their ability to cast an intelligent vote? Do debates elevate their feelings of civic duty? Viewing a debate could change any of these *mediating variables*.

Assessing the extent to which potential mediating variables explain an experimental effect can be challenging. Analytically, a single random assignment (viewing a debate vs. not viewing) makes it difficult if not impossible to isolate the mediating pathways of numerous intervening variables. To clarify such effects, a researcher needs to design several experiments, all with different kinds of treatments. In the debate example, a researcher could ask different

subjects to watch different kinds of debates, with some treatments likely to affect interest in the race and others to heighten feelings of civic duty. Indeed, an extensive series of experiments might be required before a researcher can make convincing causal claims about causal pathways.

In addition to identifying mediating variables, researchers often want to understand the conditions under which an experimental treatment affects an important outcome. For example, do debates only affect (or affect to a greater extent) political independents? Do debates matter only when held in close proximity to Election Day? These are questions about *moderation*, wherein the treatment's effect on the outcome differs across levels of other variables (e.g., partisanship, timing of debate [see Baron and Kenny 1986]). Documenting moderating relationships typically entails the use of statistical interactions between the moderating variable and the treatment. This approach, however, requires sufficient variance on the moderating variable. For example, to evaluate whether debates affect only independents, the subject population must include sufficient numbers of otherwise comparable independents and non-independents.

In practice, pinpointing mediators and moderators often requires theoretical guidance and the use of multiple experiments representing distinct conditions. This gets at one of the great advantages of experiments – they can be *replicated* and extended in order to form a body of related studies. Moreover, as experimental literatures develop, they lend themselves to *meta-analysis*, a form of statistical analysis that assesses the conditions under which effects are large or small (Borenstein et al. 2009). Meta-analyses aggregate all of the experiments on a given topic into a single dataset and test whether effect sizes vary with certain changes in the treatments, subjects, context, or manner in which the experiments were implemented. Meta-analysis can reveal statistically significant treatment effects from a set of studies that, analyzed separately,

would each generate estimated treatment effects indistinguishable from zero. Indeed, it is this feature of meta-analysis that argues against the usual notion that one should always avoid conducting experiments with low *statistical power*, or a low probability of rejecting the null hypothesis of no effect (when there is in fact an effect).^{viii} A set of low power studies taken together might have considerable power, but if no one ever launches a low power study, this needed evidence cannot accumulate (for examples of meta-analyses in political science, see Druckman 1994; Lau et al. 1999).^{ix}

Publication bias threatens the accumulation of experimental evidence through meta-analysis. Some experiments find their way into print more readily than others. Those that generate statistically significant results and show that the effect of administering a treatment is clearly non-zero are more likely to be deemed worthy of publication by journal reviewers, editors, and even authors themselves. If statistically significant positive results are published while weaker results are not, the published literature will give a distorted impression of a treatment's influence. A meta-analysis of results that have been published selectively might be quite misleading. For example, if only experiments documenting that debates affect voter opinion survive the publication process, while those that report no effects are never published, then the published literature may provide a skewed view of debate effects. For this reason, researchers who employ meta-analysis should look for symptoms of publication bias, such as the tendency for smaller studies to generate larger treatment effects.

As the discussions of validity and publication bias suggest, experimentation is no panacea.^x The interpretation of experimental results requires intimate knowledge of how and under what conditions an experiment was conducted and reported. For this reason, it is incumbent on experimental researchers to give a detailed account of the key features of their

studies, including: 1) who the subjects are and how they came to participate in the study, 2) how the subjects were randomly assigned to experimental groups, 3) what treatments each group received, 4) the context in which they received them, 5) what the outcome measures were, and 6) all procedures used to preserve comparability between treatment and control groups, such as outcome measurement that is blind to participants' experimental assignments and the management of non-compliance and attrition possibilities.

5. Ethics and Natural Experiments

Implementing experiments in ways that speak convincingly to causal questions is important and challenging. Experiments that have great clarifying potential can also be expensive and difficult to orchestrate, particularly in situations where the random assignment of treatments means a sharp departure from what would ordinarily occur. For experiments on certain visible or conflictual topics, ethical problems might also arise. Subjects might be denied a treatment that they would ordinarily seek or be exposed to a treatment they would ordinarily avoid. Even if the ethical problems are manageable, such situations might also require researchers to garner potential subjects' explicit consent to participate in the experimental activities. Subjects might refuse to consent or the consent form might prompt them to think or behave in ways they otherwise would not; in both instances challenging the external validity of the experiment. Moreover, some studies include deception, an aspect of experimental design that raises not only ethical qualms but also practical concerns about jeopardizing the credibility of the experimental instructions in future experiments.

Hence, the creative spark required of a great experimental study is not just how to test an engaging hypothesis, but how to conduct a test while effectively managing practical and ethical constraints. In some cases, researchers address such practical and ethical hurdles by searching for

and taking advantage of random assignments that occur naturally in the world. These *natural experiments* include instances where random lotteries determine which men are drafted for military service (e.g., Angrist 1990), which incoming legislators enjoy the right to propose legislation (Loewen et al. 2009), or which Pakistani Muslims obtain visas allowing them to make the pilgrimage to Mecca (Clingsmith et al. 2008). The term natural experiment is sometimes defined more expansively to include events that happen to some people and not others, but the happenstance is not random. The adequacy of this broader definition is debatable; but when the mechanism determining whether or not people are exposed to a potentially relevant stimulus is sufficiently random, then these natural experiments can provide scholars with an opportunity to conduct research on topics that would ordinarily be beyond an experimenter's reach.

6. Conclusion

That social science experiments take many forms reflects different judgments about how best to balance various research aims. Some scholars prefer laboratory experiments to field experiments on the grounds that the lab offers the researcher tighter control over the treatment and how it is presented to subjects. Others take the opposite view on the grounds that generalization will be limited unless treatments are deployed, and outcomes assessed, unobtrusively in the field. Survey experiments are sometimes preferred on the grounds that a large and representative sample of people can be presented with a broad array of different stimuli in an environment where detailed outcome measures are easily gathered. Finally, some scholars turn to natural experiments in order to study historical interventions or interventions that could not, for practical or ethical reasons, be introduced by researchers.

The diversity of experimental approaches reflects in part different tastes about which research topics are most valuable, as well as ongoing debates within the experimental community

about how best to attack particular problems of causal inference. So it is difficult to make broad claims about “the right way” to run experiments in many substantive domains. In many respects, experimentation in political science is still in its infancy, and it remains to be seen which experimental designs, or combinations of designs, provide the most reliable political insights. That said, a good working knowledge of this chapter’s basic concepts and definitions can further understanding of the reasons behind the dramatic growth in the number and scope of experiments in political science, as well as the ways in which others are likely to evaluate and learn from the experiments that a researcher develops.

Appendix: An Introduction to the Neyman-Rubin Causal Model

The logic underlying randomized experiments is often explicated in terms of a notational system that has its origins in Neyman (1923) and Rubin (1974). For each individual i let Y_0 be the outcome if i is not exposed to the treatment, and Y_1 be the outcome if i is exposed to the treatment. The treatment effect is defined as:

$$(1) \quad \tau_i = Y_{i1} - Y_{i0}.$$

In other words, the treatment effect is the difference between two potential states of the world, one in which the individual receives the treatment, and another in which the individual does not. Extending this logic from a single individual to a set of individuals, we may define the average treatment effect (ATE) as follows:

$$(2) \quad ATE = E(\tau_i) = E(Y_{i1}) - E(Y_{i0}).$$

The concept of the average treatment effect implicitly acknowledges the fact that the treatment effect may vary across individuals. The value of τ_i may be especially large, for example, among those who seek out a given treatment. In such cases, the average treatment effect in the

population may be quite different from the average treatment effect among those who actually receive the treatment.

Stated formally, the concept of the average treatment effect among the treated may be written:

$$(3) \quad ATT = E(\tau_i|T_i=1) = E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=1),$$

where $T_i=1$ when a person receives a treatment. To clarify the terminology, $Y_{i1}|T_i=1$ is the outcome resulting from the treatment among those who are actually treated, whereas $Y_{i0}|T_i=1$ is the outcome that would have been observed in the absence of treatment among those who are actually treated. By comparing equations (2) and (3), we see that the average treatment effect need not be the same as the treatment effect among the treated.

This framework can be used to show the importance of random assignment. When treatments are randomly administered, the group that receives the treatment ($T_i=1$) has the same expected outcome as the group that does not receive the treatment ($T_i=0$) would if it were treated:

$$(4) \quad E(Y_{i1}|T_i=1) = E(Y_{i1}|T_i=0).$$

Similarly, the group that does not receive the treatment has the same expected outcome, if untreated, as the group that receives the treatment, if it were untreated:

$$(5) \quad E(Y_{i0}|T_i=0) = E(Y_{i0}|T_i=1)$$

Equations (4) and (5) are termed the independence assumption by Holland (1986) because the randomly assigned value of T_i conveys no information about the potential values of Y_i . Equations (2), (4), and (5) imply that the average treatment effect may be written:

$$(6) \quad ATE = E(\tau_i) = E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=0).$$

Because $E(Y_{i1}|T_i=1)$ and $E(Y_{i0}|T_i=0)$ may be estimated directly from the data, this equation suggests a solution to the problem of causal inference. To estimate an average treatment effect, we simply calculate the difference between two sample means: the average outcome in the treatment group minus the average outcome in the control group. This estimate is unbiased in the sense that, on average across hypothetical replications of the same experiment, it reveals the true average treatment effect.

Random assignment further implies that independence will hold not only for Y_i , but for any variable X_i that might be measured prior to the administration of the treatment. For example, subjects' demographic attributes or their scores on a pre-test are presumably independent of randomly assigned treatment groups. Thus, one expects the average value of X_i in the treatment group to be the same as the control group; indeed, the entire distribution of X_i is expected to be the same across experimental groups. This property is known as *covariate balance*. It is possible to gauge the degree of balance empirically by comparing the sample averages for the treatment and control groups.

The preceding discussion of causal effects skipped over two further assumptions that play a subtle but important role in experimental analysis. The first is the idea of an *exclusion restriction*. Embedded in equation (1) is the idea that outcomes vary as a function of receiving the treatment per se. It is assumed that assignment to the treatment group only affects outcomes insofar as subjects receive the treatment. Part of the rationale for using blinded placebo groups in experimental design is the concern that subjects' knowledge of their experimental assignment might affect their outcomes. The same may be said for double-blind procedures: when those who implement experiments are unaware of subjects' experimental assignments, they cannot intentionally or inadvertently alter their measurement of the dependent variable.

A second assumption is known as the *stable unit treatment value assumption*, or SUTVA. In the notation used above, expectations such as $E(Y_{il}|T_i=t_i)$ are all written as if the expected value of the treatment outcome variable Y_{il} for unit i only depends upon whether or not the unit gets the treatment (whether t_i equals one or zero). A more complete notation would allow for the consequences of treatments T_1 through T_n administered to other units. It is conceivable that experimental outcomes might depend upon the values of $t_1, t_2, \dots, t_{i-1}, t_{i+1}, \dots, t_n$ as well as the value of t_i :

$$E(Y_{il}|T_1=t_1, T_2=t_2, \dots, T_{i-1}=t_{i-1}, T_i=t_i, T_{i+1}=t_{i+1}, \dots, T_n=t_n) .$$

By ignoring the assignments to all other units when we write this as $E(Y_{il}|T_i=t_i)$ we assume away spillovers (or multiple forms of the treatment) from one experimental subject to another.

Noncompliance

Sometimes only a subset of those who are assigned to the treatment group is actually treated, or a portion of the control group receives the treatment. When those who get the treatment differ from those who are assigned to receive it, an experiment confronts a problem of *noncompliance*. In experimental studies of get-out-the-vote canvassing, for example, non-compliance occurs when some subjects that were assigned to the treatment group remain untreated because they are not reached (see Gerber, Green, Kaplan, and Kern 2010).

How experimenters approach the problem of noncompliance depends on their objectives. Those who wish to gauge the effectiveness of an outreach program may be content to estimate the so-called *intent-to-treat effect*, that is, the effect of being randomly assigned to the treatment. The intent-to-treat effect is essentially a blend of two aspects of the experimental intervention: the rate at which the assigned treatment is actually delivered to subjects and the effect it has on those who receive it. Some experimenters are primarily interested in the latter. Their aim is to

measure the effects of the treatment on *Compliers*, people who receive the treatment if and only if they are assigned to the treatment group.

When there is noncompliance, a subject's group assignment, Z_i , is not equivalent to T_i , whether the subject gets treated or not. Let $D_1=1$ when a subject assigned to the treatment group is treated, and let $D_1=0$ when a subject assigned to the treatment group is not treated. Define a subset of the population, called *Compliers*, who get the treatment when assigned to the treatment group but not otherwise. *Compliers* are subjects for whom $D_1=1$ and $D_0=0$. Note that whether a subject is a *Complier* is a function of both subject characteristics and the particular features of the experiment and is not a fixed attribute of a subject.

When treatments are administered exactly according to plan ($Z_i = T_i, \forall i$), the average causal effect of a randomly assigned treatment can be estimated simply by comparing mean treatment group outcomes and mean control group outcomes. What can be learned about treatment effects when there is noncompliance? Angrist et al. (1996) present a set of sufficient conditions for estimating the average treatment effect for the subgroup of subjects who are *Compliers*. Here we will first present a description of the assumptions and the formula for estimating the average treatment effect for the *Compliers*. We then examine the assumptions using an example.

In order to estimate the average treatment effect among *Compliers*, we must assume that assignment Z is random. In addition, we must make four additional assumptions: the exclusion restriction, SUTVA, monotonicity, and a non-zero causal effect of the random assignment. The exclusion restriction implies that the outcome for a subject is a function of the treatment they receive but is not otherwise influenced by their assignment to the treatment group. SUTVA implies that a subject's outcomes depend only on the subject's own treatment assignment and not

on the treatment assignment of any other subjects. Monotonicity means that there are no *Defiers*, that is, no subjects who would receive the treatment if assigned to the control group and would not receive the treatment if assigned to the treatment group. The final assumption is that the random assignment has some effect on the probability of receiving the treatment. With these assumptions in place, the researcher may estimate the average treatment effect among compliers in a manner that will be increasingly accurate as the number of observations in the study increases. Thus, while the problem of experimental crossover constrains a researcher's ability to draw inferences about the average treatment effect among the entire population, accurate inferences can often be obtained with regard to the average treatment effect among Compliers.

References

- Albertson, Bethany, and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 37: 275-300.
- Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80: 313-36.
- Baron, Reuben M., and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173-82.
- Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. London: Wiley.
- Butler, Daniel M., and David W. Nickerson. 2009. "How much does Constituency Opinion Affect Legislators' Votes? Results from a Field Experiment." Unpublished manuscript, Institution for Social and Policy Studies at Yale University.
- Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54: 297-312.
- Clingingsmith, David, Asim Ijaz Khwaja, and Michael Kremer. 2008. "Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering." Unpublished paper, Harvard University.

- Druckman, Daniel. 1994. "Determinants of Compromising Behavior in Negotiation: A Meta-Analysis." *Journal of Conflict Resolution* 38: 507-56.
- Druckman, James N. 2003. "The Power of Television Images: The First Kennedy-Nixon Debate Revisited." *Journal of Politics* 65: 559-71.
- Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98: 671-86.
- Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50: 154-65.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout." *American Political Science Review* 94: 653-63.
- Gerber, Alan S., Donald P. Green, Edward H. Kaplan, and Holger L. Kern. 2010. "Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments." *Political Analysis* 18: 297-315.
- Guala, Francesco. 2005. "The Methodology of Experimental Economics." New York: Cambridge University Press.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-60.
- Hyde, Susan D. 2009. *The Causes and Consequences of Internationally Monitored Elections*. Unpublished Book Manuscript, Yale University.
- Iyengar, Shanto, Mark D. Peters, and Donald R. Kinder. 1982. "Experimental Demonstrations of the 'Not-So-Minimal' Consequences of Television News Programs." *American Political Science Review* 76: 848-58.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: The University of Chicago Press.
- Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the New South." *Journal of Politics* 59: 323-49.
- Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Review." *American Political Science Review* 93: 851-75.
- Loewen, Peter John, Royce Koop, and James H. Fowler. 2009. "The Power to Propose: A Natural Experiment in Politics." Unpublished paper, University of British Columbia.
- Lyall, Jason. 2009. "Does Indiscriminant Violence Incite Insurgent Attacks?" *Journal of Conflict Resolution* 53: 331-62.

- Mullainathan, Sendhil, Ebonya Washington, and Julia R. Azari. 2010. "The Impact of Electoral Debate on Public Opinions: An Experimental Investigation of the 2005 New York City Mayoral Election." In *Political Representation*, eds. Ian Shapiro, Susan Stokes, Elizabeth Wood, and Alexander S. Kirshner. New York: Cambridge University Press.
- McKelvey, Richard D., and Thomas R. Palfrey. 1992. "An Experimental Study of the Centipede Game." *Econometrica* 4: 803-36.
- Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.
- Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles." *Statistical Science* 5: 465-80.
- Nickerson, David W. 2008. "Is Voting Contagious?: Evidence from Two Field Experiments." *American Political Science Review* 102: 49-57.
- Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review* 86: 404-17.
- Panagopoulos, Costas, and Donald P. Green. 2008. "Field Experiments Testing the Impact of Radio Advertisements on Electoral Competition." *American Journal of Political Science* 52: 156-68.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688-701.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Sniderman, Paul M., and Thomas Piazza. 1995. *The Scar of Race*. Cambridge, MA: Harvard University Press.
- Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61: 821-40.
- Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior." *World Politics* 55: 399-422.

ⁱ We thank Holger Kern for helpful comments.

ⁱⁱ In the social sciences, in contrast to the physical sciences, experiments tend to involve use of random assignment to treatment conditions. Randomly treatments are one type of "independent variable." Another type comprises "covariates" that are not randomly assigned but nonetheless predict the outcome.

ⁱⁱⁱ Natural scientists frequently use within-subjects designs because they seldom contend with problems of memory and anticipation when working with "subjects" like electrons. Clearly, natural scientists conduct "experiments" (with interventions) even if they do not employ between-subjects random assignment. Social scientists, confronted as they are by the additional complexities of working with humans, typically rely on between-subjects experimental designs, where randomization ensures that the experimental groups are, in expectation, identical.

That said, in some cases, particularly in economics (and, hence some of the work discussed in this *Handbook*), participants are not randomly assigned to distinct between-subject treatment conditions since the

purpose of the experiment is to test well-specified theoretical predictions. Even in these cases, however, researchers randomly assign participants to distinct roles in the experiment (Guala 2005, 79; Morton and Williams 2010, 28-29). For example, the dictatorship experiment entails allowing a subject to decide how much of a fixed sum of money to keep for him or herself and how much to give to another subject. This experiment is used to test predictions about self-interest (e.g., do subjects act entirely in their self-interest or do they split the difference?), yet subjects are randomly assigned to the roles (e.g., of giving or receiving money).

^{iv} Related to internal validity is statistical conclusion validity, defined as “the validity of inferences about the correlation (covariation) between treatment and outcome” (Shadish et al. 2002, 38). Statistical conclusion validity refers specifically and solely to the “appropriate use of statistics to infer whether the presumed independent and dependent variables co-vary,” and not at all to whether a true causal relationship exists (Shadish et al. 2002, 37).

^v Internal validity is a frequent challenge for experimental research. For this reason, experimental scholars often administer *manipulation checks*, evaluations that document whether subjects experience the treatment as intended by the experimenter.

^{vi} Related is construct validity which is “the validity of inferences about the higher order constructs that represent sampling particulars” (Shadish et al. 2002, 38).

^{vii} This is related to the aforementioned internal validity concern about whether the content of the debate itself caused the reaction, or whether any such programming would have caused it. The internal validity concern is about the causal impact of the presumed stimulus – is the cause what we believe it is (e.g., the debate and not any political programming)? The external validity concern is about whether that causal agent reflects the set of causal variables to which we hope to infer (e.g., is the content of the debate representative of presidential debates?).

^{viii} Statistical power refers to the probability that a researcher will reject the null hypothesis of no effect when the alternative hypothesis is indeed true.

^{ix} Early lab and field studies of the mass media fall into this category. Iyengar et al.’s (1982) influential lab study of TV news had fewer than twenty subjects in some of the experimental conditions. Panagopoulos and Green’s (2008) study of radio advertising comprised a few dozen mayoral elections. Neither produced overwhelming statistical evidence on its own, but both have been bolstered by replications.

^x The volume does not include explicit chapters on meta-analysis or publication bias, reflecting, in part, the still relatively recent rise in experimental methods (i.e., in many areas, there is not yet a sufficient accumulation of evidence). We imagine these topics will soon receive considerably more attention within political science.

3. Internal and External Validity

Rose McDermott

One of the challenges in conducting interdisciplinary work, or in attempting to communicate across disciplinary boundaries relates to the implicit norms which infuse different fields. Much like trying to speak across cultures, it often becomes frustrating to translate or make explicit differing assumptions underlying appropriate inferential methods and strategies. To make matters worse, status differentials often exacerbate these divergences, privileging one set of disciplinary norms over another, such that decisions about ideal methods do not always rest entirely on the appropriateness of a particular technique for a given project.

Such differences clearly affect the implementation of experimental methodology across the fields of psychology, economics, and political science. One of the areas in which these biases inflict misunderstanding surround issues related to internal and external validity. In political science, concerns with external validity often border on the monomaniacal, leading to the neglect, if not the complete dismissal, of attention to the important issues involved in internal validity. In psychology, the reverse emphasis predominates. In behavioral economics, the focus depends more on the primary function of the experiment. Since both internal and external validity remain important in assessing the quality, accuracy, and utility of any given experimental design, it facilitates optimal experimental design to concentrate on attempting to maximize both, but the nature of the enterprise often requires explicit consideration of the trade-offs between them.

The purpose of an experiment informs the degree to which emphasis should be placed on internal versus external validity. In some cases, as for example when studying some universal

human experience such as vision, college students are unlikely to significantly differ from the broader population on the dimension under investigation, and therefore the additional external validity which would be provided by replication across time and population will not be as important. In other circumstances, such as a study of the effect of testosterone on decision making in combat, external validity depends on finding participants in combat, or the entire purpose of the study becomes vitiated. Of course, the primary purpose of such a study would not aim for external validity, but rather for a deeper understanding of the effects of the endocrine system on social relationships.

Recognition of the methodological goals promoted by internal and external validity remains critical to the enterprise of achieving robust experimental findings. Perhaps the best way to conceptualize the balance between internal and external validity in experimental design is to think about them in a two-step temporal sequence. Internal validity comes first, both sequentially and practically. Without first establishing internal validity, it remains unclear what process should be explored in the real world. An experimenter has to know that the conclusions result from the manipulations imposed before trying to extrapolate those findings into other contexts. External validity follows, as replications across time and populations seek to delineate the extent to which these conclusions can generalize.

This chapter proceeds in four parts. Separate discussions of internal and external validity encompass the first two sections. A brief third part notes the trade-offs in value and practical logistics between the two. A meditation on future prospects for improving validity concludes.

1. Internal Validity

Campbell (1957) considered an experiment internally valid if the experimenter finds a significant difference between the treatment and control conditions. These differences are then

assumed to provide a meaningful reflection of the causal processes at play. As long as no reason exists to assume that some extraneous mediating factor systematically influenced subjects' responses, observers can attribute changes in the dependent variable to systematic manipulations across the independent variables. From this perspective, internal validity is enhanced by experiments that are well designed, carefully controlled and meticulously measured so that alternative explanations for the phenomena under consideration can be excluded. In other words, internal validity refers to the extent to which an experimenter can be confident their findings result from their experimental manipulations, although still remain uncertain as to how the mechanism might work in various settings or across diverse individuals. Shadish, Cook, and Campbell (2002) remain careful to note that internally valid findings remain discrete to the specific experimental context in which they are explored; generalization of any uncovered causal phenomena then depends on extensions to other populations, contexts, and situations which involve attempts to achieve external validity.

Internal validity remains intrinsically tied to experimental, as opposed to mundane, realism (Aronson et al. 1990; McDermott 2002). To the extent that subjects become psychologically engaged in the process they confront, internal validity intensifies. Similarly, internal validity diminishes in the face of subject disengagement, just as one might expect any action that would distract a subject would rob the study of its ability to specify and consolidate the causal factor of interest. If subjects approach a task with skepticism or detachment, genuine responses fade and strategic incentives come to the fore. This raises the possibility that measures obtained do not accurately reflect the process being manipulated, but rather manifest a different underlying construct altogether. It does not matter if the experimental environment does not overtly mimic the real world setting as long as the subject experiences the relevant forces the

investigator seeks to elicit. Because of this, the internal experience of the experiment for the subject need not necessarily reflect outside appearances.

The success of the experiment depends on the subject taking the task seriously, and experimenters can foster such engagement to the degree they can create and establish a situation which forces psychological investment on the part of subjects. For example, if an experimenter wants to study, say, processes of cooperation, it does not matter if the subject would be unlikely to run across the actual partners presented, as long as she responds to other subjects just as she would to any other potential ally. Similarly, it should not matter in a study of aggression that subjects are unlikely to have money taken from them as a result of their behavior in a simple economic game, as long as this behavior stimulates anger in them the way an actual theft or injustice would. The critical operative feature in such experimental designs revolves around the ability of the experimenter to create a psychological situation which realistically elicits the dynamics under consideration. In other words, internal validity equates to the manipulation of a psychological response.

Comparisons with Experimental Economics

Roth (1995) described three main purposes for experiments in economics, and this analysis was extended and applied to political science by Druckman et al. (2006). These goals included extending theoretical models, which he referred to as: 1) “speaking to theorists,” 2) data generation, which he called “searching for facts,” and 3) searching for meaning or policy applications, which he described as “whispering in the ears of princes.” Importantly, each of these functions requires slightly different foci and may engender greater concern with one type of validity over another.

In speaking to theorists, at least in experimental economics, the focus typically revolves around providing an experimental test of a formal model. In this context, economists tend to secure internal validity within the experimental paradigm itself. In such cases, a formal mathematical model will be generated, and then its predictions will be tested in an experimental setting to see how well or how closely actual human behavior conforms to the hypothetical expectation. The formal model may then be adjusted to accommodate the findings brought forth by the experiment. In such tests, focus on internal validity would remain almost exclusive, since scholars remain less concerned with the extent of generalization outside the lab and more interested in the performance of the model.

In searching for facts, the purpose of the experiment revolves around generating new data. This goal can take several forms. Sometimes investigators are inspired by previous experimental results, failure, or lacunae to explore an aspect of previous studies or theory that did not make sense, or resulted in inconsistent or inconclusive findings. Often in experimental economics these studies evolve almost as conversations between scholars using different theoretical or methodological tools to examine their variables of interest from differing perspectives. Many of the studies in experimental economics which seem to undertake endless variations on the theme of how people behave in the ultimatum game are motivated, at least in part, by the desire of investigators to define precisely those conditions under which a particular behavior, like fairness, will emerge, sustain, or dissipate. Sequences of studies can generate new hypotheses, or reveal novel areas of inquiry. Fehr's work on altruistic punishment followed such a sequence of inquiry into the extent and bounds of human cooperation (Fehr and Gächter 2002; Fehr and Fischbacher 2003; de Quervain et al. 2004). Such experimental research often points to alternative explanations for enduring puzzles, as when neuroeconomists began to explore

potential biological and genetic exigencies in motivating particular economic behavior, including risk (Sokol-Hessner et al. 2009).

Similar progress can be seen in psychology within the dialogue which emerged following the original studies of judgmental heuristics and biases conducted by Tversky and Kahneman (1974). Gigerenzer (1996) and others began to challenge the assumption that such biases represented mistakes in the human inferential process, and instead demonstrated that when such dynamics were experimentally tested in ecologically valid contexts, such as when people are called upon to detect cheaters, such ostensible errors in judgment evaporate.

Finally, whispering in the ears of princes speaks to the ways in which experimental designs can be generated to address central concerns held by policy makers and other decision makers. Here the obvious emphasis would lie more in the realm of external validity, since results would need to speak to broad populations in order to be of use to policy makers. However, importantly, such studies retain no real utility to the extent that they do not first measure what they claim to examine. Herein lies an important trade-off between attempting to create an experiment whose characteristics resemble situations familiar and important to many individuals, involving perhaps choices between political candidates, financial risk, or choices over health care or employment benefits, and retaining control over the manipulation and measurement of complex or multi-dimensional variables.

Threats to Internal validity

Campbell and Stanley (1966) delineated nine primary threats to internal validity which often lie outside the experimenter's ability to control. Their discussion remains the definitive characterization of the kinds of problems which most risk confidence in attributing changes in the dependent variable to manipulations of the independent variable. These challenges include

selection, history, maturation, repeated testing, instrumentation, regression toward the mean, mortality, experimenter bias, and selection-maturation interaction. Each of these threats presents critical challenges to good experimental design. The most important for purposes of examining validity relate to attrition – or mortality effects – and subject noncompliance, because high rates of either can influence both internal and external validity.

Mortality or attrition effects occur when subjects drop out of an experiment, or are otherwise lost to follow-up. This only poses a threat to internal validity to the extent that this occurs subsequent to random assignment (Kiesler et al. 1969). If subjects drop out prior to such assignment, it may constitute a threat to the external validity of the experiment, but not to the internal validity of its findings. However, if subjects in one condition are dropping out of an experiment at a higher rate than those in another condition, it may be the case that such attrition is in fact potentiated by the treatment itself. This relates to the issue of intention-to-treat (Angrist, Imbens, and Rubin 1996) When dealing with questions of noncompliance, observers must estimate the weighted average of two different effects, where it is usual to divide the apparent effect from the proportion of the people who actually receive the treatment. This source of invalidity (nonexposure to treatment) thus remains correctable under certain assumptions (Nickerson 2005). For example, if an experimenter does not come to your door, he knows he did not have an effect on the measured outcome. Therefore if an effect is observed, then the experimenter must search for the cause elsewhere. If scholars prove willing to make some assumptions about the lack of effect on those who were not treated, it becomes possible to statistically back out the effect of the putative causal variable on those who were treated. In that way, observers can see the raw difference, and not merely the effects of biases in validity within the design itself, on the people who received it.

This can be particularly problematic in medical studies if a certain drug causes prohibitive side effects which preclude a high percentage of people from continuing treatment. In such cases, subject mortality itself constitutes an important dependent variable in the experiment. In medical research, this issue is sometimes discussed in terms of intent-to-treat effects, where the normative prescription requires analyzing data subsequent to randomization regardless of a subject's adherence to treatment or subsequent withdrawal, noncompliance or deviation from the experimental protocol. This can bias results if the withdrawal from the experiment resulted from the treatment itself and not from some other extraneous factor.

Of course, subject mortality poses a much more severe threat in between-subject designs, especially in matched or blocked designs where the loss of one subjects becomes equivalent to the loss of two. This does not mean that a block design allows experimenters to drop blocks in the face of attrition. Drawing on examples from field experiments in voter mobilization campaigns, Nickerson (2005) demonstrates how this strategy produces bias except under special circumstances.

Proxy outcomes do not always constitute a valid measure of the topic of interest. For example, in AIDS research, focusing on numbers of T-cells as indicators of immune system function may prove helpful for researchers, but only remains significant for patients to the extent that these values correlate significantly with clinical outcomes of interest, such as quality of life, or longevity itself. In medicine, it is often the case that proxy measures which are assumed to correlate with outcome measures of concern actually represent orthogonal values; the best recent example relates to findings that even well controlled blood sugar does not fully mitigate the risk of heart disease among diabetics. Political discussions suffer from a similar dynamic. In work on the influence of negative advertising on voting, the important effect exerts itself not only to the

immediate influence on voter choice, but also relates to downstream effects, such as suppression of overall voter turnout. In turn, scholars might not be so interested in this variable if they were not focused on vote choice, but remained more concerned with the sources of public opinion on its own. Often researchers care about the reaction to the stimuli as much as they care about its immediate effect. Whether questions surrounding internal or external validity in circumstances involving proxy, intervening and intermediate variables remain problematic often depends on whether a scholar is interested in the immediate or the downstream effect of a variable.

In addition, mortality often arises as an issue in iterated contexts not so much because subjects die, but rather because long term follow-ups are often so costly and difficult that interim measures are instituted as dependent variables to replace the real variables of interest. In the statistics literature, these are referred to as “principal surrogates.” So, for example, in medical experiments, blood sugar is treated as the variable of interest in studies of diabetes rather than longevity. Transitional voter turnout remains monotonically related to political legitimacy; however, it does not appear obvious that an observer gets more purchase on such larger overall issues using proxy variables. Such a focus hid the finding that many blood sugar medications, which indeed controlled blood sugar, nonetheless potentiated higher rates of cardiovascular related deaths than uncontrolled blood sugar caused.

Another problem is active or passive forms of noncompliance. Many standard forms of analysis assume that everyone who receives treatment experiences similar levels of engagement with the protocol, but this assumption often remains faulty, since noncompliance rates can be nontrivial. Noncompliance raises the prospect of artificially inducing treatment values into the estimated outcome effects. If this happens, it may not necessarily introduce systematic bias, but it may also not provide any useful information about the process the experimenter seeks to

understand. If subjects are not paying attention, or thinking about something unrelated to the task at hand, then it will remain unclear whether or not the manipulation actually exerted an effect. Null results under such conditions may represent true negatives, implying no effect where one may exist if subjects attended to the manipulation as intended. Angrist et al (1996) propose use of instrumental variables with the Rubin Causal Model in order to circumvent this problem. Arceneaux, Gerber, and Green (2006) demonstrate the superiority of the instrumental variables approach over a matching analysis in a large-scale voter mobilization experiment.

Instrumental variables are sometimes used when there is concern that the treatment and unobserved factors that might affect the outcome (i.e., the disturbance term) might be correlated in some significant way. To be clear, the issue is not the assigned treatment, but rather the actual treatment as received and experienced by the subject. These effects can differ for a number of reasons, not least among them subject noncompliance. When this happens, the concern arises that the treatment received by the subject is somehow related to the disturbance term. If this occurs, we could not know the true effect regardless of the analysis, because the disturbance remains unobserved. An instrumental variable is one which exerts its effect through an influence on the independent variable but has no direct effect on the dependent variable, nor is it systematically related to unobserved causes of the dependent variable. In other words, it is only related to the dependent variable through the mediating effect of the endogenous independent variable. Gerber and Green (2000) provide an example of this with regard to voter turnout effects. In their study, random assignment to treatment determines whether a subject is successfully canvassed, which in turn appears to affect turnout. Assignment to the condition of being canvassed, which is random, remains unrelated to the disturbance term. The assignment to the condition of being canvassed only influences turnout through actual contact with voters; no

backdoor paths exist that would cause people to be affected, either directly or indirectly, by their assignment to being canvassed or not, other than their contact with the canvasser. Because of the expense involved in large experimental studies, researchers sometimes use instrumental observational variables to gain traction on the problem at hand. In other words, they tend to be used to try to infer causality by imagining that the independent variable is near random (Sovey and Green 2010) when experimental studies designed to determine true causation might not be possible for one reason or another.

Of course, the key to success depends on selecting valid variables. This can be difficult, but often nature or government can supply a useful instrument to use. For example, Miguel, Sayanth, and Sergenti (2004) use weather as an instrumental variable to examine the relationship between economic shocks and civil conflict. Certainly hurricanes, fires, or policy changes in large programs such as welfare might serve similar purposes. Such instrumental variables are feasible to the extent that the independent variable provides consistent estimates of causal effects when the instruments are independent of the disturbance term and correlated in a substantial way with the endogenous independent variable of interest. Successfully identifying such an instrument can help triangulate on the relationships and treatments of interest, but often finding such instruments can prove challenging. Alternative strategies for dealing with problems where the treatment and the outcome may be correlated exist, including intent-to-treat effects, which were discussed above.

Other, unrelated concerns, can also compromise prospects for the development of an effective experimental protocol and merit some consideration as well. Importantly, different kinds of experimental design pose greater risks in some areas than in others, and recognizing which designs present which challenges can prevent inadvertent error.

Several particularly problematic areas exist. Pseudoexperimental designs, where the researcher is not able to manipulate the independent variable, as well as experimental designs, which do not allow for the randomization of subjects across condition for practical or ethical reasons, present greater challenges for internal validity than more controlled laboratory experiments. In addition, field experiments raise the specter of subject noncompliance to a higher degree than more restricted laboratory settings.

Further, certain content areas of investigation may pose greater threats to internal validity than other topics. Honesty in investigating socially sensitive subjects, such as race or sex, may be compromised by subjects' desire for positive impression management. They may not want to admit the extent to which they harbor or espouse views that they know others may find offensive. Less obtrusive measurements, such as those involving reaction time tests, or implicit association measure, may help circumvent this problem. Alternatively, techniques which do not rely on subject report, such as analyses of brain waves or hormonal or genetic factors, may obviate the need for subject honesty, depending on the topic under investigation.

Regardless, certain realities inevitably constrain the ability of an investigator to know whether or not subjects are sufficiently engaged in an experimental task so as to justify reliance on the data generated by them. Any systematic restrictions in the performance of subjects can potentially contaminate the internal validity of the results obtained. The following discussion highlights some of the ways in which subjects can, intentionally or otherwise, impede internal validity in experimental tests. Some of these categories overlap with Campbell and Stanley, while others introduce additional concerns.

Good experimentalists should strive for the goal of trying to design an experiment from the subject's perspective, with an eye toward understanding what the person will see, hear and

experience in the setting created, and not with the singular purpose of achieving the fastest, most efficient way to collect the data they need. This becomes particularly important in behaviorally oriented tasks. Subject involvement is not an altruistic goal, but rather one that should be motivated entirely by enlightened self-interest. To the extent that the experimentalist can create an engaging, involving, and interesting task environment, many of the following issues may be ameliorated. True, most subjects remain motivated to participate in experiments because of the incentives offered, whether money, credit, or some other benefit. But it behooves any conscientious experimenter to keep in mind that many, if not most, subjects will want to try to figure out what the experiment is “really” about, and strive to discover what the experimenter wants, either to comply, resist, or simply as a matter of curiosity. The job of the experimenter is to make the task sufficiently absorbing that the subject finds it more interesting to concentrate on the task at hand than to try to game the experiment. One of the most effective and efficient ways to achieve this goal is to engage in pre-experimental pilot testing to see which stimuli or tasks elicit the most subject engagement. Particularly entrepreneurial experimenters can corral friends and relatives to test alternative scenarios to enhance subjects’ psychological involvement in a study. If an experimentalist invokes this strategy of pilot testing, it becomes absolutely imperative to solicit as much information from subjects in post-test debriefing in order to learn how they perceived the situation, how they understood their task, what systematic biases or misinterpretations might have emerged from the experimenter’s perspective, and how the procedure might be improved. Asking pilot subjects what they think might have increased their interest can prove a disarmingly straightforward and surprisingly successful strategy for enhancing future subject engagement.

Problems affecting prospects for internal validity arise when anything interferes with the ability to attribute changes in the dependent variables to manipulations in the independent variable. Sometimes, but not always, this can occur if subjects intentionally change their behavior to achieve a particular effect. Perhaps they want to give the experimenter what they think she wants, although they can easily be wrong about the experimenter's goals. Or they may want to intentionally thwart the investigator's purpose. Intentional subject manipulation of outcomes can even induce treatment values into the experiment artificially.

Most of these concern related to subjects strategically trying to manipulate their responses for reasons having nothing to do with the experimental treatment can be obviated by randomization, except to the extent that such attempts at deceiving the experimenter are either systematic or widespread in effect. Sometimes such efforts only affect the inferential process to the extent that subjects are able to successfully guess the investigator's hypotheses. Often it does not matter if the subject knows the purpose of an experiment; observers want to test straightforward conscious processes. However, if knowledge of the experimenter's hypothesis encourages subjects to consciously attempt to override their more natural instincts within the confines of the experiment, then the possibility of systematic interference with internal validity arises. Obviously, this is most problematic under conditions which require experimental deception. Since over 80 percent of psychology experiments in top journals utilize some kind of deception and almost no experiments in economics journals do, the issue of subjects guessing an experimenter's hypothesis remains more problematic in some disciplines than in others.

If this is a concern, one way to potentially control for such effects is to probe subjects after the experiment to see if they guessed deception was operative or discerned the true purpose of the experiment. Every effort should be made to keep subjects within the analysis, but skeptical

subjects can be compared separately with more susceptible subjects to determine any differences in response. If no differences exist, data can be collapsed; if differences emerge, they might be reported, especially if they suggest a bias in treatment effect resulting from individual differences in acceptance of the protocol. If there is no deception but the subject knows what the experiment is about, the problem of subjects intentionally trying to manipulate results is not eliminated; however, again, to the extent that such efforts are small and random, their should be minimized by processes of randomization across condition.

In addition, of course, any of these challenges to internal validity can be exacerbated to the extent that they occur concomitantly or interact in unexpected or unpredictable ways.

Ways to Improve

Many of the ways to circumvent challenges to internal validity have been alluded to in the course of the discussion above. Well designed experiments with strong control, careful design and systematic measurement go a long way toward alleviating many of these concerns.

Perhaps the single most important strategy experimenters can employ to avoid risks to internal validity is develop procedures to optimize experimental realism for subjects. Designing an experiment which engages subjects' attention and curiosity will ensure that the dynamic processes elicited in the experimental condition mimic those which are evoked under similar real world conditions. No amount of time that goes into trying to develop an involving experiment from the perspective of the subject will be wasted in terms of maximizing the resemblance between the psychological experience in the experiment and that of the unique real world environment they both inhabit and create.

2. External Validity

While psychologists pay primary attention to issues associated with internal validity, political scientists tend to focus, almost exclusively, on problems associated with external validity. External validity refers to the generalizability of findings from a study, or the extent to which conclusions can be applied across different populations or situations. Privileging of external validity often results from a misunderstanding that generalizability can result from, or be contained within, a single study, as long as it is large enough, or broad enough. This is almost never true. External validity results primarily from *replication* of particular experiments across diverse populations and different settings, using a variety of methods and measures. As Aronson et al. (1990) state succinctly: “No matter how similar or dissimilar the experimental context is to a real-life situation, it is still only one context: we cannot know how far the results will generalize to other contexts unless we carry on an integrated program of systematic replication” (77).

Some of the reason for the difference in disciplinary emphasis results from divergent purposes. Most psychologists, like many economists, use experiments primarily to test theory, rather than to make generalizations of such theory to broader populations. Their primary research goal focuses on explicating and elucidating basic operating principles underlying common human behaviors, such as cooperation or discrimination, and then distilling these processes through the crystalline filter of replication to delineate the boundaries of their manifestation and expression in real world contexts. Replication which establishes external validity can, and should, take many forms. If a genuine cause and effect relationship exists across variables, it should emerge over time, within different context, using various methods of measurement, and across population groups, or the boundaries of their operation should become defined (Smith and Mackie 1995). Aronson et al. describe this process best:

Bringing the research out of the laboratory does not necessarily make it more generalizable or “true”; it simply makes it different. The question of which method—“artificial” “laboratory experiments *versus* experiments conducted in the real world—will provide the more generalizable results is simply the wrong question. The generalizability of *any* research finding is limited. This limitation can be explicated only by systematically testing the robustness of research results across different empirical realizations of both the independent and dependent variables via systematic replication to test the extent to which different translations of abstract concepts into concrete realizations yield similar results (Aronson et al. 1990, 82).

Of course it remains important to examine the extent to which the outcomes measured in a laboratory setting find analogues in real world contexts. External validity can be examined in various ways, including measuring various treatment effects in real world environments, exploring the diverse context in which these variables emerges, investigating the various populations it affects, and looking at the way basic phenomena might change in response to different situations. Some of these factors can be explored in the context of a controlled laboratory setting, but some might be more profitably addressed in field contexts. However, experimenters should remain aware of the trade-offs involved in the ability to control and measure carefully defined variables with a richer understanding of the extent to which these factors might interact with other unknowns outside the laboratory setting.

While the concerns regarding external validity certainly remain legitimate, it is important to keep it mind that they should only arise to the extent that sufficient prior attention has been paid to assuring that a study embodies internal validity first. As Aronson et al. (1990), rightly state: “internal validity is, of course, the more important, for if random or systematic error makes it impossible for the experimenter even to draw any conclusions from the experiment, the question of the generality of these conclusions never arises” (75).

Threats to External Validity

Most concerns that political scientists express regarding external validity reflect their recognition of the artificial nature of the laboratory setting. The notion here is that the trivial tasks presented to subjects offer a poor analogue to the real world experiences that individuals confront in trying to traverse their daily political and social environments. This characterization of a controlled laboratory experiment, while often accurate, reflects a privileging of mundane as opposed to experimental realism. The benefit of such a stripped-down stylized setting is that it offers the opportunity to carefully operationalize and measure the variables of interest, and then, through multiple tests on numerous populations, begin to define the conditions under which generality might obtain. The reason it becomes so critical to uncover these mechanisms is because unless an investigator knows the underlying principles operating in a given dynamic, it will prove simply impossible to ascertain which aspect of behavior is causing which effect within the context of real world settings where many other variables and interactions occur simultaneously. One of the most dramatic examples of this process occurred in the famous Milgram (1974) experiment; Milgram set out to explain the compliance of ordinary Germans with Nazi extermination of the Jews. Testing at Yale was designed to provide a control condition for later comparison with German and Japanese subjects. Prior to the experiment, every psychiatrist consulted predicted that only the worst, most rare psychopaths would administer the maximum amount of shock. But the careful design of the experiment allowed Milgram to begin to uncover the subtle and powerful effects of obedience on behavior.

Experimental realism remains more important than mundane realism in maximizing prospects for internal validity because it is more likely to elicit the critical dynamic under investigation; more highly stylized or abstract experimental protocols can risk both internal and external validity by failing to engage subjects' attention or interest. Creative design can

sometimes overcome these difficulties, even in laboratory settings, as clever experimentalists have found ways to simulate disrespect, for example, by having a confederate “accidentally” bump into a subject rudely (Cohen and Nisbett 1994) or simulating an injury (Darley and Latane 1968), for example.

Restricted subject populations can also limit the degree of potential generalizability from studies as well, although the degree to which this problem poses a serious threat varies with the topic under investigation. While in general it is better to have more subjects across a wider demographic range, depending on the content of study, it may be more important to obtain more subjects, rather than explicitly diverse ones. Common sense, in combination with practical logistics such as costs, should guide judgment concerning how best to orchestrate this balance.

Several other threats to external validity exist, some of which mirror those which can compromise internal validity. Subject mortality raises a concern for external validity to the extent that such mortality takes place prior to randomization; recall that mortality subsequent to randomization compromises internal validity. Prior to randomization, subject mortality may compromise the representativeness of the study population.

Selection bias, in terms of nonrandom sampling, represents another threat to external validity which also threatens internal validity as described above. If subjects are drawn from too restrictive a sample, or an unrepresentative sample, then obviously more replication will be required in order to generalize the results with confidence. This is becoming an increasing concern with the huge increase in Internet samples, where investigator knowledge and control of their subject populations can become extremely restricted. It can be virtually impossible to know whether the person completing the Internet survey is who they say they are, much less whether they are attending to the tasks in any meaningful way. With unrepresentative samples of

students, it is still possible to reason about the ways in which they may not be representative of a larger population, by having superior abstract cognitive skills for example. But with an Internet sample, even this ability to determine the ways in which one sample may differ from another becomes extremely challenging.

The so-called *Hawthorne effect* poses another threat to external validity. This phenomenon, named after the man who precipitated its effect when it was first recognized, refers to the way in which people change their behavior simply because they know they are being monitored (Roethlisberger and Dickson 1939). Surveillance alone can change behavior in ways which can influence the variable being measured. Without such monitoring and observation, behavior, and thus results, might appear quite different. Sometimes this effect can be desirable, such as when observation is used to enforce compliance with particular protocols, reminding subjects to do a certain thing in a particular way or at a specific time. Technology such as personal handheld devices can help facilitate this process as well. But other times such self consciousness can affect outcome measures in more biased ways.

External interference, or interference between units, presents a complex and nuanced potential confound (see Sinclair's chapter in this volume). Behavior in the real world operates differently than in more sanitized settings precisely because there is more going on at any given time and, often, more is at stake in the psychic world of the subject. Moreover, the same variables may not operate the same way in two different situations precisely because other factors may exacerbate or ameliorate the appearance of any given response within a particular situation. Interference thus can occur not only as a result of what happens within a given experimental context, but also as a consequence of the way such responses can change when

interacting with diverse and unpredictable additional variables in real world contexts which can operate to suppress, potentiate, or otherwise overwhelm the expression of the relevant processes.

Perhaps the biggest concern that political scientists focus on with regard to external validity revolves around issues related to either the artificiality or triviality of the experimental situation, although certainly it is possible for the opposite criticism, that subjects pay too much attention to stimulus materials in experiments, to be leveled as well. For example, in looking at the effect of television advertising, subjects may pay much closer attention to such ads in the lab than they would in real life, where they might be much more likely to change channels or walk out of the room when the ads come on. Clearly, it would be next to impossible for experimenters to replicate most aspects of real life in a controlled way. Time constraints, cultural norms, and subject investment will preclude such mirroring (Walker 1976). However, it is often the case that such cloning is not necessary in order to study a particular aspect of human behavior, and the ability to isolate such phenomena, and explore its dimensions, can compensate for the more constrained environmental setting by allowing investigators to delineate the precise microfoundational mechanisms underlying particular attitudes and behaviors of interest. The benefits that can derive from locating such specificity in the operation of the variable under investigation can make the ostensible artificiality worthwhile.

Ways to Improve

Given that political scientists tend to be united in their concern for politically relevant contexts (Druckman and Lupia 2006), external validity will continue to serve as a central focus of concern for those interested in experimental relevance for broader societal contexts. Several strategies can help maximize the potential for increasing such relevance and broader applicability, first and foremost being reliance on replication across subjects, time, and situation.

In general, anything that multiplies the ways in which a particular dynamic is investigated can facilitate prospects for external validity. To be clear, external validity occurs primarily as a function of this strategy of systematic replication. Conducting a series of experiments which include different populations, involve different situations, and utilize multiple measurements establishes the fundamental basis of external validity. A single study, no matter how many subjects it encompasses, or how realistic the environment, cannot alone justify generalization outside the population and domain in which it was conducted.

One of the most important ways to enhance external validity involves increasing the heterogeneity of the study populations, unless of course one is only trying to generalize to homogenous population, such as veterans, or republicans, or women, in which case the study of focal populations remains optimal. Including subjects from different age groups, sexes, various races, and diverse socio-economic or educational statuses, for example, increases the representativeness of the sample, and potentiates prospects for generalization. Again, common sense should serve as a guide as to which populations should be studied for any given topic. Studies involving facial recognition of emotion, for example, can benefit greatly from employing subjects with focal brain lesions because their deficits in recognition can inform researchers as to the processes necessary for intact processing of these attributes. In this case, fewer subjects with particularly illuminating characteristics can provide greater leverage than a larger number of less informative ones.

Increasing the diversity of circumstances or situations in which a particular phenomenon is investigated can also heighten external validity. Exploring a particular process, such as cooperation, in a variety of settings can prove particularly helpful for discovering contextual

boundaries on particular processes, locating environmental cues which trigger such dynamics, and illustrating the particular dimensions of its operation.

Using multiple measures, or multiple types of measures, as well as doing everything possible to improve the quality of measures employed, can greatly enhance external validity as well. This might involve finding multiple dependent measures to assess downstream effects, either over time or across space (Green and Gerber 2002). In many medical studies, intervening variables are often used as proxies to determine intermediary effects if the critical outcome variable does not occur frequently, or takes a long time to manifest, although some problems associated with this technique were noted above as well. Proper and careful definition of the variables under consideration, both those explicitly being studied and measured, as well as those expected to impact these variables differentially in a real world setting, remains crucial to isolating the conditions under which particular processes are predicted to occur.

3. Balance between Internal and External Validity

Obviously it goes without saying that it is best to strive to maximize both internal and external validity. But sometimes this is not possible within the practical and logistical constraints of a given experimental paradigm. Maximizing internal validity may diminish the ability to extrapolate the findings to situations and populations outside those specifically studied. Privileging external validity often neglects important aspects of internal experimental control so that the true cause of reported findings remains unclear. It remains important to explicitly and clearly recognize the inherent nature of the trade-offs between them.

Two principle trade-offs exist between internal and external validity. First, the balance between these types of validity clearly reflects a difference in value. Attention to internal validity optimizes the ability of an investigator to achieve confidence that changes in the dependent

variables truly resulted from the manipulation of the independent variable. In other words, methodological and theoretical clarity emerge from careful and conscientious documentation of variables and measures. The experimenter can rest assured that the processes investigated returned the results produced; in other words, investigators can believe that they studied what they intended, and that any effect was produced by the manipulated purported cause.

On the other hand, concentration on external validity by expanding subject size or representativeness can increase confidence in generalizability, but only to the extent that extraneous or confounding hypotheses can be eliminated or excluded from contention. If sufficient attention has gone into securing details assuring internal validity, then the window between the laboratory and the outside world can become more transparent.

Second, trade-offs between internal and external validity exist in practical terms as well. Internal validity can take time and attention to detail in operationalizing variables, comparing measures and contrasting the implications of various hypotheses. Most of this effort takes place prior to actually conducting the experiment. Working toward enhancing external validity requires more enduring effort, since by definition the effort must sustain beyond a single study and encompass a sequence of experiments. In addition, securing additional populations or venues may take time after the experiment is designed.

Such trade-offs between internal and external validity emerge inevitably over the course of experimental work. Depending on topic, a given experimenter may concentrate on maximizing one concern over the other within the context of any particular study. But awareness of the requisite trade-offs in value and practice can be important in balancing the intent and implementation of any given experiment.

4. Future Work

Striving to maximize internal and external validity in every experiment remains a laudable goal, even if optimizing both can sometimes prove unrealistic in any particular study. Certain things can be done to improve and increase validity in general depending on the substance of investigation.

Particular methodological techniques or technologies may make it easier to enhance validity in part by making measures less obtrusive, allowing fewer channels for subjects to consciously accommodate or resist experimental manipulation. These can become particularly useful when studying socially sensitive topics such as race or sex. Such strategies include implicit association tests (IAT) and other implicit measures of attention (see Lodge and Taber's chapter in this volume). Instruments such as these can allow investigators to obtain responses from subjects without the subjects necessarily being aware of either the topic, or being able to control their reactions consciously. Similarly, reaction time tests can also provide measures of speed and accuracy of association in ways that can bypass subjects' conscious attempts to deceive or manipulate observers. While some subjects may be able to consciously slow their response to certain stimuli, although it is unclear why they would choose to do so, it may be impossible for them to perform more rapidly than their inherent capacity allows.

Of course, other forms of subliminal tests exist, although they tend to be less widely known or used in political science than in other fields such as psychology or neuroscience. Neuroscientists for example often use eye tracking devices in order to follow what a subjects observes without having to rely on less accurate self-report measures. Physiological measures, including heart rate, galvanic skin response, or eye blink function can also be employed for this purpose. Clearly, the most common technology at the moment in both psychology and neuroscience involve functional magnetic resonance imagery to locate particular geographies in

the brain. Blood and saliva tests for hormonal or genetic analysis can also provide useful and effective, if more intrusive, indirect measures of human functioning. I have leveraged these measures in my own work exploring the genetic basis of aggression (McDermott et al. 2009) and sex differences in aggression (McDermott and Cowden 2001); in this way, such biological measures can be used to explore some factors underlying conflict. These technologies offer the advantage of circumventing notoriously unreliable or deceptive self-report to obtain responses which can be compared either within, or between, subjects in determining potential sources for particular attitudes and behaviors of interest. Such efforts can enhance prospects for internal validity and increase the ease and speed with which external validity can be achieved as well.

References

- Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-55.
- Arceneaux, Kevin, Alan Gerber, and Donald Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 37-62.
- Aronson, Elliott, Phoebe C. Ellsworth, J. Merrill Carlsmith, and Marti Hope Gonzales. 1990. *Methods of Research in Social Psychology*, 2nd Ed. New York: McGraw-Hill.
- Campbell, Donald T. 1957. "Factors Relevant to Validity of Experiments in Social Settings." *Psychological Bulletin* 54: 297-312.
- Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.
- Cohen, Dov, and Richard E. Nisbett. 1994. "Self-Protection and the Culture of Honor: Explaining Southern Violence." *Personality and Social Psychology Bulletin* 20: 551-67.
- Darley, John M., and Bibb Latane. 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility." *Journal of Personality and Social Psychology* 8: 377-83.
- Druckman, James N., Donald P. Green, James H. Kuklinski., and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100: 627-35.

- Druckman, James N., and Arthur Lupia. 2006. "Mind, Will and Choice." In *The Oxford Handbook on Contextual Political Analysis*, eds. Charles Tilly, and Robert E. Goodin. Oxford: Oxford University Press.
- de Quervain, Dominique J.-F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. "The Neural Basis of Altruistic Punishment." *Science* 27 305(5688): 1254-8
- Fehr, Ernst, and Urs Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425: 785-91.
- Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137-40.
- Gerber, Alan, and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653-63
- Gigerenzer, Gerd. 1996. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103: 592-6.
- Green, Donald P., and Alan Gerber. 2002. "The Downstream Benefits of Experimentation." *Political Analysis* 10: 394-402.
- Kiesler, Charles A., Barry E. Collins, and Norman Miller. 1969. *Attitude Change: A Critical Analysis of Theoretical Approaches*. New York: Wiley.
- McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10: 325-42.
- McDermott, Rose, and Jonathan A. Cowden. 2001. "The Effects of Uncertainty and Sex in a Simulated Crisis Game." *International Interactions* 27: 353-80.
- McDermott, Rose, Dustin Tingley, Jonathan A. Cowden, Giovanni Frazzetto, and Dominic D. P. Johnson. 2009. "Monoamine Oxidase A Gene (MAOA) Predicts Behavioral Aggression Following Provocation." *Proceedings of the National Academy of Sciences* 106: 2118-23.
- Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variable Approach." *Journal of Political Economy* 112: 725-53.
- Milgram, Stanley. 1974. *Obedience to Authority*. New York: Harper & Row.
- Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233-52
- Roethlisberger, Fritz J., and William J. Dickson. 1939. *Management and the Worker*. Cambridge, MA: Harvard University Press.

- Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *Handbook of Experimental Economics*, eds. J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.
- Smith, Eliot R., and Diane M. Mackie. 1995. *Social Psychology*. New York: Worth Publishers.
- Sokol-Hessner, Peter, Ming Hsu, Nina G. Curley, Mauricio R. Delgado, Colin F. Camerer, and Elizabeth A. Phelps. 2009. "Thinking Like a Trader Selectively Reduces Individuals' Loss Aversion." *Proceedings of the National Academy of Sciences* 106: 5035-40.
- Sovey, Allison J., and Donald P. Green. 2010. "Instrumental Variables Estimation in Political Science: A Reader's Guide." Unpublished manuscript, Yale University.
- Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185: 1124-31.
- Walker, T. 1976. "Microanalytic Approaches to Political Decision Making." *American Behavioral Science* 20: 93-110.

4. Students as Experimental Participants: A Defense of the “Narrow Data Base”

James N. Druckman and Cindy D. Kamⁱ

An experiment entails randomly assigning participants to various conditions or manipulations. Given common consent requirements, this means experimenters need to recruit participants who, in essence, agree to be manipulated. The ensuing practical and ethical challenges of subject recruitment have led many researchers to rely on convenience samples of college students. For political scientists who put particular emphasis on generalizability, the use of student participants often constitutes a critical, and according to some reviewers, fatal problem for experimental studies.

In this chapter, we investigate the extent to which using students as experimental participants creates problems for causal inference. First, we discuss the impact of student subjects on a study’s internal and external validity. In contrast to common claims, we argue that student subjects do *not* intrinsically pose a problem for a study’s external validity. Second, we use simulations to identify situations when student subjects are likely to constrain experimental inferences. We show that such situations are relatively limited; any convenience sample poses a problem *only* when the size of an experimental treatment effect depends upon a characteristic on which the convenience sample has virtually no variance. Third, we briefly survey empirical evidence that provides guidance on when researchers should be particularly attuned to taking steps to ensure appropriate generalizability from student subjects. We conclude with a discussion of the practical implications of our findings. In short, we argue that student subjects are not an inherent problem to experimental research; moreover, the burden of proof—of student subjects being a problem—should lie with critics rather than experimenters.

1. The “Problem” of Using Student Subjects

Although internal validity may be the “sine qua non” of experiments, most researchers use experiments to make *generalizable* causal inferences (Shadish et al. 2002, 18-20). For example, suppose one implements a laboratory study with students and finds a causal connection between the experimental treatment (e.g., a media story about welfare) and an outcome of interest (e.g., support for welfare). An obvious question is whether the relationship found in the study exists within a heterogeneous population, in various contexts (e.g., a large media marketplace), over time. This is an issue of external validity, which refers to the extent to which the “causal relationship holds over variations in persons, settings, treatments [and timing], and outcomes” (Shadish et al. 2002, 83). McDermott (2002) explains that “External validity... tend[s] to preoccupy critics of experiments. This near obsession... tend[s] to be used to dismiss experiments” (334).

A point of particular concern involves generalization from the sample of experimental participants – especially when, as is often the case, the sample consists of students – to a larger population of interest. Indeed, this was the focus of Sears’ (1986) widely cited article, “College Sophomores in the Laboratory: Influences of a Narrow Data base on Social Psychology’s View of Human Nature.”ⁱⁱ Many political scientists employ “the simplistic heuristic of ‘a student sample lacks external generalizability’” (Kam et al. 2007, 421). Gerber and Green (2008) explain that “If one seeks to understand how the general public responds to social cues or political communication, the external validity of lab studies of undergraduates has inspired skepticism (Sears 1986; Benz and Meier 2008)” (358). In short, social scientists in general and political scientists in particular view student subjects as a major hindrance to drawing inferences from experimental studies.

Assessing the downside of using student subjects has particular current relevance. First, many political science experiments use student subjects; for example, Kam et al. report that from 1990 through 2006, a quarter of experimental articles in general political science journals relied on student subjects while over seventy percent did so in more specialized journals (Kam et al. 2007, 419-420); see also Druckman et al. 2006). Are the results from these studies of questionable validity? Second, there are practical issues. A common rationale for moving away from laboratory studies, in which student subjects are relatively common, to survey and/or field experiments is that these latter venues facilitate using nonstudent participants. When evaluating the pros and cons of laboratory versus survey or field experiments, should substantial weight be given to whether participants are students? Similarly, those implementing lab experiments have increasingly put forth efforts (and paid costs) to recruit nonstudent subjects (e.g., Lau and Redlawsk 2006, 65-66; Kam 2007). Are these costs worthwhile? To address these questions, we next turn to a broader discussion of what external validity demands.

The Dimensions of External Validity

To assess the external validity or generalizability of a causal inference, one must consider *from what* we are generalizing and *to what* we hope to generalize. When it comes to “from what,” a critical, albeit often neglected, point is that external validity is best understood as being assessed over a range of studies on a single topic (McDermott 2002, 335). Assessment of any single study, regardless of the nature of its participants, must be done in light of the larger research agenda to which it hopes to contribute.ⁱⁱⁱ

Moreover, when it comes to generalization from a series of studies, the goal is to generalize across *multiple* dimensions. External validity refers to generalization not only of individuals but also across settings/contexts, times, and operationalizations. There is little doubt

that institutional and social contexts play a critical role in determining political behavior and that, consequently, they can moderate causal relationships. One recent powerful example comes from the political communication literature; a number of experiments, using both student *and* nonstudent subjects, show that when exposed to political communications (e.g., in a laboratory), individuals' opinions often reflect the content of those communications (e.g., Kinder 1998; Chong and Druckman 2007b). The bulk of this work, however, ignores the contextual reality that people outside of the controlled study setting have choices (i.e., they are not captive). Arceneaux and Johnson (2008) show that as soon as participants in communication experiments can choose whether to receive a communication (i.e., the captive audience constraint is removed), results about the effects of communications drastically change (i.e., the effects become less dramatic). In this case, ignoring the contextual reality of choice appears to have constituted a much greater threat to external validity than the nature of the subjects.^{iv}

Timing also matters. Results from experiments implemented at one time may not hold at other times given the nature of world events. Gaines, Kuklinski, and Quirk (2007) further argue that survey experiments in particular may misestimate effects due to a failure to consider what happened prior to the study (also see Gaines and Kuklinski's chapter in this volume). Building on this insight, Druckman (2009) asked survey respondents for their opinions about a publicly owned gambling casino, which was a topic of "real world" ongoing political debate. Prior to expressing their opinions, respondents randomly received no information (i.e., control group) or information that emphasized either economic benefits or social costs (e.g., addiction to gambling). Druckman shows that the opinions of attentive respondents (i.e., respondents who regularly read newspaper coverage of the campaign) in the economic information condition did not significantly differ from attentive individuals in the control group.^v The non-effect likely

stemmed from the economic information – which was available outside the experiment in ongoing political discussion – having already influenced all respondents. Another exposure to this information in the experiment did not add to the prior, pre-treatment effect. In other words, the ostensible non-effect lacked external validity – not because of the sample – but because it failed to account for the timing of the treatment and what had occurred prior to that time (also see Slothuus 2009).^{vi}

A final dimension of external validity involves how concepts are employed. Finding support for a proposition means looking for different ways of administering and operationalizing the treatment (e.g., delivering political information via television ads, newspaper stories, interpersonal communications, survey question text) and operationalizing the dependent variables (e.g., behavioral, attitudinal, physiological, implicit responses).

In short, external validity does *not* simply refer to whether a specific study, if re-run on a different sample, would provide the same results. It refers more generally to whether “conceptually equivalent” (Anderson and Bushman 1997) relationships can be detected across people, places, times, and operationalizations. This introduces the other end of the generalizability relationship – that is, “equivalent” to what? For many, the “to what” refers to behavior as observed outside of the study, but this is not always the case. Experiments have different purposes; Roth (1995) identifies three non-exclusive roles that experiments can play: “search for facts,” “speaking to theorists,” or “whispering in the ears of princes,” (22) which facilitates “the dialogue between experimenters and policymakers” (see also Guala 2005, 141-160). These types likely differ in the target of generalization. Of particular relevance is that theory-oriented experiments typically are not meant to “match” behaviors observed outside the study *per se*, but rather the key is to generalize to the precise parameters put forth in the given

theory. Plott (1991) explains that “The experiment should be judged by the lessons it teaches about the theory and not by its similarity with what nature might have happened to have created” (906). This echoes Mook’s (1983) argument that much experimental work is aimed at developing and/or testing a theory, not at establishing generalizability. Even experiments that are designed to demonstrate “what can happen” (e.g., Milgram 1963, Zimbardo 1973) can still be useful, even if they do not mimic everyday life.^{vii} In many of these instances, the nature of the subjects in the experiments are of minimal relevance, particularly given experimental efforts to ensure their preferences and/or motivations match those in the theory (e.g., see Dickson’s chapter in this volume).

Assessment of how student subjects influence external validity depends on three considerations: (1) the research agenda on which the study builds (e.g., has prior work already established relationship with student subjects, meaning incorporating other populations may be more pressing?), (2) the relative generalizability of the subjects, compared to the setting, timing, and operationalizations (e.g., a study using students may have more leeway to control these other dimensions), and (3) the goal of the study (e.g., to build a theory or to generalize one).

Evaluating External Validity

The next question is how to evaluate external validity. While this is best done over a series of studies, we acknowledge the need to assess the strengths of a particular study with respect to external validity. Individual studies can be evaluated in at least two ways (Aronson and Carlsmith 1968; Aronson, Brewer, and Carlsmith 1998). First, experimental realism refers to whether “an experiment is realistic, if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has impact on them” (Aronson et al. 1985, 485). Second, mundane realism concerns “the extent to which events occurring in the research setting are likely to occur

in the normal course of the subjects' lives, that is, in the 'real world'" (Aronson et al. 1985, 485).^{viii}

Much debate about samples focuses on mundane realism. When student subjects do not match the population to which a causal inference is intended, many conclude that the study has low external validity. Emphasis on mundane realism, however, is misplaced (e.g., McDermott 2002; Morton and Williams 2008, 345); of much greater importance is experimental realism. Failure of participants to take the study and treatments "seriously" compromises internal validity, which in turn, renders external validity of the causal relationship meaningless (e.g., Dickhaut et al. 1972, 477; Liyanarachchi 2007, 56).^{ix} In contrast, at worst, low levels of mundane realism simply constrain the breadth of any generalization but do not make the study useless.

Moreover, scholars have yet to specify clear criteria for assessing mundane realism, and, as Liyanarachchi (2007) explains, "any superficial appearance of reality (e.g., a high level of mundane realism) is of little comfort, because the issue is whether the experiment 'captures the intended essence of the theoretical variables' (Kruglanski 1975, 106)" (57).^x That said, beyond superficiality, we recognize student subjects – while having no ostensibly relevant connection with experimental realism^{xi} – may limit mundane realism that constrains generalizations of a particular study. This occurs when characteristics of the subjects affect the nature of the causal relationship being generalized. When this occurs, and with what consequences, are questions to which we now turn.

2. Statistical Framework

In this section, we examine the use of student samples from a statistical point of view. This allows us to specify the conditions under which student samples might constrain causal generalization (in the case of a single experiment). Our focus, as in most political science

analyses of experimental data, is on the magnitude of the effect of some experimental treatment, T , on an attitudinal or behavioral dependent measure, y . Suppose, strictly for presentational purposes, we are interested in the effect of a persuasive communication (T) on a subject's post-stimulus policy opinion (y) (we could use virtually any example from any field). T takes on a value of 0 for subjects randomly assigned to the control group and takes on a value of 1 for subjects randomly assigned to the treatment group.^{xii} Suppose the true data generating process features a homogeneous treatment effect:

$$y_i = \beta_0 + \beta_T T_i + \varepsilon_i \quad (1)$$

Assuming that all assumptions of the classical linear regression model are met, the OLS estimate for β_T is unbiased, consistent, and efficient.^{xiii} The results derived from estimation on a given sample would be fully generalizable to those that would result from estimation on any other sample.

Specific samples will differ in their distributions of individual covariates. Continuing with our running example, samples may differ in the distribution of *attitude crystallization* (i.e., an attitude is increasingly crystallized when it is stronger and more stable).^{xiv} Student samples may yield a disproportionately large group of subjects that are low in crystallization. A random sample from the general population might generate a group that is normally distributed and centered at the middle of the range. A sample from politically active individuals (such as conventioners) might result in a group that is disproportionately high in crystallization.^{xv}

Consider the following samples with varying distributions on attitude crystallization. In all cases, $N=200$ and treatment is randomly assigned to half of the cases. Attitude crystallization ranges from 0 (low) to 1 (high). Consider a "Student Sample" where ninety percent of the sample is at a value of "0" and ten percent of the sample is at a value of "1". Consider a "Random

Sample” where the sample is normally distributed and centered on 0.5 with standard deviation of 0.165. Finally, consider a “Conventioneers Sample” where ten percent of the sample is at a value of “0” and ninety percent of the sample is at a value of “1”.^{xvi}

Suppose the true treatment effect (β_T) takes a value of “4”. We set up a Monte Carlo experiment that estimated Equation [1] 1,000 times, each time drawing a new ε term. We repeated this process for each of the three types of samples (student, random, and conventioneers). The sampling distributions for b_T appear in Figure 4-1.

[Figure 4-1 about here]

The results demonstrate that when the true data generating process produces a single treatment effect, estimates from *any* sample will produce an unbiased estimate of the true underlying treatment effect. Perhaps this point seems obvious, but we believe it has escaped notice from many who criticize experiments that rely on student samples. We repeat: *If the underlying data generating process is characterized by a homogeneous treatment effect (i.e., the treatment effect is the same across the entire population), then any convenience sample should produce an unbiased estimate of that single treatment effect, and, thus, the results from any convenience sample should generalize easily to any other group.*

Suppose, however, the “true” underlying data generating process contains a heterogeneous treatment effect: that is, the effect of the treatment is moderated^{xvii} by individual-level characteristics. The size of the treatment effect might depend upon some characteristic, such as gender, race, age, education, sophistication, etc. Another way to say this is that there may be an “interaction of causal relationship with units” (Shadish et al. 2002, 87).

As one method of overcoming this issue, a researcher can randomly sample experimental subjects. By doing so, the researcher can be assured that:

the average causal relationship observed in the sample will be the same as (1) the average causal relationship that would have been observed in any other random sample of persons of the same size from the same population and (2) the average causal relationship that would be observed across *all* other persons in that population who were not in the original random sample (Shadish et al. 2002, 91).

Although random sampling has advantages for external validity, Shadish et al. (2002) note that “it is so rarely feasible in experiments” (91). The way to move to random sampling might be to use survey experiments, where respondents are (more or less) a random sample of some population of interest. We will say a bit more about this possibility below. For now, let us assume that a given researcher has a specific set of reasons for not using a random sample (cost, instrumentation, desire for laboratory control, etc.), and let’s examine the challenges a researcher using a convenience sample might face in this framework.

We revise our data generating process to reflect a heterogeneous treatment effect by taking Equation (1) and modeling how some individual-level characteristic, Z (e.g., attitude crystallization), influences the magnitude of the treatment effect:

$$\beta_1 = \gamma_{10} + \gamma_{11}Z_i \quad (2)$$

We also theorize that Z might influence the intercept:

$$\beta_0 = \gamma_{00} + \gamma_{01}Z_i$$

Substituting into (1):

$$y_i = (\gamma_{00} + \gamma_{01}Z_i) + (\gamma_{10} + \gamma_{11}Z_i)T_i + \epsilon_i$$

$$y_i = \gamma_{00} + \gamma_{01}Z_i + \gamma_{10}T_i + \gamma_{11}Z_i * T_i + \epsilon_i \quad (3)$$

If our sample includes sufficient variance on this moderator, and we have *ex ante* theorized that the treatment effect depends upon this moderating variable, Z , then we can (and should) *estimate* the interaction. If, however, the sample does not contain sufficient variance, not only can we not

identify the moderating effect, but we may misestimate the on-average effect—depending on what specific range of Z is present in our sample.

The question of generalizing treatment effects reduces to asking if there is a single treatment effect or a set of treatment effects, the size of which depends upon some (set of) covariate(s). Note that this is a *theoretically oriented* question of generalization. It is not just whether “student samples are generalizable” but rather, what particular characteristics of student samples might lead us to wonder whether the causal relationship detected in a student sample experiment would be systematically different from the causal relationship in the general population.

Revisiting our running example, suppose we believe that a subject’s level of attitude crystallization (Z) influences the effect of a persuasive communication (T) on a subject’s post-stimulus policy opinion (y). The more crystallized someone’s attitude is, the smaller the treatment effect should be. The less crystallized someone’s attitude is, the greater the treatment effect should be. Using this running example, based on equation (3), assume that the true relationship has the following (arbitrarily selected) values:

$$\gamma_{00} = 0$$

$$\gamma_{01} = 0$$

$$\gamma_{10} = 5$$

$$\gamma_{11} = -5$$

Let Z , attitude crystallization, range from 0 (least crystallized) to 1 (most crystallized).

γ_{10} tells us the effect of the treatment when $Z=0$, that is, the treatment effect among the least crystallized subjects. γ_{11} tells us how crystallization moderates the effect of the treatment.

First, consider what happens when we estimate (1), the simple (but theoretically incorrect, given it fails to model the moderating effect) model that looks for the “average” treatment effect. We estimated this model 1,000 times, each time drawing a new ϵ term. We repeated this process for each of the three samples. The results appear in Figure 4-2.

[Figure 4-2 about here]

When we estimate a “simple” model, looking for an average treatment effect, our estimates for β_1 diverge from sample to sample. In cases where we have a student sample, and where low levels of crystallization increase the treatment effect, we systematically overestimate the treatment effect relative to what we would get in estimating the same model on a random sample with moderate levels of crystallization. In the case of a Conventioneer Sample, where high levels of crystallization depress the treatment effect, we systematically underestimate the treatment effect, relative to the estimates obtained from the general population.

We have obtained three different results across the samples because we have estimated a model based on Equation (1). Equation (1) should only be estimated when the data generating process produces a *single* treatment effect: the value of β_1 . However, we have “mistakenly” estimated Equation (1) when the true data generating process produces a series of treatment effects (governed by the function: $\beta_1 = 5 - 5Z_i$). The sampling distributions in Figure 4-2 provide the “average” treatment effect, which depends directly upon the mean value of Z within a given sample: $5 - 5 * E(Z)$.

Are the results from one sample more trustworthy than the results from another sample? As Shadish et al (2002) note, conducting an experiment on a random sample will produce an “average” treatment effect; hence, to some degree the results from the Random Sample might be more desirable than the results from the other two convenience samples. However, all three sets

of results reflect a fundamental disjuncture between the model that is estimated and the true data generating process. If we have a theoretical reason to believe that the data generating process is more complex (i.e., the treatment depends on an individual level moderator), then we should embed this *theoretical model* into our *statistical model*.

To do so, we returned to Equation (3) and estimated the model 1,000 times, each time drawing a new ϵ term. We repeated this process three times, for each of the three samples. The results appear in Figure 4-3.

[Figure 4-3 about here]

First, notice that the sampling distributions for b_T are all centered on the same value: 5, and the sampling distributions for b_{TZ} are also all centered on the same value: -5. In other words, Equation (3) produces *unbiased point estimates* for β_T and β_{TZ} , regardless of which sample is used. We uncover unbiased point estimates even where only 10% of the sample provides key variation on Z (Student Sample and Conventioneers Sample).

Next, notice the spread of the sampling distributions. We have the most certainty about b_T in the Student Sample and substantially less certainty in the Random Sample and the Conventioneers Sample. The greater degree of certainty in the Student Sample results from the greater mass of the sample that is located at 0 in the Student Sample (since the point estimate for β_T , the un-interacted term in Equation (3), represents the effect of T when Z happens to take on the value of 0).

For the sampling distribution of b_{TZ} , we have higher degrees of certainty (smaller standard errors) in the Student Sample and the Conventioneers Sample. This is an interesting result. By using samples that have higher variation on Z , we yield more precise point estimates

of the heterogeneous treatment effect.^{xviii} Moreover, we are still able to uncover the interactive treatment effect, since these samples still contain some variation across values of Z .

How much variation in Z is sufficient? So long as Z varies to any degree in the sample, the estimates for b_T and b_{TZ} will be *unbiased*. Being “right on average” may be of little comfort if the degree of uncertainty around the point estimate is large. If Z does not vary very much in a given sample, then the estimated standard error for b_{TZ} will be large. But concerns about uncertainty are run-of-the-mill when estimating a model on any dataset: more precise estimates arise from analyzing datasets that maximize variation in our independent variables.

Our discussion thus suggests that experimentalists (and their critics) need to consider the underlying data generating process: that is, *theory* is important. If a single treatment effect is theorized, then testing for a single treatment effect is appropriate. If a heterogeneous treatment effect is theorized, then researchers should explicitly theorize how the treatment effect should vary along a specific (set of) covariate(s), and researchers can thereby estimate such relationships so long as there is sufficient variation in the specific (set of) covariate(s) in the sample. We hope to push those who launch vague criticisms regarding the generalizability of student samples to offer more constructive, more theoretically oriented critiques that reflect the possibility that student samples may be problematic if the magnitude and direction of the treatment effect depends upon a particular (set of) covariate(s) that are peculiarly distributed within a student sample.

In sum, we have identified three distinct situations. First, in the homogeneous case – where the data generating process produces a single treatment effect – we showed the estimated treatment effect derived from a student sample is an unbiased estimate of the true treatment effect. Second, when there is a heterogeneous case (where the treatment effect is moderated by

some covariate Z) and the researcher fails to recognize the contingent effect, a student sample (indeed, any convenience sample) may misestimate the average treatment effect if the sample is non-representative on the particular covariate Z . However, in this case, even a representative sample would mis-specify the treatment effect due to a failure to model the interaction. Third, when the researcher appropriately models the heterogeneity with an interaction, then the student sample, even if it is non-representative on the covariate Z , will misestimate the effect only if there is virtually no variance (i.e., literally almost none) on the moderating dynamic. Moreover, a researcher can empirically assess the degree of variance on the moderator within a given sample, and/or use simulations to evaluate whether limited variance poses a problem for uncovering the interactive effect. An implication is that the burden, to some extent, falls on an experiment's critic to identify the moderating factor and demonstrate it lacks variance in an experiment's sample.

3. Contrasting Student Samples with Other Samples

We have argued that a given sample constitutes only one – and arguably not the critical one – of many considerations when it comes to assessing external validity. Further, a student sample only creates a problem when a researcher: 1) fails to model a contingent causal effect (when there is an underlying heterogeneous treatment effect), and 2) the students differ from the target population with regard to the distribution of the moderating variable. This situation, which we acknowledge does occur with non-trivial frequency, leads to the question of just how often student subjects empirically differ from representative samples. The greater such differences, the more likely problematic inferences will occur.

Kam (2005) offers some telling evidence comparing student and nonstudent samples on two variables that can affect information processing: political awareness and need for cognition.

She collected data from a student sample using the exact same items as are used in the American National Election Study's (ANES) representative sample of adult citizens. She finds the distributions for both variables in the student sample closely resemble those in the 2000 ANES. This near identical match in distribution, then, allowed Kam (2005) to more broadly generalize results from an experiment, on party cues, she ran with the student subjects.

Kam focuses on awareness and need for cognition because these variables plausibly moderate the impact of party cues—as explained, in comparing student and nonstudent samples, one should focus on possible differences that *are relevant to the study in question*. Of course, one may nonetheless wonder whether students differ in others ways that *could* matter (e.g., Sears 1986, 520). This requires a more general comparison, which we undertake by turning to the 2006 Civic and Political Health of the Nation Dataset (collected by CIRCLE) (for a similar exercise, see Kam et al. 2007).

These data consist of telephone and web interviews with 2,232 individuals age 15 years and older living in the continental US. We limited the analysis to individuals aged 18 years and over. We selected all ostensibly politically relevant predispositions available in the data,^{xix} and then compared individuals currently enrolled in college against the general population. The web appendix^{xx} contains question wording for each item.

[Table 4-1 about here]

As we can see from Table 4-1, in most cases, the difference in means for students and the nonstudent general population are indistinguishable from zero. Students and the nonstudent general population are, on average, indistinguishable when it comes to partisanship (we find this for partisan direction and intensity), ideology, the importance of religion, belief in limited government, views about homosexuality as a way of life, the contributions of immigrants to

society, social trust, degree of following and discussing politics, and overall media use. Students are distinguishable from nonstudents in religious attendance, in level of political information as measured in this dataset,^{xxi} and in specific types of media use. Overall, however, we are impressed by just how similar students are to the nonstudent general population on key covariates often of interest to political scientists.

In cases where samples differ on variables that are theorized to influence the size and direction of the treatment effect, the researcher should, as we have noted above, model the interaction. The researcher also might consider cases where students – despite differing on relevant variables – might be advantageous. In some situations, students facilitate testing a causal proposition. Students are relatively educated, in need of small amounts of money, and accustomed to following instructions (e.g., from professors) (Guala 2005, 33-4). For these reasons, student samples may enhance the experimental realism of experiments that rely on induced value theory (where monetary payoffs are used to induce preferences) and/or involve relatively complicated, abstract instructions (Friedman and Sunder 1994, 39-40).^{xxii} The goal of many of these experiments is to test theory and, as mentioned, the match to the theoretical parameters (e.g., the sequence of events if the theory is game theoretic) is of utmost importance (rather than mundane realism).

Alternatively, estimating a single treatment effect upon a student sample subject pool can sometimes make it *harder* to find effects. For example, studies of party cues examine the extent to which subjects will follow the advice given to them by political parties. Strength of party identification might be a weaker cue for student subjects, whose party affiliations are still in the formative stages (Campbell et al. 1960). If this were the case, then the use of a student sample would make it even more difficult to discover party cue effects. To the extent that party cues

work among student samples, these likely *underestimate* the degree of cue-taking that might occur among the general population, whose party affiliations are more deeply grounded. Similarly, students seem to exhibit relatively lower levels of self-interest and susceptibility to group norms (Sears 1986, 524) meaning that using students in experiments on these topics increases the challenge of identifying treatment effects.^{xxiii}

Finally, it is worth mentioning that if the goal of a set of experiments is to generalize a theory, then testing the theory across a set of carefully chosen convenience samples may even be superior to testing the theory within a single random sample.^{xxiv} A theory of the moderating effect of attitude crystallization on the effects of persuasive communications might be better tested on a series of different samples (and possibly different student samples) that vary on the key covariate of interest.

Researchers need to consider what particular student sample characteristics might lead a causal relationship discovered in the sample to systematically differ from what would be found in the general population. Researchers then need to elaborate upon the direction of the bias: the variation might facilitate the assessment of causation, and/or it might lead to *either* an overestimation *or* an underestimation of what would be found in the general population.

4. Conclusion

As mentioned, political scientists are guilty of a “near obsession” with external validity (McDermott 2002, 334). And, this obsession with external validity focuses nearly entirely upon a single dimension of external validity: who is studied. Our goal in this chapter has been to situate the role of experimental samples within a broader framework of how one might assess the generalizability of an experiment. Our key points are, as follows:

- The external validity of a single experimental study must be assessed in light of an entire research agenda, *and* in light of the goal of the study (e.g., testing a theory or searching for facts).
- Assessment of external validity involves multiple-dimensions including the sample, context, time, and conceptual operationalization. There is no reason *per se* to prioritize the sample as the source of an inferential problem. Indeed, we are more likely to lack variance on context and timing since these are typically constants in the experiment.
- In assessing the external validity of the sample, experimental realism (as opposed to mundane realism) is critical, and there is nothing inherent to the use of student subjects that reduces experimental realism.
- The nature of the sample—and the use of students—matters in certain cases. However, a necessary condition is a heterogeneous (or moderated) treatment effect. Then the impact depends on:
 - If the heterogeneous effect is theorized, the sample only matters if there is virtually no variance on the moderator. If there is even scant variance, the treatment effect not only will be correctly estimated but may be estimated with greater confidence. The suitability of a given sample can be assessed (e.g., empirical variance can be analyzed).
 - If the heterogeneous effect is not theorized, it may be misestimated. However, even in this case, evaluating the bias is not straightforward because any sample will be inaccurate (since the “correct” moderated relationship is not being modeled).

- The range of heterogeneous, non-theorized cases may be much smaller than is often thought. Indeed, when it comes to a host of politically relevant variables, student samples do not significantly differ from nonstudent samples.
- There are cases where student samples are desirable since they facilitate causal tests or make for more challenging assessments.

Our argument has a number of practical implications. First, we urge researchers to attend more to the potential moderating effects of the other dimensions of generalizability: context, time, and conceptualization. The last decade has seen an enormous increase in survey experiments, due in no small way to the availability of more representative samples. Yet scholars must account for the distinct context of the survey interview (e.g., Converse and Schuman 1974; Zaller 1992, 28). Sniderman et al. (1991) elaborates that “the conventional survey interview, though well equipped to assess variations among individuals, is poorly equipped to assess variation across situations” (265). Unlike most controlled lab settings, researchers using survey experiments have limited ability introduce contextual variations.

Second, we encourage the use of dual samples of students and nonstudents. The discovery of differences should lead to serious consideration of what drives distinctions (i.e., what is the underlying moderating dynamic and can it be modeled?). The few studies that compare samples (e.g., Gordon et al. 1986; James and Sonner 2001; Peterson 2001; Mintz et al. 2006; Depositario et al. 2009; Henrich et al. 2009), while sometimes reporting differences, rarely explore the nature of the differences.^{xxv} When dual samples are not feasible, researchers can take a second-best approach by utilizing question wordings that match those in general surveys (thereby facilitating comparisons).

Third, we hope for more discussion about the pros and cons of alternative modes of experimentation, which may be more amenable to utilizing nonstudent subjects. While we recognize the benefits of using survey and/or field experiments, we should not be overly sanguine about their advantages. For example, the control available in laboratory experiments enables researchers to maximize experimental realism (e.g., by using induced value or simply by more closely monitoring the subjects). Similarly, there is less concern in laboratory settings about compliance \times treatment interactions that become problematic in field experiments or spillover effects in survey experiments (Transue et al. 2009; also see Sinclair’s chapter in this volume). The increased control offered by the laboratory setting often affords greater ability to manipulate context and time, which, we have argued, deserve much more attention. Finally, when it comes to the sample, attention should be paid to the nature of any sample and not just student samples. This includes consideration of non-response biases in surveys (see Groves and Peytcheva 2008) and the impact of using “professional” survey respondents that are common in many web-based panels.^{xxvi} In short, the nature of any particular sample needs to be assessed in light of various tradeoffs including consideration of an experiment’s goal, costs of different approaches, other dimensions of generalizability, and so on.

We have made a strong argument for the increased usage and acceptance of student subjects, suggesting that the burden of proof be shifted from the experimenter to the critic (also see Friedman and Sunder 1994, 16). We recognize that many will not be persuaded; however, at the very least, we hope to have stimulated increased discussion about why and when student subjects may be problematic.

References

- Anderson, Craig A., and Brad J. Bushman. 1997. "External Validity of 'Trivial' Experiments: The Case of Laboratory Aggression." *Review of General Psychology* 1: 19-41.
- Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. New York: Free Press.
- Arceneaux, Kevin, and Martin Johnson. 2008. "Choice, Attention, and Reception in Political Communication Research." Presented at the annual meeting of the International Society for Political Psychology, Paris, France.
- Aronson, Elliot, Marilynn B. Brewer, and J. Merrill Carlsmith. 1985. "Experimentation in Social Psychology." In *Handbook of Social Psychology*, eds. Gardner Lindzey, and Elliot Aronson. 3rd Edition. New York: Random House.
- Aronson, Elliot, and J. Merrill Carlsmith. 1968. "Experimentation in Social Psychology." In *Handbook of Social Psychology*, eds. Gardner Lindzey, and Elliot Aronson. 2nd Ed. Reading, MA: Addison-Wesley.
- Aronson, Elliot, Timothy D. Wilson, Marilynn B. Brewer. 1998. "Experimentation in Social Psychology." In *The Handbook of Social Psychology*, eds. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. 4th Ed. Boston: McGraw-Hill.
- Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104: 226-42.
- Baron, Reuben M., and David A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51: 1173-82.
- Benz, Matthias, and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field?: Evidence from Donations." *Experimental Economics* 11: 268-81.
- Berkowitz, Leonard, and Edward Donnerstein. 1982. "External Validity is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments." *American Psychologist* 37: 245-57.
- Campbell, Angus, Philip Converse, Warren Miller, and Donald Stokes. 1960 [1980 reprint]. *The American Voter*. Chicago: University of Chicago Press.
- Campbell, Donald T. 1969. "Prospective: Artifact and Control." In *Artifact in Behavioral Research*, eds. Robert Rosenthal, and Robert Rosnow. New York: Academic Press.
- Chong, Dennis, and James N. Druckman. 2007a. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101: 637-55.
- Chong, Dennis, and James N. Druckman. 2007b. "Framing Theory." *Annual Review of Political*

Science 10: 103-26.

Converse, Jean M., and Howard Schuman. 1974. *Conversations at Random: Survey Research as Interviewers See It*. New York: Wiley.

Depositario, Dinah Pura T., Rodolfo M. Nayga Jr., Ximing Wu, and Tiffany P. Laude. 2009. "Should Students Be Used as Subjects in Experimental Auctions?" *Economic Letters* 102: 122-4.

Dickhaut, John W., J. Leslie Livingstone, and David J. H. Watson. 1972. "On the Use of Surrogates in Behavioral Experimentation." *The Accounting Review* 47, Supplement: 455-71.

Druckman, James N. 2001. "On The Limits of Framing Effects." *Journal of Politics* 63: 1041-66.

Druckman, James N. 2009. "Competing Frames in a Campaign." Unpublished Paper, Northwestern University.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research Political Science." *American Political Science Review* 100: 627-36.

Franklin, Charles H. 1991. "Efficient Estimation in Experiments." *The Political Methodologist* 4: 13-5.

Friedman, Milton. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Economics: A Primer for Economists*. New York: Cambridge University Press.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.

Gerber, Alan, James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2007. "The Influence of Television and Radio Advertising on Candidate Evaluations: Results from a Large Scale Randomized Experiment." Unpublished Paper, Yale University.

Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.

Gordon, Michael E., L. Allen Slade, and Neal Schmitt. 1986. "The 'Science of the Sophomore' Revisited: From Conjecture to Empiricism." *Academy of Management Review* 11: 191-207.

- Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72: 167-89.
- Guala, Francesco. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2009. "The Weirdest People in the World: How Representative Are Experimental Findings from American University Students? What Do We Really Know About Human Psychology?" Unpublished Paper, University of British Columbia.
- James, William L., and Brenda S. Sonner. 2001. "Just Say No to Traditional Student Samples." *Journal of Advertising Research* 41: 63-71.
- Kam, Cindy D. 2005. "Who Toes the Party Line?: Cues, Values, and Individual Differences." *Political Behavior* 27: 163-82.
- Kam, Cindy D. 2007. "When Duty Calls, Do Citizens Answer?" *Journal of Politics* 69: 17-29.
- Kam, Cindy, and Robert J. Franzese. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: University of Michigan Press.
- Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415-40.
- Kinder, Donald R. 1998. "Communication and Opinion." *Annual Review of Political Science* 1: 167-97.
- Kruglanski, Arie W. 1975. "The Human Subject in the Psychology Experiment: Fact and Artifact." In *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz. New York: Academic Press.
- Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. Cambridge, UK: Cambridge University Press.
- Liyanarachchi, Gregory A. 2007. "Feasibility of Using Student Subjects in Accounting Experiments: A Review." *Pacific Accounting Review* 19: 47-67.
- MacDonald, Paul. 2003. "Useful Fiction or Miracle Maker: The Competing Epistemological Foundations of Rational Choice Theory." *American Political Science Review* 97: 551-65.
- McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10: 325-42.

- Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67: 371-378.
- Mintz, Alex, Steven B. Redd, and Arnold Vedlitz. 2006. "Can We Generalize From Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?" *Journal of Conflict Resolution* 50: 757-76.
- Mook, Douglas G. 1983. "In Defense of External Invalidity." *American Psychologist* 38: 379-87.
- Morton, Rebecca B. and Kenneth C. Williams. 2008. "Experimentation in Political Science." In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.
- Peterson, Robert A. 2001. "On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-analysis." *Journal of Consumer Research* 28: 450-61.
- Plott, Charles R. 1991. "Will Economics Become an Experimental Science?" *Southern Economic Journal* 57: 901-19.
- Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press.
- Sears, David O. 1986. "College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51: 515-30.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Simon, Herbert A. 1963. "Problems of Methodology Discussion." *American Economic Review Proceedings* 53: 229-31.
- Simon, Herbert A. 1979. "Rational Decision Making in Business Organizations." *American Economic Review* 69: 493-513.
- Slothuus, Rune. 2009. "The Political Logic of Party Cues in Opinion Formation." Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Sniderman, Paul M., Richard A. Brody, and Philip E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. Cambridge: Cambridge University Press.
- Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion*, eds. Willem E. Saris, and Paul M. Sniderman. Princeton, NJ: Princeton University Press.

Stevens, Charles D. and Ronald A. Ash. 2001. "The Conscientiousness of Students in Subject Pools: Implications of 'Laboratory' Research." *Journal of Research in Personality* 35: 91-7.

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17: 143-61.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zimbardo, Phillip. "A Pirandellian Prison," *New York Times Magazine* April 8, 1973.

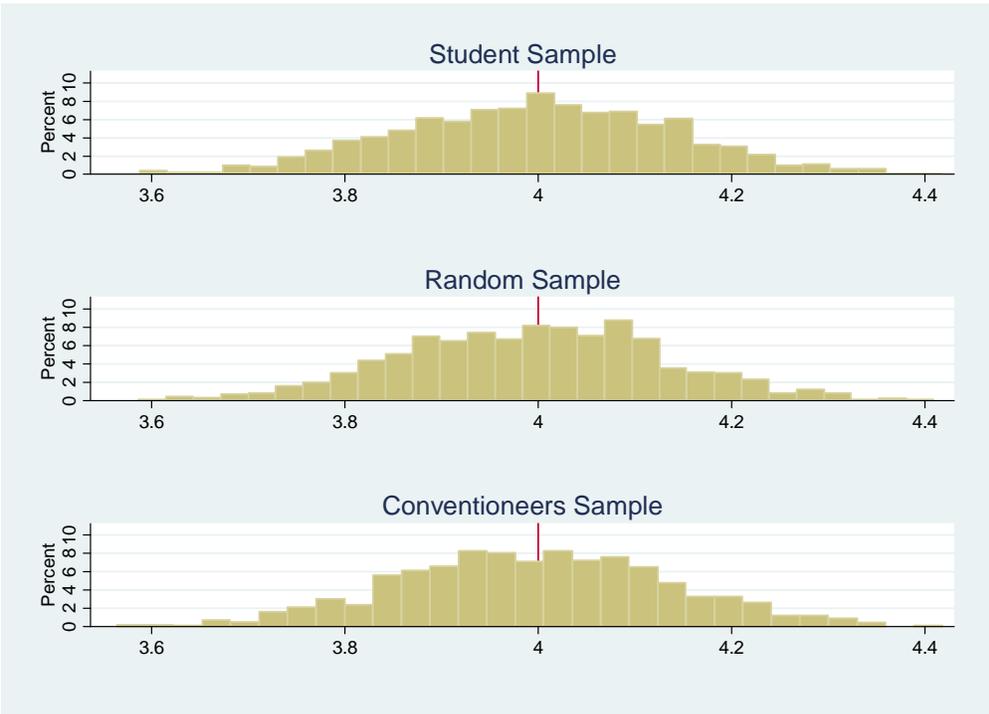
Table 4-1. Comparison of students versus nonstudent general population

	Students	Nonstudent General Population	<i>p-value</i>
Partisanship	0.47 (0.02)	0.45 (0.01)	<i>ns (not significant)</i>
Ideology	0.50 (0.01)	0.52 (0.01)	<i>ns</i>
Religious Attendance	0.56 (0.02)	0.50 (0.01)	<0.01
Importance of Religion	0.63 (0.02)	0.62 (0.02)	<i>ns</i>
Limited Government	0.35 (0.03)	0.33 (0.02)	<i>ns</i>
Homosexuality as a way of life	0.60 (0.03)	0.62 (0.02)	<i>ns</i>
Contribution of immigrants to society	0.62 (0.03)	0.63 (0.02)	<i>ns</i>
Social trust	0.34 (0.03)	0.33 (0.02)	<i>ns</i>
Follow politics	0.68 (0.02)	0.65 (0.01)	<i>ns</i>
Discuss politics	0.75 (0.01)	0.71 (0.01)	<i>ns</i>
Political information (0 to 6 correct)	2.53 (0.11)	1.84 (0.07)	<0.01
Newspaper use (0 to 7 days)	2.73 (0.14)	2.79 (0.11)	<i>ns</i>
National TV news (0 to 7 days)	3.28 (0.15)	3.63 (0.10)	<0.05
News radio (0 to 7 days)	2.47 (0.16)	2.68 (0.11)	<i>ns</i>
Web news (0 to 7 days)	3.13 (0.16)	2.18 (0.10)	<0.01
Overall media use	2.90 (0.09)	2.83 (0.06)	<i>ns</i>

Weighted analysis. Means with standard errors in parentheses. See the appendix for variable coding and question text.

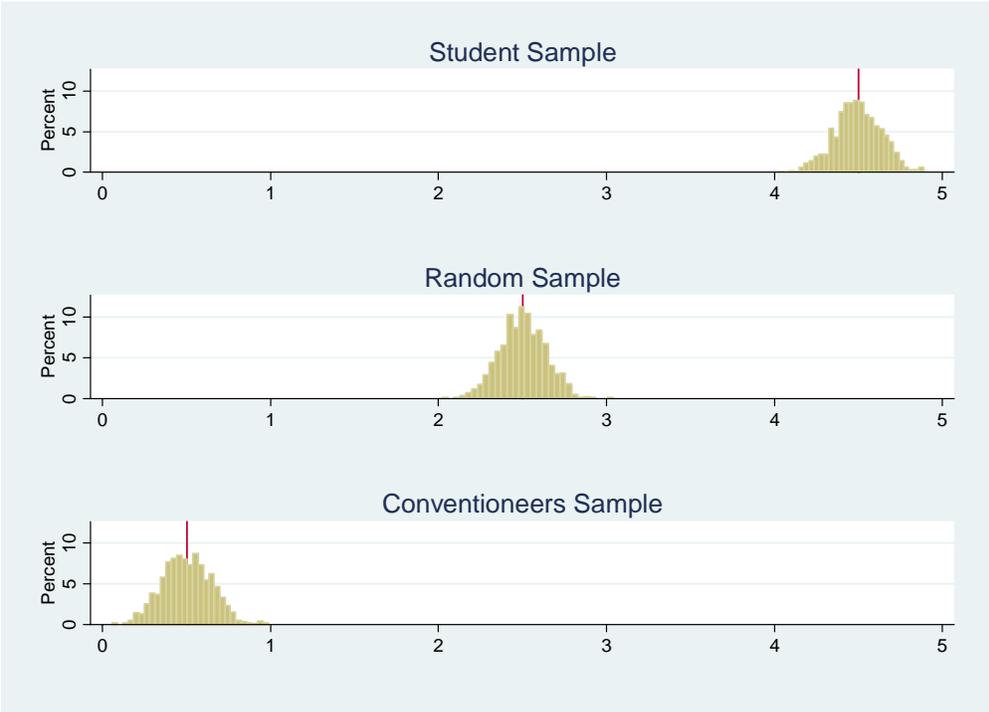
Source: 2006 Civic and Political Health Survey.

Figure 4-1. Sampling distribution of b_T , single treatment effect



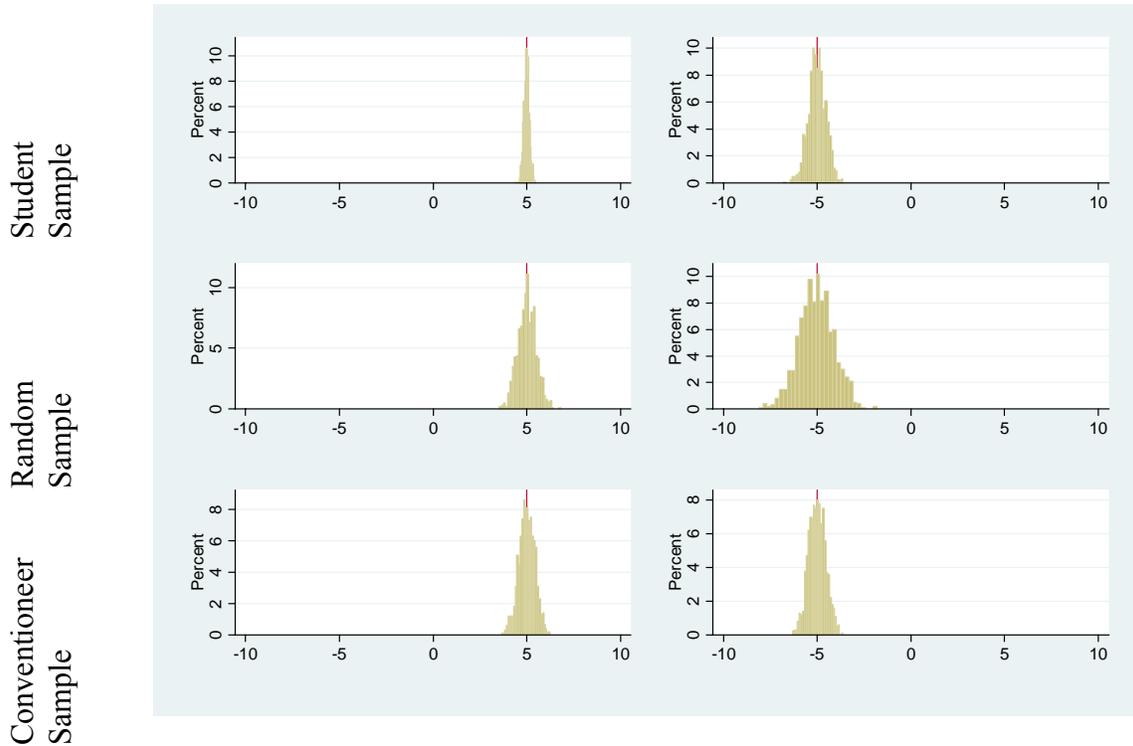
Note: 1,000 iterations, estimated using Eq [1]

Figure 4-2. Sampling distribution of b_T , heterogeneous treatment effects



Note: 1,000 iterations, estimated using Eq [1]

Figure 4-3. Sampling distributions of b_T and b_{TZ} , heterogeneous treatment effects
 Sampling Distribution of b_T Sampling Distribution of b_{TZ}



Note: 1,000 iterations, estimated using Eq [3]

ⁱ We thank Kevin Arceneaux, Don Green, Jim Kuklinski, Peter Loewen, and Diana Mutz for helpful advice, and Samara Klar and Thomas Leeper for research assistance.

ⁱⁱ Through 2008, Sears’ (1986) article has been cited an impressive 446 times according to the Social Science Citation Index. It is worth noting that Sears’ argument is conceptual – he does not offer empirical evidence that student subjects create problems.

ⁱⁱⁱ This is consistent with a Popperian approach to causation that suggests causal hypotheses are never confirmed and evidence accumulates via multiple tests, even if all of these tests have limitations. Campbell (1969) offers a fairly extreme stance on this when he states, “...had we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Wistar staring white rats” (361).

^{iv} A related example comes from Barabas and Jerit’s (2010) study that compares the impact of communications in a survey experiment against analogous dynamics that occurred in actual news coverage. They find the survey experiment vastly over-stated the effect, particularly among certain sub-groups. Sniderman and Theriault (2004) and Chong and Druckman (2007a) also reveal the importance of context; both studies show that prior work that limits competition between communications (i.e., by only providing participants with a single message rather than a mix that is typically found in political contexts) likely misestimate the impact of communications on public opinion.

^v For reasons explained in his paper, Druckman (2009) also focuses on individuals more likely to have formed prior opinions about the casino.

^{vi} Another relevant timing issue concerns the duration of any experimental treatment effect (see, e.g., Gaines et al. 2007; Gerber et al. 2007).

^{vii} Aronson et al. (1998) explain that it “is often assumed (perhaps mindlessly!) that all studies should be as high as possible in external validity, in the sense that we should be able to generalize the results as much as possible across populations and settings and time. Sometimes, however, the goal of the research is different” (132).

^{viii} A third evaluative criterion is psychological realism, which refers to “the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life” (Aronson et al. 1998, 132). The relevance of psychological realism is debatable and depends on one’s philosophy of science (c.f., Friedman 1953; Simon 1963, 1979, 475-476; also see MacDonald 2003).

^{ix} By “seriously,” we mean analogous to how individuals treat the same stimuli in the settings to which one hopes to generalize (and not necessarily “serious” in a technical sense). We do not further discuss steps that can be taken to ensure experimental realism, as this moves into the realm of other design issues (e.g., subject payments, incentives; see Dickson’s chapter in this volume).

^x Berkowitz and Donnerstein (1982) explain that “The meaning the subjects assign to the situation they are in and the behavior they are carrying out [i.e., experimental realism] plays a greater part in determining generalizability of an experiment’s outcome than does the sample’s demographic representatives or the setting’s surface realism” (249).

^{xi} This claim is in need of empirical evaluation, as it may be that students are more compliant and this may impact realism.

^{xii} For ease of exposition, our example only has one treatment group. The lessons easily extend to multiple treatment groups.

^{xiii} We could have specified a data generating process that also includes a direct relationship between y and some individual-level factors such as partisanship or sex (consider a vector of such variables, \mathbf{X}). Under random assignment, the expected covariance between the treatment and \mathbf{X} is zero. Hence, if we were to estimate the model without \mathbf{X} , omitted variable bias would technically not be an issue. If the data generating process does include \mathbf{X} , and even though we might not have an omitted variable bias problem, including \mathbf{X} in the model may still be advisable. Inclusion of relevant covariates (that is, covariates that, in the data generating process, actually have a nonzero effect on y) will reduce e_i (the difference between the observed and predicted y), which in turn will reduce s^2 , resulting in more precise estimated standard errors for our coefficients (see Franklin 1991). Moreover, it is only in *expectation* that $\text{Cov}(\mathbf{X}, T) = 0$. In any given sample, $\text{Cov}(\mathbf{X}, T)$ may not equal zero. Inclusion of covariates can mitigate against incidental variation in cell composition. In advising inclusion of control variables, Ansolabehere and Iyengar (1995) note, “...randomization does not always work. Random assignment of treatments provides a general safeguard against biases but it is not foolproof. By chance, too many people of a particular type may end up in one of the treatment groups, which might skew the results” (172; see also Bowers’ chapter in this volume).

^{xiv} This example is inspired by Sears’ (1986) discussion of “Uncrystallized Attitudes.”

^{xv} And, of course, crystallization might vary across different types of issues. On some issues (e.g., financial aid policies), students might have highly crystallized views, whereas conventioners might have less crystallized views.

^{xvi} Now, if our goal was to use our three samples to make descriptive inferences about the general population’s mean level of attitude crystallization, then both the Student Sample and the Conventioneers Sample would be inappropriate. The goal of an experimental design is expressly *not* to undertake this task. Instead, the goals of an experimental design are to estimate the causal effect of some treatment and then to generalize it.

^{xvii} See Baron and Kenny (1986) for the distinction between moderation and mediation. Psychologists refer to the case where Z affects the *effect* of X as moderation (i.e., an interaction effect). Psychologists refer to mediation when some variable X influences the *level* of some variable Z , whereby X affects Y through its effect on the level of Z . For an extended treatment of interaction effects in regression analysis, see Kam and Franzese (2007). For a discussion of mediation, see Bullock and Ha’s chapter in this volume.

^{xviii} Uncovering more certainty in the Student and Conventioneers Samples (compared to the Random Sample) derives from the specific ways in which we have constructed the distributions of Z . If the Random Sample were, say, uniformly distributed rather than normally distributed along Z , then the same result would not hold. The greater precision in the estimates depends upon the underlying distribution of Z in a given sample.

^{xix} We did this before looking at whether there were differences between students and the nonstudent general population sample; that is, we did not selectively choose variables.

^{xx} Available at <http://faculty.wcas.northwestern.edu/~jnd260/publications.html>.

^{xxi} The measure of political information in this dataset is quite different from that typically found in the ANES; it is heavier on institutional items and relies on more recall than recognition.

^{xxii} We suspect that this explains why the use of student subjects seems to be much less of an issue in experimental economics (e.g., Guala 2005).

^{xxiii} As explained, students also tend to be more susceptible to persuasion (Sears 1986). This makes them a more challenging population on which to experiment if the goal is to identify conditions where persuasive messages fail (e.g., Druckman 2001).

^{xxiv} Convenience samples might be chosen to represent groups that are high and low on a particular covariate of interest. This purposive sampling might yield more rewards than using a less informative random sample.

^{xxv} For example, Mintz et al. (2006) implemented an experiment, with both students and military officers, about counterterrorism decision-making. They find the two samples significantly differed, on average, in the decisions they made, the information they used, the decision strategies they employed, and the reactions they displayed. They (Mintz et al. 2006) conclude that “student samples are often inappropriate, as empirically they can lead to divergence in subject population results” (769). We would argue that this conclusion is pre-mature. While their results reveal on average differences between the samples, the authors leave unanswered why the differences exist. Mintz et al. (2006, 769) speculate that the differences may stem from variations in expertise, age, accountability, and gender. A thorough understanding of the heterogeneity in the treatment effects (which, as explained, is the goal of any experiment) would, thus, require exploration of these moderators. Our simulation results suggest that even if the student sample exhibited limited variation on these variables, it could have isolated the same key treatment dynamics as would be found in the military sample.

^{xxvi} The use of professional, repeat respondents raises similar issues to those caused by repeated use of participants from a subject pool (e.g., Stevens and Ash 2001).

5. Economics vs. Psychology Experiments: Stylization, Incentives, and Deception

Eric S. Dickson

In this chapter, I follow other authors (e.g., Kagel and Roth 1995; McDermott 2002; Camerer 2003; Morton and Williams 2009) in focusing on a few key dimensions of difference between experiments in the economic and psychological traditions.

The first section of this chapter considers the level of *stylization* typical in economics and psychology experimentation. While research in the political psychology tradition tends to place an emphasis on the descriptive realism of laboratory scenarios, work in experimental economics tends to proceed within a purposefully abstract, “context free” environment.

The second section of this chapter considers the kinds of *incentives* offered to subjects by experimentalists from these two schools of thought. Experimental economists generally offer subjects *monetary incentives* that depend on subjects’ choices in the laboratory – and, in game-theoretic experiments, the choices of other subjects as well. In contrast, psychology research tends not to offer inducements that are conditional on subjects’ actions, instead giving subjects fixed cash payments or fixed amounts of course credit.

The third section of this chapter considers the use of *deception*. The psychological school tends to see deception as a useful tool in experimentation, at times a necessary one; in contrast, the economic school by and large considers deception to be taboo.

These basic differences in research style highlight the historical divide between psychological and economic – alternatively, behavioral and rational choice – scholarship in political science. Over the years, scholars have tended to peer across this divide with more mistrust than understanding, and intellectual interchange between the different schools has been

lamentably limited in scope. The difference in approaches between psychologists and economists reflects more than the sociology of their respective traditions, however; many of the norms characteristic of each field have evolved in response to the specific nature of theory and of inquiry within the separate disciplines.

To say that each school of experimentation has categorical strengths and weaknesses would perhaps be too strong a claim. Rather, in this chapter I will argue that the advantages and disadvantages associated with specific design choices may play out differently, depending on the nature of the research question being posed, the theory being tested, and even of the results that are ultimately obtained.

In this chapter, I organize my discussion around the logic of inference in economics- and psychology-style experiments. Going down this path leads me to several conclusions that may at first seem counterintuitive. For example, I will argue that stylized, economics-style experimentation can sometimes be particularly valuable in the study of essentially psychological research questions. Contrary to the way our discipline has traditionally been organized around separate schools of methodological practice, strategy and psychology are inextricably bound up together in virtually all of the political phenomena that we desire to understand. The multifaceted nature of our objects of study, along with the varying strengths and weaknesses of different research methods in attacking different problems, together highlight the advantages of methodological pluralism in building an intellectually cumulative literature in experimental political science.

1. Stylized versus Contextually Rich Experimental Scenarios

A first salient dimension of difference between economics and psychology experiments is rooted in the basic nature of the experimental scenarios presented to subjects. With some exceptions,

economics experiments tend to be carried out in a highly stylized environment, in which the scenarios presented to subjects are purposefully abstract, while experiments in psychology tend to evoke more contextually rich settings. Because the economic style of experimentation is likely to be more foreign to many readers, this discussion begins by describing some arguments that have been given in support of stylization in laboratory experiments.

The Logic of Stylization

Research in the economic style tends to frame experimental scenarios in an abstract rather than in a naturalistic manner. The roles assumed by subjects, and the alternatives subjects face, are generally described using neutral terminology with a minimum of moral or emotional connotations; experimental instructions are written in a technocratic style. For example, in their landmark study of punishment in games of public goods provision, Fehr and Gächter (2000) employ an experimental frame using strictly neutral language, never once mentioning the word punishment or other potentially-charged terms such as fairness or revenge. In a similar way, Levine and Palfrey (2007) use the labels X and Y, rather than terms like vote and abstain, in their experimental study on voter turnout; the cost of voting is translated into a “Y bonus” accruing only to individuals who choose Y – that is, do not vote. In their study of deliberation, Dickson, Hafer, and Landa (2008) model individual decisions to communicate in a stylized environment; the “arguments” exchanged during deliberation are represented using simple single-digit numbers.

The abstract experimental tasks associated with this form of stylization are used in part because of a desire to maintain experimental control. Researchers in this tradition generally believe that the use of normatively-charged terms such as punishment, fairness, or revenge may evoke reactions in subjects whose source the analyst cannot fathom and which the analyst cannot

properly measure. Experimental economists would generally argue that such loss of control would limit the generalizability, and thus the usefulness, of their findings in the laboratory.

According to this argument, the descriptively-appealing complexity of highly contextual experiments comes with strings attached when it comes to inference. Suppose that a particular effect is measured in a contextually rich setting. More or less by definition, contextually rich settings contain many features that could potentially claim subjects' attention or influence subjects' behavior or cognition. Given this, how could we know *which* feature of the setting – or which combination of features – led to the effect that we observed?

In contrast, it is argued that a similar effect measured in a stylized setting may have wider lessons to teach. One argument for this claim can be explicated through the use of two examples. First, consider the Fehr and Gächter (2000) experiment, which demonstrated that many experimental subjects are willing to undertake costly punishment of counterparts who fail to make adequate contributions to a public good, even under conditions where such punishment is costly and no benefit from punishment can accrue to the punisher. Because this result was obtained in such an abstract choice environment, which did not directly prime subjects to think in terms of punishment or fairness, the result seems unlikely to be merely an artifact of some abstruse detail of the experimental frame presented to subjects. A more natural interpretation of the study's findings is that a willingness to punish the violation of norms is a basic feature of human nature that comes to be expressed even in novel settings in which subjects lack experience or obvious referents. As such, the use of an abstract, stylized environment in the study arguably strengthens rather than weakens the inferences we make from its result. Second, Dickson et al. (2008) demonstrate that many subjects “overspeak” compared to a benchmark equilibrium prediction – that is, that subjects often choose to exchange arguments during the

course of deliberation even when they are more likely to alienate listeners than persuade them. This finding suggests that deliberation may unfold in a manner more compatible with the deliberative democratic ideal of a “free exchange of arguments” than a fully-strategic model would be likely to predict. In their study, stylization has at least two distinct advantages. The use of a stylized, game-theoretic environment allowed for the definition of a rational-choice benchmark in the first place – without which overspeaking could not have been defined or identified. And, the finding that individuals overspeak even in a stylized environment without obvious normative referents underscores the behavioral robustness of individual willingness to exchange arguments with others.

Such arguments in favor of stylization have, in fact, even been employed from time to time within social psychology itself. The minimal group experimental paradigm (Tajfel et al. 1971; Tajfel and Turner 1986) demonstrated that social identities can motivate individual behavior even when those social identities were somewhat laughable constructs artificially induced within a stylized setting: for instance, dividing subjects based on their tendency to overcount or undercount dots on a screen or their preference for paintings by one abstract painter (Klee) over another (Kandinsky). The finding that even these social identities could affect behavior helped to establish social identity theory and to motivate a vast field of research.

The Limits of Stylization

The first and perhaps most obvious point is that certain research questions – particularly certain research questions in political psychology -- cannot reasonably be posed both in stylized and in contextually rich settings. Just to take one clear-cut example, Brader (2005) studies the effects of music within political advertisements on voters’ propensities to turn out, seek additional political information, and other dependent variables. It would obviously make little sense to attempt to

translate such a study into a highly stylized setting, because the psychological mechanisms Brader explores are so deeply rooted in the contextual details of his experimental protocol.

Many other research questions, however, could potentially lend themselves to exploration either in stylized or in highly contextual contexts. In considering the advantages and disadvantages of stylization in such cases, a natural question to ask is whether or not experimental results obtained using both methods tend to lead to similar conclusions.

For at least some research questions, the evidence suggests that stylization may lead to conclusions that are misleading or at least incomplete. A classic example comes from the psychology literature on the Wason selection task. In Wason's (1968) original study, subjects were given a number of cards, each of which had a number on side and a letter on the other, and a rule that had to be tested: namely, that every card with a vowel on one side has an even number on the other side. Given a selection of cards, labeled E, K, 4, and 7, subjects were required to answer *which* cards must be turned over in order to test the rule. In this study, only a small fraction of subjects gave the correct answer (E and 7); especially few noted that the rule could be falsified by turning over the 7 and finding a vowel, while others included 4 in their answer in an apparent search for information confirming the rule. This finding is often taken as clear evidence for a confirmatory bias in hypothesis testing.

The Wason selection task became a popular paradigm in the aftermath of the original study, and parallel versions have been carried out in many different settings. Interestingly, subjects' performance at the task appears to be highly variable, depending on the context in which the task is presented. In another well-known study, Griggs and Cox (1982) present subjects with a selection task logically equivalent to Wason's, but rather than using abstract letters and numbers as labels, the task is framed as a search for violators of a social norm:

underage drinking. In this study, most subjects are readily able to answer correctly that people who are drinking and people who are known to be underage are the ones whose age or behavior need to be examined when searching for instances of underage drinking.

Results such as these suggest that subjects may sometimes think about problems quite differently, depending on the frame in which the problem is presented, an intuition that seems natural to scholars with a background in psychology. At the same time, such results by no means imply that stylized studies yield different results from highly contextual ones more generally. To take framing effects themselves as an example, parallel literatures within economics and psychology suggest that frames can affect choice behavior in similar ways both in stylized and highly contextual environments.

As of now, there is nothing like a general theory that would give experimentalists guidance as to when stylization might pose greater problems for external validity. Many scholars find that stylization can be beneficial, given their research questions – because of a perceived higher degree of experimental control, because stylization can sometimes allow for a clearer definition of theoretical benchmarks than might be the case in a highly contextual environment, or because stylized environments can sometimes pose a “tough test” for measuring behavioral or psychological phenomena, as in the Fehr and Gächter (2000) and Dickson et al. (2008) studies. At the same time, a literature consisting wholly of such studies would widely be met with justifiable skepticism about external validity. At least for many research areas within political science, the best progress is likely to be made most quickly when research in both traditions is carried out – and when scholars communicate about their findings across traditional dividing lines. When research using different techniques tends to point in the same direction, we can have more confidence in the results than we could have if only one research method had been

employed. When research using different techniques instead points in different directions, the details of these discrepancies may prove invaluable in provoking new theoretical explanations for the phenomenon at hand, as scholars attempt to understand the discrepancies' origins.

2. The Use of Monetary Incentives

In most economics experiments, subjects receive cash payments that depend on their own choices in the laboratory and, in the case of game-theoretic experiments, on the choices of other people. In contrast, subjects who take part in political psychology experiments are generally compensated in a way that does not depend on the choices they make, typically either a fixed cash payment or a fixed amount of course credit. What motivates experimentalists from these two traditions to take different approaches to motivating subjects?

The most obvious point to make is that many research studies in political psychology are not well-suited to the use of monetary incentives because the relevant quantities of interest cannot be monetized in a reasonable way. For example, in a framing study by Druckman and Nelson (2003), subjects report their attitudes on political issues after exposure to stimuli in the form of newspaper articles; clearly, in studies with a dependent variable like this one, offering subjects financial incentives to report one opinion as opposed to another would be of no help whatsoever in studying framing effects or the formation of public opinion.

Of course, the same is not true of *all* research questions of interest to political experimentalists, political psychologists included. As such, experimenters sometimes have a real choice to make in deciding whether to motivate subjects with monetary incentives. In considering the implications of this choice, it is useful to review some of the varied purposes for which monetary incentives have been used in experiments.

Monetary Incentives as a Means of Rewarding Accuracy or Reducing Noise

One potential use for monetary incentives in experiments is to reward accuracy. Experimentalists wish to ensure that subjects actually pay attention and properly engage the tasks they are meant to perform. In settings where a “right answer” is both definable and, at least in principle, achievable by the subject – a setting very unlike the Druckman and Nelson (2003) article cited above – financial inducements can help fulfill this role. For example, in a survey experiment on political knowledge, Prior and Lupia (2008) find that monetary rewards motivate subjects to respond more accurately and to take more time considering their responses. This result suggests that financial inducements can sometimes help elicit more accurate measures of knowledge and reduce levels of noise in survey responses.

A natural, and related, setting for the use of such methods in political experiments involves the study of political communication. Scholars want to understand how individuals learn from the political communications to which they are exposed – and whether citizens are actually able to learn what they need to in order to make reasoned choices (Lupia and McCubbins 1998). In pursuit of these objectives, a number of scholars have devised stylized experimental settings in which subjects receive messages whose informational value can be objectively weighed using Bayes’ Rule in the context of a signaling game equilibrium (e.g., Lupia and McCubbins 1998; Dickson in press). Subjects then receive monetary rewards that depend on the degree of fit between their own posterior beliefs and the “correct” beliefs implied by Bayesian rationality in equilibrium.

Monetary Incentives as a Means of Controlling for Preferences

Many experiments in political economy focus on the effects of institutions in shaping individual behavior. Such experiments are typically organized as tests of predictions from game-theoretic models. Of course, actors’ preferences over different possible outcomes are primitive elements of

such models. As such, in order to expose a game-theoretic model to an experimental test, it must be that there is some means of inducing subjects to share the preferences of actors in the theoretical model. In economics experiments, this is done through the use of monetary incentives for subjects.

It is instructive to highlight the difference between this approach and typical research methods in the psychological tradition. In political psychology experiments, direct inquiry into the nature of individual motivations, preferences, and opinions is often the goal. In contrast, for the purposes of testing a game-theoretic model, economics experiments generally prefer to control for individual motivations by manipulating them exogenously, to the extent that this is possible. By controlling for preferences using monetary incentives, experimental economists attempt to focus on testing other aspects of their theoretical models, such as whether actors make choices that are consistent with a model's equilibrium predictions, or the extent to which actors' cognitive skills enable them to make the optimal choices predicted by theory.

Monetary Incentives as a Means of Measuring Social Preferences

Finally, it might also be noted that the use of monetary incentives can be beneficial for the study of subjects' intrinsic motivations. Consider, for example, the Fehr and Gächter (2000) study cited earlier. Subjects interacted within a stylized environment, making public goods contributions decisions and choosing whether or not to punish others based on their behavior. In the experiment, both kinds of decisions were associated with monetary incentives; a decision to punish another subject, for example, came at a (monetary) cost to the punisher. That individuals are willing to engage in punishment even when this has a monetary cost and when no future monetary benefit can possibly accrue strengthens our sense of how strong subjects' intrinsic motivations to punish may be. Certainly this finding is more telling than would be a parallel

result from an analogous experiment in which subjects' decisions were hypothetical and they did not bear any personal material cost for punishing others. In principle, this methodology can potentially allow us to measure the strength of this intrinsic motivation by varying the scale of the monetary incentives. Thus, studies such as Fehr and Gächter can allow us to learn about individuals' intrinsic motivations by observing deviations from game theoretic predictions about how completely (monetarily) self-interested actors would behave.

Other studies have taken a similar approach, allowing for inquiry into traditionally psychological topics within the context of game-theoretic experiments. A prominent example is Chen and Li (2009), who translate the study of social identities into a lab environment where subjects play games for monetary incentives, thereby offering a novel tool for measuring the strength of identities and the effects of identities on social preferences.

Does the Scale of Monetary Incentives Matter?

If an experimentalist decides that motivating subjects with monetary incentives is appropriate for her study, one basic question of implementation involves the appropriate *scale* for monetary incentives. It is not unusual for experimental economics labs to have informal norms that subjects' expected earnings should not fall below some minimum rate of compensation; the maintenance of a willing subject pool requires that "customers" be reasonably happy overall with their experiences in the lab. Morton and Williams (2010) summarize existing norms by estimating that payments are typically structured to average around 50 to 100 percent above the minimum wage for the time spent in the lab. Such considerations aside, resource constraints give experimentalists a natural incentive to minimize the scale of payoffs in order to maximize the amount of data that can be selected – so long as the payments subjects that receive are sufficient to motivate them in the necessary way.

A recent voting game study by Bassi, Morton, and Williams (2008) suggests that the scale of financial incentives can affect experimental results. In their study, the inducements offered to subjects varied across three treatments, involving a flat fee only, a scale typical of many experimental economics studies, and a larger scale offering subjects twice as much. The fit between subjects' behavior and game-theoretic predictions became monotonically stronger as incentives increased; suggestively, this pattern was found to be most prominent for the most cognitively challenging tasks faced by subjects. These results suggest that, at least in some settings, higher rates of payment to subjects can increase subjects' level of attention to the experiment in a way that may affect behavior, a result consistent with intuitions derived from Prior and Lupia (2008), as well as related studies in economics (e.g. Camerer and Hogarth 1999).

Gneezy and Rustichini (2000) carried out a study on IQ test performance that communicates a compatible message. Their experiment varied financial incentives for correct answers across four distinct treatments. They found performance to be identical in the two treatments offering the least incentives for performance (one of which simply involved a flat show-up fee), performance to also be identical in the two treatments offering the highest incentives, but performance in the higher-incentive treatments exceeded that in the lower-incentive treatments. This finding, along with Prior and Lupia (2008) and Bassi et al. (2008), suggests that a higher scale of incentives can increase attention, at least up to a point; and that higher attention can increase performance, at least up to a point that is determined in part by the difficulty of the problem.

This pattern has implications for the kinds of inferences that can be made from studies employing monetary incentives. The nature of these implications can reasonably be expected to differ depending on the nature of the experimental findings. Consider some of the political

communication studies cited above. In the scenarios of Lupia and McCubbins (1998), for example, subjects are quite good at inferring the informational content of communications they receive from strategically motivated speakers. In such instances, confidence in a result's external validity may depend to some extent on the "calibration" between the financial incentives in play and the stakes involved in receiving analogous communications in the real world. The incentives offered by Lupia and McCubbins appear to be quite appropriate in scale. However, consider a counterfactual experiment in which the monetary stakes for subjects were much larger. If, in this counterfactual experiment, subjects were substantially more motivated to pay attention and make proper inferences by the monetary inducements in the laboratory than they would have been by naturalistic considerations in the real world, then a clear issue would arise in extrapolating from "good" performance in the laboratory to predictions about real world performance. In contrast, in the cheap-talk-and-coordination scenario of Dickson (In press), subjects systematically fail fully to account for a speaker's strategic incentives when inferring the information content of communications. Of course, proper calibration of financial incentives to real world motivations would always be an ideal. However, for a study whose central result demonstrates "poor" performance or the existence of a "bias" in subject behavior, confidence in external validity is likely to be stronger when the experimenter errs on the side of making financial incentives too large rather than too small – that is, our confidence that a particular form of bias actually exists will be stronger if it persists even when subjects have extra incentives to perform a task well in the laboratory relative to the weaker incentives they face in real world settings. This logic underscores the extent to which simple decisions of experimental design may have powerful effects on the inferences we can draw from an experiment, even when the results are the same across different designs. A given finding will generally be more impressive when the

experimental design is more heavily stacked against the emergence of that finding.

Potential Problems with the Use of Monetary Incentives

As noted above, monetary incentives may be a non-starter for some research questions, but there may be arguments in favor of their use for other research questions. Are there potential problems with the use of monetary incentives that may argue against their use in certain settings?

One potential issue involves interactions between subjects' intrinsic motivations and the external motivation they receive from financial incentives. Some research in psychology suggests that financial incentives can "crowd out" intrinsic motivations, leading to somewhat counterintuitive patterns of behavior. Among the best known examples of crowding out comes from Titmuss (1970), who showed that offering financial compensation for blood donations can lead to lower overall contribution levels. The standard interpretation is that individuals who donate blood are typically motivated to do so for altruistic reasons; when financial incentives are offered, individuals' mode of engagement with the blood donation system changes, with marketplace values coming to the fore while intrinsic motivations such as altruism are crowded out.

Whether crowding out poses a problem for the use of monetary incentives is likely to depend on the nature of the research question being explored. For the purposes of game theory testing, crowding out of intrinsic motivations can often actually be considered desirable, because the experimenter wishes exogenously to assign preferences to subjects in order to instantiate the experimental game in the laboratory. On the other hand, suppose that social interactions within some real world setting of interest are believed to depend heavily on individuals' intrinsic motivations. In translating this real world setting into the laboratory, injudicious use of monetary incentives could potentially crowd out the intrinsic motivations that are central to the

phenomenon being studied.

This potential problem with the use of monetary incentives is in some instances a challenging one, because it may be difficult to anticipate to what extent such incentives might cause a transformation in subjects' modes of engagement with the experimental scenario. This concern goes hand in hand with understandable questions about the extent to which stylized economic and contextually rich psychological experiments actually investigate the same cognitive mechanisms, an important and understudied matter which may be illuminated more thoroughly in the future by across-school collaborations as well as by neuroscientific and other frontier research methods.

3. The Use of Deception

In few regards is the difference between the economic and psychological schools as stark as in attitudes about deceiving subjects. The more-or-less consensus view on deception in the experimental economics subfield is simple: just don't do it. In contrast, deception has been and has remained fairly commonplace within the political psychology research tradition. This section describes potential advantages and disadvantages of using deception from a methodological and inferential perspective; ethical considerations are not discussed here because of space limitations (for a recent review, see Morton and Williams 2010).

The Lack of Deception in Experimental Economics

Deep-seated opposition to the use of deception has become a feature of various institutions within the economics discipline. It is common for experimental economics laboratories to publicize and enforce bans on deceiving subjects; a strong norm among practitioners and journal editors makes experiments employing deception de facto unpublishable in major economics journals.

Before describing the motivations for these norms, it is worth describing what “deception” means, and does not mean, to experimental economists. A rough distinction can be made between sins of *commission* and sins of *omission*. Describing features of the experimental scenario in a way that is either explicitly dishonest or actively misleading – a sin of commission – would straightforwardly be considered a taboo act of deception by experimental economists. In contrast, a failure fully to describe some features of the experimental scenario – a sin of omission – would not necessarily be counted as a deceptive act. As Hey (1998) puts it, “there is a world of difference between not telling subjects things and telling them the wrong things. The latter is deception, the former is not” (397). Thus, in several studies of public goods provision, experimentalists employ a “surprise re-start,” in which a second, previously unannounced public goods game is played after the completion of the first. So long as subjects are not actively misled by the wording of the experimental protocol, such a procedure is not considered to be deceptive. And, of course, few scholars would argue that it is necessary explicitly to inform subjects about the purpose of the study in which they are taking part.

What arguments do experimental economists present against the use of deception? Both Bonetti (1998) and Morton and Williams (2010) cite Ledyard (1995) as offering a standard line of reasoning:

It is believed by many undergraduates that psychologists are intentionally deceptive in most experiments. If undergraduates believe the same about economists, we have lost control. It is for this reason that modern experimental economists have been carefully nurturing a reputation for absolute honesty in all their experiments... (I)f the data are to be valid. Honesty in procedures is absolutely crucial. Any deception can be discovered and contaminate a subject pool not only for the experimenter but for others. Honesty is a methodological public good and deception is not contributing (134).

At the heart of this case is the fear that the use of deception will lead to a loss of experimental control; as we have seen, many features of economics-style experimentation, including the use of

stylized experimental scenarios and the use of monetary incentives, are designed to help maintain experimental control of different kinds. Hey (1991) articulates the specific nature of this concern:

(I)t is crucially important that economics experiments actually do what they say they do and that subjects believe this. I would not like to see experiments in economics degenerate to the state witnessed in some areas of experimental psychology where it is common knowledge that the experimenters say one thing and do another...[O]nce subjects start to distrust the experimenter, then the tight control that is needed is lost (171-3).

This kind of concern about experimental control is quite natural given the typical nature of research questions in experimental economics. As noted above, most economics experiments either test the predictions of game-theoretic models or explore the nature of behavior in game-theoretic settings. Crucially, the most common concepts of equilibrium in games, from which predictions are derived, assume that actors share common knowledge about basic features of the game being played. Of course, experimental subjects learn about “the rules of the game” through the experimenter. If researchers indeed do, as Hey fears, develop a reputation for employing deception in their experiments, then subjects may develop heterogeneous beliefs about what is really going on in the laboratory – while also being aware that other subjects are doing the same. At the end of the day, subjects could well effectively find themselves playing a wholly different game than the one the experimenter had intended. The conjectures within subjects’ minds about the true nature of the game would, of course, be essentially unknowable not only to one another, but also to the analyst.

Ledyard’s opinion also reflects a common viewpoint among experimental economists: namely that a lab can benefit from maintaining a reputation for transparency with its subject pool. Such a reputation, it is argued, could quickly be squandered if deception takes place in the laboratory; the subject pool may become “tainted” with subjects who have either themselves experienced deception or who have been told about it by friends.

This argument is reasonable, but the question it bears on is ultimately an empirical one. Relatively little systematic research has explored this point, but there is some evidence that the experience of deception in the laboratory may affect individual subjects' propensities to participate in future experiments as well as their behavior in future experiments (Jamison, Karlan, and Schechter 2008). To my knowledge, there has been no systematic research into a related question: the extent to which experimental economics laboratories who do ban deception actually attain the reputations to which they aspire – that is, to what extent subjects are aware of lab policies on deception in general or actually believe that they are never being deceived while taking part in particular experiments in no-deception labs. Economists' arguments about the sanctity of subject pools further tend to presuppose that psychology departments do not exist, or at least that they draw from a disjoint set of participants. If psychology and economics labs operate simultaneously at the same university, to what extent do undergraduate subjects actually perceive them as separate entities, with distinct reputations? Does the physical proximity of the labs to another affect subject perceptions – for example, if they are in the same building as opposed to different buildings? It would appear that such questions remain to be answered.

The Use of Deception in Experimental Political Psychology

In contrast, the use of deception is quite common in political psychology, as it is in social psychology. As we have seen before, the reasons for this difference can be understood as springing from the distinctive natures of inquiry and theory testing in the two schools.

Importantly, the ability to induce common knowledge of an experimental scenario within a group of subjects is usually not nearly so crucial for experiments in the political psychology tradition, which typically do not involve tests of game-theoretic models. This subsection reconsiders the advantages and disadvantages of deception in the context of political psychology research

questions.

One prominent class of examples can be found in the study of political communication, in which scholars quite frequently present subjects with stimuli that are fabricated or falsely attributed. Thus, Brader (2005) presents experimental political advertisements to subjects as though they were genuine ads from a real, ongoing campaign; meanwhile, Druckman and Nelson (2003) present experimental newspaper stories to subjects as though they came from well-known outlets such as the *New York Times*.

In the following paragraphs, I use these articles as examples in discussing potential advantages of deception. Throughout, I take as the salient alternative an otherwise identical experimental design in which the same stimuli are presented to subjects, but explicitly labeled as “hypothetical” campaign ads, newspaper stories, etc. Of course, in certain circumstances different counterfactual designs might also reasonably be considered.

In judging the potential usefulness of deception, then, a natural question to ask is whether an individual’s mode of psychological engagement with a stimulus depends on whether that stimulus is framed as being “real” as opposed to hypothetical. If the answer to this question is “yes” – and if this would make a substantial enough difference for measurements of the quantities of interest – then at the least a benefit from deception will have been identified. Ultimately, of course, in any given setting it is an empirical question whether the answer will be “yes” or “no.” To my knowledge, however, no systematic studies have been carried out measuring the effects, if any, of choosing deceptive as opposed to explicitly hypothetical experimental scenarios.

Taking Druckman and Nelson’s design as an example, though, it at least seems plausible that the difference may sometimes be considerable. An individual picking up what she believes

to be an article from the *New York Times* will respond to frames and other cues in a way that depends directly on her relationship with the *New York Times* – her sense of the newspaper’s reliability, the fit of its ideology with her own, and so forth. In contrast, a hypothetical exercise of the form “suppose the *New York Times* reported...” could insert in the subject’s mind a mysterious intermediary between the newspaper and the subject. Who is it that is doing this supposing, and what are they up to? Alternatively, the subject may simply attend differently to the article, paying it less heed or greeting it with less trust, if she knows from the offset that it is a fiction. Under such circumstances, it would not be unreasonable to suppose that a given article might have less of an effect than it would have had it been described as a “real” article. While economically inclined scholars might tend to doubt whether experiments employing deception can ever gain a full measure of experimental control, it is arguable in this setting that more control might be lost with an explicitly hypothetical stimulus than with a deceptive one. Whether this is true, of course, depends on the extent to which subjects were actually successfully deceived. This, however, is the sort of question that can often be addressed through the use of simple manipulation checks by the experimenter. At least in this example, the treatment effects in Druckman and Nelson’s findings very strongly suggest that the deceptive manipulation did indeed have the desired effect on subjects.

In a similar way, it seems plausible that deception may be a useful element of Brader’s design. In part this is arguable because of the nature of some of Brader’s dependent variables. Among other things, Brader shows that the use of music in contrived political advertising can affect subjects’ self-reported level of inclination to seek more information about an election campaign; the idea of asking subjects to report their level of inclination to seek more information about a hypothetical campaign that means nothing to them seems straightforwardly problematic.

These examples suggest that deception may offer access to certain research questions that would remain inaccessible in its absence. Psychologists also claim that deception may be necessary at times to conceal the purpose of an experiment from subjects (Bortolotti and Mameli 2006); psychologists are frequently concerned about the possibility of “Hawthorne effects,” through which subjects attempt to meet whatever they perceive the experimenter’s expectations to be. Such effects can be particularly worrisome in sensitive research areas, such as the study of racial politics.

Finally, it could be argued that the use of deception can sometimes strengthen the inferences that are possible from a given piece of research. Among the most famous experiments in social psychology is the seminal Milgram (1974) experiment on obedience and authority. In the experiment, subjects were deceived into believing that they could, with the twist of a knob, deliver electric shocks of increasing magnitude to another person; an authority figure urged subjects to deliver such shocks in the context of a staged scenario. In the end, a large fraction of subjects did conform to the authority figure’s commands, to the point of delivering highly dangerous voltages.

This is a rather shocking result, one which had a profound effect on the study of authority and on social psychology more generally. Its power, of course, comes from our sense that subjects really did believe – at least to some considerable extent – that their actions were causing actual bodily harm to another human being. An otherwise comparable study involving an explicitly hypothetical scenario would, for obvious reasons, have been far less convincing, even if it yielded the same results. It could be easily argued that Milgram’s act of deception was central to the lasting influence of Milgram’s study.

References

- Bassi, Anna, Rebecca Morton, and Kenneth Williams. 2008. "The Effects of Identities, Incentives, and Information on Voting." Working Paper.
- Bonetti, Shane. 1998. "Experimental Economics and Deception." *Journal of Economic Psychology* 19: 377-95.
- Bortolotti, Lisa, and Matteo Mameli. 2006. "Deception in Psychology: Moral Costs and Benefits of Unsought Self-Knowledge." *Accountability in Research* 13: 259-75.
- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49: 388-405.
- Camerer, Colin F. 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.
- Camerer, Colin, and Robin Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19: 7-41.
- Chen, Yan, and Sherry Xin Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99: 431-57.
- Dickson, Eric S. In press. "Leadership, Followership, and Beliefs About the World: An Experiment." *British Journal of Political Science*.
- Dickson, Eric S., Catherine Hafer, and Dimitri Landa. 2008. "Cognition and Strategy: A Deliberation Experiment." *Journal of Politics* 70: 974-89.
- Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47: 729-45.
- Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90: 980-94.
- Gneezy, Uri, and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* 115: 791-810.
- Griggs, Richard A., and James R. Cox. 1982. "The Elusive Thematics Material Effect in Wason's Selection Task." *British Journal of Psychology* 73: 407-20.
- Hey, John D. 1998. "Experimental Economics and Deception: A Comment." *Journal of Economic Psychology* 19: 397-401.
- Jamison, Julian, Dean Karlan, and Laura Schechter. 2008. "To Deceive or Not to Deceive: The Effect of Deception on Behavior in Future Laboratory Experiments." *Journal of Economic Behavior and Organization* 68: 477-88.
- Kagel, John H., and Alvin E. Roth. 1995. *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics*, eds., John H. Kagel, and Alvin E. Roth. Princeton, NJ: Princeton University Press.
- Levine, David K., and Thomas R. Palfrey. 2007. "The Paradox of Voter Participation: A Laboratory Study." *American Political Science Review* 101: 143-58.
- Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge: Cambridge University Press.
- McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10: 325-42.
- Milgram, Stanley. 1974. *Obedience and Authority*. New York: Harper and Row.
- Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.
- Prior, Markus, and Arthur Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing

- Quick Recall and Political Learning Skills.” *American Journal of Political Science* 52: 169-183.
- Tajfel, Henri, Michael Billig, R. Bundy, and Claude L. Flament. 1971. “Social Categorization and Inter-Group Behavior.” *European Journal of Social Psychology* 1: 149–77.
- Tajfel, Henri, and John Turner. 1986. “The Social Identity Theory of Intergroup Behavior.” In *Psychology of Intergroup Relations*, eds., Stephen Worchel, and William G. Austin. Chicago: Nelson-Hall.
- Titmuss, Richard M. 1970. *The Gift Relationship: From Human Blood to Social Policy*. London: George Allen and Unwin.
- Wason, P.C. 1968. “Reasoning About a Rule.” *Quarterly Journal of Experimental Psychology* 20: 273-81.

II. The Development of Experiments in Political Science

6. Laboratory Experiments in Political Science

Shanto Iyengar

Until the middle of the twentieth century, the discipline of political science was primarily qualitative – philosophical, descriptive, legalistic, and typically reliant on case studies that failed to probe causation in any measurable way. The word “science” was not entirely apt.

In the 1950s, the discipline was transformed by the behavioral revolution, spearheaded by advocates of a more social scientific, empirical approach. Even though experimentation was the *sine qua non* of research in the hard sciences and in psychology, the method remained a mere curiosity among political scientists. For behavioralists interested in individual-level political behavior, survey research was the methodology of choice on the grounds that experimentation could not be used to investigate real-world politics (for more detailed accounts of the history of experimental methods in political science, see Bositis and Steinel 1987; Kinder and Palfrey 1993; Green and Gerber 2003). The consensus view was that laboratory settings were too artificial and that experimental subjects were too unrepresentative of any meaningful target population for experimental studies to be valid. Further, many political scientists viewed experiments -- which typically necessitate the deception of research subjects -- as an inherently unethical methodology.

The bias against experimentation began to weaken in the 1970s when the emerging field of political psychology attracted a new constituency for interdisciplinary research. Laboratory experiments gradually acquired the aura of legitimacy for a small band of scholars working at the intersection of the two disciplines.¹ Most of these scholars focused on the areas of political behavior, public opinion and mass communication, but there were also experimental forays into

the fields of international relations and public choice (Hermann and Herman 1967; Riker 1967). Initially, these researchers faced significant disincentives to applying experimental methods -- most importantly, research-based on experiments was unlikely to see the light of day simply because there were no journals or conference venues that took this kind of work seriously.

The first major breakthrough for political scientists interested in applying the experimental method occurred with the founding of the journal *Experimental Study of Politics (ESP)* in 1970. The brainchild of the late James Dyson (then at Florida State University) and Frank Scioli (then at Drew University and now at the National Science Foundation), *ESP* was founded as a boutique journal dedicated exclusively to experimental work. The coeditors and members of their editorial board were committed behavioralists who were convinced that experiments could contribute to more rigorous hypothesis testing and thereby to theory building in political science (Scioli 2009). As stated by the editors, the mission of the journal was to “provide an outlet for the publication of materials dealing with experimental research in the shortest possible time, and thus to aid in rapid dissemination of new ideas and developments in political research and theory (Scioli 2009).”

ESP served as an important, albeit specialized, outlet for political scientists interested in testing propositions about voting behavior, presidential popularity, mass communication and campaigns, or group decision making. The mere existence of a journal dedicated to experimental research (with a masthead featuring established scholars from highly ranked departments)ⁱⁱ provided a credible signal to graduate students and junior faculty (this author included) that it might just be possible to publish (rather than perish) and build a career in political science on the basis of experimental research.

Although *ESP* provided an important “foot in the door,” the marginalized status of experiments in political science persisted during the 1970s. Observational methods, most notably, survey research, dominated experimentation even among the practitioners of political psychology. One obvious explanation for the slow growth rate in experimental research was the absence of necessary infrastructure. Experiments are typically space-, resource-, and labor-intensive. Laboratories with sophisticated equipment or technology, and trained staff were nonexistent in political science departments, with one notable exception, namely, the State University of New York at Stony Brook.

When SUNY–Stony Brook was established in the early 1960s, the political science department was given a mandate to specialize in behavioral research and experimental methods. In 1978, the department moved into a new building with state-of-the-art experimental facilities including laboratories for measuring psychophysiological responses (modeled on the psychophysiology labs at Harvard), cognitive or information-processing labs for tracking reaction time, and an array of social psychological labs modeled on the lab run by the eminent Columbia psychologist Stanley Schachter.ⁱⁱⁱ Once these labs were put to use by the several prominent behaviorists who joined the Stony Brook political science faculty in the early 1970s (including Milton Lodge, Joseph Tanenhaus, Bernard Tursky and John Wahlke), the department would play a critical role in facilitating and legitimizing experimental research.^{iv}

The unavailability of suitable laboratory facilities was but one of several obstacles facing the early experimentalists. An equally important challenge was the recruitment of experimental subjects. Unlike the field of psychology, where researchers could draw on a virtually unlimited captive pool of student subjects, experimentalists in political science had to recruit volunteer

(and typically unpaid) subjects on their own initiative. Not only did this add to the costs of conducting experiments, it also ensured that the resulting samples would be far from typical.

In the early 1980s, experimental methods were of growing interest to researchers in several subfields of the discipline. Don Kinder and I were fortunate enough to receive generous funding from the National Institutes of Health and the National Science Foundation for a series of experiments designed to assess the effects of network news on public opinion. These experiments, most of which were administered in a dilapidated building on the Yale campus, revealed that contrary to the conventional wisdom at the time, network news exerted significant effects on the viewing audience. We reported the full set of experimental results in *News That Matters* (Iyengar and Kinder 1987). The fact that the University of Chicago Press published a book based exclusively on experiments demonstrated that experiments could be harnessed to address questions of political significance. That the book was generally well received demonstrated that a reliance on experimental methodology was no longer stigmatized in political science.

By the end of the 1980s, laboratory experimentation had become sufficiently recognized as a legitimate methodology in political science for mainstream journals to regularly publish papers based on experiments (see Druckman, Green, Kuklinski, and Lupia 2006). Despite the significant diffusion of the method, however, two key concerns contributed to continued scholarly skepticism. First, experimental settings were deemed lacking in mundane realism -- the experience of participating in an experiment was sufficiently distinctive to preclude generalizing the results to real-world settings. Second, student-based and other volunteer subject pools were considered unrepresentative of any broader target population of interest (i.e. registered voters or individuals likely to engage in political protest). To this day, the problem of external validity or

questionable generalizability continues to impede the adoption of experimentation in political science.

In this chapter I begin by describing the inherent strengths of the experiment as a basis for causal inference, using recent examples from my own work in political communication. I argue that the downside of experiments -- the standard “too artificial” critique -- has been weakened by several developments, including the use of more realistic designs that move experiments outside of a laboratory environment and the technological advances associated with the Internet. The online platform is itself now entirely realistic (given the extensive daily use of the Internet by ordinary individuals); it also allows researchers to overcome the previously profound issue of sampling bias. All told, these developments have gone a long way toward alleviating concerns about the validity of experimental research -- so much so that I would argue that experiments now represent a dominant methodology for researchers in several fields of political science.

1. Causal Inference: The Strength of Experiments

The principal advantage of the experiment over the survey or other observational methods -- and the focus of the discussion that follows -- is the researcher’s ability to isolate and test the effects of specific components of specific causal variables. Consider the case of political campaigns. At the aggregate level, campaigns encompass a concatenation of messages, channels, and sources, all of which may influence the audience, often in inconsistent directions. The researcher’s task is to identify the potential causal mechanisms and delineate the range of their relevant attributes. Even at the relatively narrow level of campaign advertisements, for instance, there are virtually an infinite number of potential causal forces, both verbal and visual. What was it about the infamous "Willie Horton" advertisement that is thought to have moved so many

American voters away from Michael Dukakis during the 1988 presidential campaign? Was it, as widely alleged during the campaign, that Horton was African American (see Mendelberg 2001)? Or was it the violent and brutal nature of his described behavior, the fact that he was a convict, or something else entirely? Experiments make it possible to isolate the attributes of messages that move audiences, whether these are text-based or nonverbal cues. Surveys, on the other hand, can only provide indirect evidence on self-reported exposure to the causal variable in question.

Of course, experiments not only shed light on treatment effects but also enable researchers to test more elaborate hypotheses concerning moderator variables by assessing interactions between the treatment factors and relevant individual-difference variables. In the case of persuasion, for instance, not all individuals are equally susceptible to incoming messages (see Zaller 1992). In the case of the 1988 campaign, perhaps Democrats with a weak party affiliation and strong sense of racial prejudice were especially likely to sour on Governor Dukakis in the aftermath of exposure to the Horton advertisement.

In contrast with the experiment, the inherent weaknesses of the survey design for isolating the effects of causal variables have been amply documented. In a widely cited paper, Hovland (1959) identified several problematic artifacts of survey-based studies of persuasion, including unreliable measures of media exposure. Clearly, exposure is a necessary precondition for media influence, but self-reported exposure to media coverage is hardly equivalent to actual exposure. People have notoriously weak memories for political experiences (see, for instance, Pierce and Lovrich 1982; Bradburn, Rips and Shevell 1987). In the Ansolabehere and Iyengar experiments on campaign advertising (which spanned the 1990, 1992, and 1994 election cycles), over fifty percent of the participants who were exposed to a political advertisement were unable, *some thirty minutes later*, to recall having seen the advertisement (Ansolabehere and Iyengar

1998). In a more recent example, Vavreck found that nearly half of a control group not shown a public service message responded either that they couldn't remember or that they *had* seen it (Vavreck 2007; also see Prior 2003). Errors of memory also compromise recall-based measures of exposure to particular news stories (see Gunther 1987) or news sources (Price and Zaller 1993). Of course, since the scale of the error in self-reports tends to be systematic (respondents are prone to overstate their media exposure), survey-based estimates of the effects of political campaigns are necessarily attenuated (Bartels 1993; Prior 2003).

An even more serious obstacle to causal inference in the survey context is that the indicators of the causal variable (self-reported media exposure in most political communication studies) are typically endogenous to a host of outcome variables researchers seek to explain (such as candidate preference). Those who claim to read newspapers or watch television news on a regular basis, for instance, differ systematically (in ways that matter to their vote choice) from those who attend to the media less frequently. This problem has become especially acute in the aftermath of the revolution in "new media." In 1968, approximately seventy-five percent of the adult viewing audience watched one of the three network evening newscasts, but by 2008 the combined audience for network news was less than thirty-five percent of the viewing audience. In 2008, the only people watching the news were those with a keen interest in politics; most everyone else had migrated to more entertaining, nonpolitical programming alternatives (Prior 2007).

The endogeneity issue has multiple ramifications for political communication research. First, consider those instances where self-reported media exposure is correlated with political predispositions but actual exposure is not. This is generally the case with televised political advertising. Most voters encounter political ads unintentionally, in the course of watching their

preferred television programs in which the commercial breaks contain a heavy dose of political messages. Thus, actual exposure is idiosyncratic (based on the viewer's preference for particular television programs), while self-reported exposure is based on political predispositions.

The divergence in the antecedents of self-reported exposure has predictable consequences for research on effects. In experiments that manipulated the tone of campaign advertising, Ansolabehere and Iyengar (1995) found that actual exposure to negative messages demobilized voters, i.e., discouraged intentions to vote. However, on the basis of self-reports, survey researchers concluded that exposure to negative campaign advertising stimulated turnout (Wattenberg and Briars 1999). But was it recalled exposure to negative advertising that prompted turnout, or was the greater interest in campaigns among likely voters responsible for their higher level of recall? When recall of advertising in the same survey was treated as endogenous to vote intention and the effects reestimated using appropriate two-stage methods, the sign of the coefficient for recall was reversed: those who recalled negative advertisements were less likely to express an intention to vote (see Ansolabehere, Iyengar and Simon 1999).^v Unfortunately, most survey-based analyses fail to disentangle the reciprocal effects of self-reported exposure to the campaign and partisan attitudes and behaviors. As this example suggests, in cases where actual exposure to the treatment is less selective than self-reported exposure, self-reports may prove especially biased.

In other scenarios, however, the tables may be turned and the experimental researcher may actually be at a disadvantage. Actual exposure to political messages in the real world is typically not analogous to random assignment. People who choose to participate in experiments on campaign advertising are likely to differ from those who choose to watch ads during campaigns (for a general discussion of the issue, see Gaines and Kuklinski 2008). Unlike

advertisements, news coverage of political events can be avoided by choice, meaning that exposure is limited to the politically engaged strata. Thus, as Hovland (1959) and others (Heckman and Smith 1995) have pointed out, manipulative control actually weakens the ability to generalize to the real world where exposure to politics is typically voluntary. In these cases, it is important that the researcher use designs that combine manipulation with self-selected exposure.

One other important aspect of experimental design that contributes to strong causal inference is the provision of procedures to guard against the potential contaminating effects of “experimental demand” -- cues in the experimental setting or procedures that convey to participants what is expected of them (for the classic account of demand effects, see Orne 1962). Demand effects represent a major threat to internal validity: participants are motivated to respond to subtle cues in the experimental context suggesting what is wanted of them rather than to the experimental manipulation itself.

The standard precautions against experimental demand include disguising the true purpose of the study by providing participants with a plausible (but false) description,^{vi} using relatively unobtrusive outcome measures, and maximizing the “mundane realism” of the experimental setting so that participants are likely to mimic their behavior in real-world settings. (I will return to the theme of realism in the section on generalizability.)

In the campaign advertising experiments described I describe in the following section, for instance, the researchers inserted manipulated political advertisements into the ad breaks of the first ten minutes of a local newscast. Study participants were diverted from the researchers’ intent by being misinformed that the study was about “selective perception of television news.” The use of a design in which the participants answered the survey questions only after exposure

to the treatment further guarded against the possibility that they might see through the cover story and infer the true purpose of the study.

In summary, the fundamental advantage of the experimental approach -- and the reason experimentation is the methodology of choice in the hard sciences -- is the researcher's ability to isolate causal variables, which constitute the basis for experimental manipulations. In the next section, I describe manipulations designed to assess the effects of negative advertising campaigns, racial cues in television news coverage of crime, and the physical similarity of candidates to voters.

Negativity in Campaign Advertising

At the very least, establishing the effects of negativity in campaign advertising on voters' attitudes requires varying the tone of a campaign advertisement while holding all other attributes of the advertisement constant. Despite the significant increase in scholarly attention to negative advertising, few studies live up to this minimal threshold of control (for representative examples of survey-based analyses see Finkel and Geer 1998; Freedman and Goldstein 1999; Kahn and Kenney 1999.)

In a series of experiments conducted by Ansolabehere and Iyengar, the researchers manipulated negativity by unobtrusively varying the text (soundtrack) of an advertisement while preserving the visual backdrop (Ansolabehere and Iyengar 1995). The negative version of the message typically placed the sponsoring candidate on the unpopular side of some salient policy issue. Thus, during the 1990 California gubernatorial campaign between Pete Wilson (Republican) and Dianne Feinstein (Democrat), the treatment ads positioned the candidates either as opponents or proponents of offshore oil drilling and thus as either friends or foes of the environment. This manipulation was implemented by simply substituting the word "yes" for the

word “no.” In the positive conditions, the script began as follows: “When federal bureaucrats asked for permission to drill for oil off the coast of California, Pete Wilson/Dianne Feinstein said no” In the negative conditions, we substituted “said yes” for “said no.” An additional substitution was written into the end of the ad when the announcer stated that the candidate in question would either work to “preserve” or “destroy” California’s natural beauty. Given the consensual nature of the issue, negativity could be attributed to candidates who claimed their opponent was soft on polluters.^{vii}

The results from these studies (which featured gubernatorial, mayoral, senatorial, and presidential candidates) indicated that participants exposed to negative rather than positive advertisements were less likely to say they intended to vote. The demobilizing effects of exposure to negative advertising were especially prominent among viewers who did not identify with either of the two political parties (see Ansolabehere and Iyengar 1995).

Racial Cues in Local News Coverage of Crime

As any regular viewer of television will attest to, crime is a frequent occurrence in broadcast news. In response to market pressures, television stations have adopted a formulaic approach to covering crime, an approach designed to attract and maintain the highest degree of audience interest. This “crime script” suggests that crime is invariably violent and those who perpetrate crime are disproportionately nonwhite. Because the crime script is encountered so frequently (several times each day in many cities) in the course of watching local news, it has attained the status of common knowledge. Just as we know full well what happens when one walks into a restaurant, we also know -- or at least think we know -- what happens when crime occurs (Gilliam and Iyengar 2000).

In a series of recent experiments, researchers have documented the effects of both

elements of the crime script on audience attitudes (see Gilliam et al. 1996; Gilliam, Valentino and Beckman 2002). For illustrative purposes, I focus here on the racial element. In essence, these studies were designed to manipulate the race/ethnicity of the principal suspect depicted in a news report while maintaining all other visual characteristics. The original stimulus consisted of a typical local news report, which included a close-up still mug shot of the suspect. The picture was digitized, adjusted to alter the perpetrator's skin color, and then reedited into the news report. As shown in Figure 6-1, beginning with two different perpetrators (a white male and a black male), the researchers were able to produce altered versions of each individual in which their race was reversed, but all other features remained identical. Participants who watched the news report in which the suspect was thought to be nonwhite expressed greater support for punitive policies (e.g., imposition of "three strikes and you're out" remedies, treatment of juveniles as adults, and support for the death penalty). Given the precision of the design, these differences in the responses of the subjects exposed to the white or black perpetrators could only be attributed to the perpetrator's race (see Gilliam and Iyengar 2000).

[Figure 6-1 about here]

Facial Similarity as a Political Cue

A consistent finding in the political science literature is that voters gravitate to candidates who most resemble them on questions of political ideology, issue positions, and party affiliation. But what about physical resemblance; are voters also attracted to candidates who look like them?

Several lines of research suggest that physical similarity in general, and facial similarity in particular, is a relevant criterion for choosing between candidates. Thus, frequency of exposure to any stimulus -- including faces -- induces a preference for that stimulus over other, less familiar stimuli (Zajonc 2001). Moreover, evolutionary psychologists argue that physical

similarity is a kinship cue and there is considerable evidence that humans are motivated to treat their kin preferentially (see, for instance, Burnstein, Crandall, and Kitayama 1994; Nelson 2001).

In order to isolate the effects of facial similarity on voting preferences, researchers obtained digital photographs of 172 registered voters selected at random from a national Internet panel (for details on the methodology, see Bailenson et al. 2009). Participants were asked to provide their photographs approximately three weeks in advance of the 2004 presidential election. One week before the election, these same participants were asked to participate in an online survey of political attitudes that included a variety of questions about the presidential candidates (President George W. Bush and Senator John Kerry). The screens for these candidate questions included photographs of the two candidates displayed side by side. Within this split-panel presentation, participants had their own face either morphed with Bush or Kerry at a ratio of sixty percent of the candidate and forty percent of themselves.^{viii} Figure 6-2 shows two of the morphs used in this study.

[Figure 6-2 about here]

The results of the face morphing study revealed a significant interaction between facial similarity and strength of the participant's party affiliation. Among strong partisans, the similarity manipulation had no effect; these voters were already convinced of their vote choice. But weak partisans and independents -- whose voting preferences were not as entrenched -- moved in the direction of the more similar candidate (see Bailenson et al. 2009). Thus, the evidence suggests that nonverbal cues can influence voting, even in the most visible and contested of political campaigns.^{ix}

In short, as these examples indicate, the experiment provides unequivocal causal evidence because the researcher is able to isolate the causal factor in question, manipulate its

presence or absence, and hold other potential causes constant. Any observed differences between experimental and control groups, therefore, can only be attributed to the factor that was manipulated.

Not only does the experiment provide the most convincing basis for causal inference, experimental studies are also inherently replicable. The same experimental design can be administered independently by researchers in varying locales with different stimulus materials and subject populations. Replication thus provides a measure of the reliability or robustness of experimental findings across time, space, and relatively minor variations in study procedure.

Since the first published reports on the phenomenon of media priming -- the tendency of experimental participants to weigh issues they have been exposed to in experimental treatments more heavily in their political attitudes -- the effect has been replicated repeatedly. Priming effects now apply to evaluations of public officials and governmental institutions, to vote choices in a variety of electoral contests, to stereotypes, group identities, and any number of other attitudes. Moreover, the finding has been observed across an impressive array of political and media systems (for a recent review of priming research, see Roskos-Ewoldsen, Roskos-Ewoldsen and Carpentier 2005).

2. The Issue of Generalizability

The problem of limited generalizability, long the bane of experimental design, is manifested at multiple levels: the realism of the experimental setting, the representativeness of the participant pool, and the discrepancy between experimental control and self-selected exposure to media presentations.

Mundane Realism

Because of the need for tightly controlled stimuli, the setting in which the typical laboratory experiment occurs is often quite dissimilar from the setting in which subjects ordinarily experience the target phenomenon. Concern over the artificial properties of laboratory experiments has given rise to an increased use of designs in which the intervention is nonobtrusive and the settings more closely reflect ordinary life.⁴

One approach to increasing experimental realism is to rely on interventions with which subjects are familiar. The Ansolabehere/Iyengar campaign experiments were relatively realistic in the sense that they occurred during ongoing campaigns characterized by heavy levels of televised advertising (see Ansolabehere and Iyengar 1995). The presence of a political advertisement in the local news (the vehicle used to convey the manipulation) was hardly unusual or unexpected since candidates advertise most heavily during news programs. The advertisements featured real candidates -- Democrats and Republicans, liberals and conservatives, males and females, incumbents and challengers -- as the sponsors. The materials that made up the experimental stimuli were selected either from actual advertisements used by the candidates during the campaign, or were produced to emulate typical campaign advertisements. In the case of the latter, the researchers spliced together footage from actual advertisements or news reports making the treatment ads representative of the genre. (The need for control made it necessary for the treatment ads to differ from actual political ads in several important attributes, including the absence of music and the appearance of the sponsoring candidate.)

Realism also depends upon the physical setting in which the experiment is administered. Asking subjects to report to a location on a university campus may suit the researcher but may make the experience of watching television equivalent to the experience of visiting the doctor. A

more realistic strategy is to provide subjects with a milieu that closely matches the setting of their home television-viewing environment. The fact that the advertising research lab was configured to resemble a typical living or family room setting (complete with reading matter and refreshments) meant that participants did not need to be glued to the television screen. Instead, they could help themselves to cold drinks, browse through newspapers and magazines, or engage in small talk with fellow participants.^x

A further step toward realism concerns the power of the manipulation (also referred to as experimental realism). Of course, the researcher would like for the manipulation to have an effect. At the same time, it is important that the required task or stimulus not overwhelm the subject (as in the Milgram obedience studies where the task of administering an electric shock to a fellow participant proved overpowering and ethically suspect). In the case of the campaign advertising experiments, we resolved the experimental realism versus mundane realism tradeoff by embedding the manipulation in a commercial break of a local newscast. For each treatment condition, the stimulus ad appeared with other nonpolitical ads and subjects were led to believe that the study was about “selective perception of news,” so they had no incentive to pay particular attention to ads. Overall, the manipulation was relatively small, amounting to thirty seconds of a fifteen-minute videotape.

In general, there is a significant tradeoff between experimental realism and manipulative control. In the aforementioned advertising studies, the fact that subjects were exposed to the treatments in the company of others meant that their level of familiarity with fellow subjects was subject to unknown variation. And producing experimental ads that more closely emulated actual ads (e.g. ads with musical background included and featuring the sponsoring candidate) would necessarily have introduced a series of confounding variables

associated with the appearance and voice of the sponsor. Despite these tradeoffs, however, it is still possible to achieve a high degree of experimental control with stimuli that closely resemble the naturally occurring phenomenon of interest.

Sampling Bias

The most widely cited limitation of experiments concerns the composition of the subject pool (Sears 1986). Typically, laboratory experiments are administered upon captive populations - college students who must serve as guinea pigs in order to gain course credit. College sophomores may be a convenient subject population for academic researchers, but are they comparable to "real people?"^{xi}

In conventional experimental research, it is possible to broaden the participant pool but at considerable cost/effort. Locating experimental facilities at public locations and enticing a quasi-representative sample to participate proves both cost- and labor-intensive. Typical costs include rental fees for an experimental facility in a public area (such as a shopping mall), recruitment of participants, and training and compensation of research staff to administer the experiments. In our local news experiments conducted in Los Angeles in the summer and fall of 1999, the total costs per subject amounted to approximately forty-five dollars. Fortunately, and as I will describe, technology has both enlarged the pool of potential participants and reduced the per capita cost of administering an experimental study.

Today, traditional experimental methods can be rigorously and far more efficiently administered using an online platform. Utilizing the Internet as the experimental site provides several advantages over conventional locales, including the ability to reach diverse populations without geographic limitations. Diversity is important not only to enhance generalizability, but also to mount more elaborate tests of mediator or moderator variables. In experiments featuring

racial cues, for instance, it is imperative that the study participants include a nontrivial number of minorities. Moreover, with the ever-increasing use of the Internet, not only are the samples more diverse but the setting in which participants encounter the manipulation (surfing the Web on their own) is also more realistic.

“Drop-in” Samples

The Political Communication Laboratory (PCL) at Stanford University has been administering experiments over the Internet for nearly a decade. One of the Lab’s more popular online experiments is “whack-a-pol” (<http://pcl.stanford.edu/exp/whack/polm>), modeled on the well-known whack-a-mole arcade game. Ostensibly, the game provides participants with the opportunity to “bash” well-known political figures.

Since going live in 2001, over 2500 visitors have played whack-a-pol. These “drop in” subjects found the PCL site on their own initiative. How does this group compare with a representative sample of adult Americans with home access to the Internet, and a representative sample of all voting-age adults? First, we gauged the degree of divergence between drop-in participants and typical Internet users. The results suggested that participants in the online experiments reasonably approximated the online user population at least with respect to race/ethnicity, education, and party identification. The clearest evidence of selection bias emerged with age and gender. The mean age of study participants was significantly younger and participants were also more likely to be male. The sharp divergence in age may be attributed to the fact that our studies are launched from an academic server that is more likely to be encountered by college students -- and also to the general “surfing” proclivities of younger users. The gender gap is more puzzling and may reflect differences in political interest or greater enthusiasm for online games among males.

The second set of comparisons assesses the overlap between our self-selected online samples and all voting-age adults (these comparisons are based on representative samples drawn by Knowledge Networks 2000). Here the evidence points to a persisting digital divide in the sense that major categories of the population remain underrepresented in online studies. In relation to the broader adult population, our experimental participants were significantly younger, more educated, more likely to be white males, and less apt to identify as a Democrat.

Although these data make it clear that people who participate in online media experiments are no microcosm of the adult population, the fundamental advantage of online over conventional field experiments cannot be overlooked. Conventional experiments recruit subjects from particular locales; online experiments draw subjects from across the country. The Ansolabehere/Iyengar campaign advertising experiments, for example, recruited subjects from a particular area of southern California (greater Los Angeles). The online experiments, in contrast, attracted a sample of subjects from thirty different American states and several countries.

Expanding the Pool of Online Participants

One way to broaden the online subject pool is by recruiting participants from more well-known and frequently visited websites. News sites that cater to political junkies, for example, may be motivated to increase their circulation by collaborating with scholars whose research studies focus on controversial issues. While the researcher obtains data which may be used for scholarly purposes, the website gains a form of interactivity through which the audience may be engaged. Playing an arcade game or watching a brief video clip may pique participants' interest thus encouraging them to return to the site and boosting the news organization's online traffic.

In recent years, PCL has partnered with Washingtonpost.com to expand the reach of online experiments. Studies designed by PCL -- focusing on topics of interest to people who read

Washingtonpost.com -- are advertised on the Website's politics section. Readers who click on a link advertising the study in question are sent directly to the PCL site, where they complete the experiment, and are then returned to Washingtonpost.com. The results from these experiments were then described in a newspaper story and online column. In cases where the results were especially topical (e.g., a study of news preferences showing that Republicans avoided CNN and NPR in favor of Fox News), a correspondent from Washingtonpost.com hosted an online "chat" session to discuss the results and answer questions.

To date, the Washingtonpost.com – PCL collaborative experiments have succeeded in attracting relatively large samples, at least by the standards of experimental research.⁶ Experiments on especially controversial or newsworthy subjects attracted a high volume of traffic (on some days exceeding 500). In other cases, the rate of participation slowed to a trickle, resulting in a longer period of time to gather the data.

Sampling from Online Research Panels

Even though drop-in online samples provide more diversity than the typical college sophomore sample, they are obviously biased in several important respects. Participants from Washingtonpost.com, for instance, included very few conservatives or Republicans. Fortunately, it is now possible to overcome issues of sampling bias -- assuming the researcher has access to funding -- by administering online experiments to representative samples. In this sense, the lack of generalizability associated with experimental designs is largely overcome.

Two market research firms have pioneered the use of web-based experiments with fully representative samples. Not surprisingly, both firms are located in the heart of Silicon Valley. The first is Knowledge Networks based in Menlo Park, and the second is Polimetrix (recently purchased by the UK polling company of YouGov) based in Palo Alto.

Knowledge Networks has overcome the problem of selection bias inherent to online surveys (which reach only that proportion of the population that is both online and inclined to participate in research studies) by recruiting a nationwide panel through standard telephone methods. This representative panel (including over 150,000 Americans between the ages of sixteen and eighty-five) is provided free access to the Internet via a WebTV. In exchange, panel members agree to participate (on a regular basis) in research studies being conducted by Knowledge Networks. The surveys are administered over the panelist's WebTV. Thus, in theory Knowledge Networks can deliver samples that meet the highest standards of probabilistic sampling. In practice, because their panelists have an obligation to participate, Knowledge Networks also provides relatively high response rates (Dennis, Li and Chatt 2004).

Polimetrix uses a novel matching approach to the sampling problem. In essence, they extract a quasi-representative sample from large panels of online volunteers. The process works as follows. First, Polimetrix assembles a very large pool of opt-in participants by offering small incentives for study participation (e.g. the chance of winning an iPod). As of November of 2007, the number of Polimetrix panelists exceeded 1.5 million Americans. In order to extract a representative sample from this pool of self-selected panelists, Polimetrix uses a two-step sampling procedure. First, they draw a conventional random sample from the target population of interest (i.e. registered voters). Second, for each member of the target sample, Polimetrix substitutes a member of the opt-in panel who is similar to the corresponding member of the target sample on a set of demographic characteristics such as gender, age, and education. In this sense, the matched sample consists of respondents who represent the respondents in the target sample. Rivers (2006) describes the conditions under which the matched sample approximates a true random sample.

The Polimetrix samples have achieved impressive rates of predictive validity, thus bolstering the claims that matched samples emulate random samples.^{xii} In the 2005 California special election, Polimetrix accurately predicted the public's acceptance or rejection of all seven propositions (a record matched by only one other conventional polling organization) with an average error rate comparable to what would be expected given random sampling (Rivers and Bailey 2009).

3. Conclusion

The standard comparison of experiments and surveys favors the former on the grounds of precise causal inference and the latter on the grounds of greater generalizability. As I have suggested, however, traditional experimental methods can be effectively and just as rigorously replicated using online strategies. Web experiments eliminate the need for elaborate lab space and resources; all that is needed is a room with a server. These experiments have the advantage of reaching a participant pool that is more far flung and diverse than the pool relied on by conventional experimentalists. Online techniques also permit a more precise targeting of recruitment procedures so as to enhance participant diversity. Banner ads publicizing the study and the financial incentives for study participants can be placed in portals or sites that are known to attract underrepresented groups. Female subjects or African Americans, for instance, could be attracted by ads placed in sites tailored to their interests. Most recently, the development of online research panels has made it possible to administer experiments on broad cross-sections of the American population. All told, these features of web experiments go a long way toward neutralizing the generalizability advantage of surveys.

Although web experiments are clearly a low cost, effective alternative to conventional experiments, they are hardly applicable to all arenas of behavioral research. Most notably, web-

based experiments provide no insight into group dynamics or interpersonal influence. Web use is typically a solitary experience and web experiments are thus entirely inappropriate for research that requires placing individuals in some social or group milieu (e.g. studies of opinion leadership or conformity to majority opinion).

A further frontier for web experimentalists will be cross-national research. Today, experimental work in political science is typically reliant on American stimuli and American subjects. The present lack of cross-national variation in the subject pool makes it impossible to contextualize American findings,^{xiii} and also means that the researcher is unable to rule out a family of alternative explanations for any observed treatment effects having to do with subtle interactions between culture and treatment (see Juster et al. 2001). Happily, the rapidity with which public access to the web has diffused on a global basis now makes it possible to launch online experiments on a cross-national basis. Fully operational online opt-in research panels are already available in many European nations including Belgium, Britain, Denmark, Finland, Germany, the Netherlands, Norway, and Sweden. Efforts to establish and support infrastructure for administering and archiving cross-national laboratory experiments are underway at several universities including the Nuffield Centre for Experimental Social Sciences and the Zurich Program in the Foundations of Human Behavior.^{xiv} I suspect that by 2015, it will be possible to deliver online experiments to national samples in most industrialized nations. Of course, given the importance of economic development to web access, cross-national experiments administered online -- at least in the near term -- will be limited to the “most similar systems” design.

In closing, it is clear that information technology has removed the traditional barriers to experimentation in political science, including the need for lab space, convenient access to

diverse subject pools, and skepticism over the generalizability of findings. The web makes it possible to administer realistic experimental designs on a world-wide scale with a relatively modest budget. Given the advantages of online experiments, I expect a bright future for laboratory experiments in political science.

References

- Ansolabehere, Stephen D., and Shanto Iyengar. 1995. *Going Negative: How Political Ads Shrink and Polarize the Electorate*. New York: Free Press.
- Ansolabehere, Stephen D., and Shanto Iyengar. 1998. "Messages Forgotten: Misreporting in Surveys and the Bias Towards Minimal Effects." Unpublished manuscript, University of California – Los Angeles.
- Ansolabehere, Stephen D., Shanto Iyengar, and Adam Simon. 1999. "Replicating Experiments Using Aggregate and Survey Data." *American Political Science Review* 93: 901-10.
- Bailenson, Jeremy, Shanto Iyengar, Nick Yee, and Nathan Collins. 2009. "Facial Similarity between Candidates and Voters Causes Influence." *Public Opinion Quarterly* 72: 935-61.
- Bartels, Larry. 1993. "Messages Received: The Political Impact of Media Exposure." *American Political Science Review* 87: 267-85.
- Bositis, David A., and Douglas Steinel. 1987. "A Synoptic History and Typology of Experimental Research in Political Science." *Political Behavior* 9: 263-84.
- Bradburn, Norman M., Lance J. Rips, and Stephen K. Shevell. 1987. "Answering Autobiographical Questions: The Impact of Memory and Inference in Surveys." *Science* 236: 157-61.
- Burnstein, Eugene, Christian Crandall, and Shinobu Kitayama. 1994. "Some Neo-Darwinian Decision Rules for Altruism: Weighing Cues for Inclusive Fitness as a Function of the Biological Importance of the Decision." *Journal of Personality and Social Psychology* 67: 773-89.
- Dennis, J. Michael, Rick Li, and Cindy Chatt. 2004. "Benchmarking Knowledge Networks' Web-Enabled Panel Survey of Selected GSS Questions Against GSS In-Person Interviews." Knowledge Networks Technical Report.
- Druckman, James N., Donald P. Green, James H. Kuklinski, J., and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100: 627-35.

- Finkel, Steven E., and John G. Geer. 1998. "A Spot Check: Casting Doubt on the Demobilizing Effect of Attack Advertising." *American Journal of Political Science* 42: 573-95.
- Freedman, Paul, and Kenneth Goldstein. 1999. "Measuring Media Exposure and the Effects of Negative Campaign Ads." *American Journal of Political Science* 43: 1189-208.
- Gaines, Brian J., and James H. Kuklinski. 2008. "A Case for Including Self-Selection Alongside Randomization in the Assignment of Experimental Treatments." Presented at the annual meeting of the Midwestern Political Science Association, Chicago, IL.
- Gilliam, Franklin Jr., and Shanto Iyengar. 2000. "Prime Suspects: The Influence of Local Television News on the Viewing Public." *American Journal of Political Science* 44: 560-73.
- Gilliam, Franklin Jr., Shanto Iyengar, Adam Simon, and Oliver Wright. 1996. "Crime in Black and White: The Violent, Scary World of Local News." *Harvard International Journal of Press/Politics* 1: 6-23.
- Gilliam, Franklin Jr., Nicholas A. Valentino, and Matthew Beckman. 2002. "Where You Live and What You Watch: The Impact of Racial Proximity and Local Television News on Attitudes About Race and Crime." *Political Research Quarterly* 55: 755-80.
- Green, Donald P., and Alan S. Gerber. 2003. "The Under-Provision of Experiments in Political and Social Science." *Annals of the American Academy of Political and Social Science* 589: 94-112.
- Gunther, Barrie. 1987. *Poor Reception: Misunderstanding and Forgetting Broadcast News*. Hillsdale, NJ: Lawrence Erlbaum.
- Heckman, James J., and Jeffrey P. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9: 85-110.
- Hermann, Charles F., and Margaret G. Hermann. 1967. "An Attempt to Simulate the Outbreak of World War I." *American Political Science Review* 61: 400-16.
- Hill, Seth, Lo, James, Vavreck, Lynn, and John R. Zaller. 2007. "The Opt-in Internet Panel: Survey Mode, Sampling Methodology and the Implications for Political Research." Unpublished paper, University of California-Los Angeles. Retrieved from <http://web.mit.edu/polisci/portl/cces/material/HillLoVavreckZaller2007.pdf>.
- Hovland, Carl I. 1959. "Reconciling Conflicting Results Derived From Experimental and Survey Studies of Attitude Change." *American Psychologist* 14: 8-17.
- Iyengar, Shanto, and Donald R. Kinder. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.
- Juster, Thomas F. Richard Blundell, Richard V. Burkhauser, Graziella Caselli, Linda P. Fried, Albert I. Hermalin, Robert L. Kahn, Arie Kapteyn, Michael Marmot, Linda G. Martin,

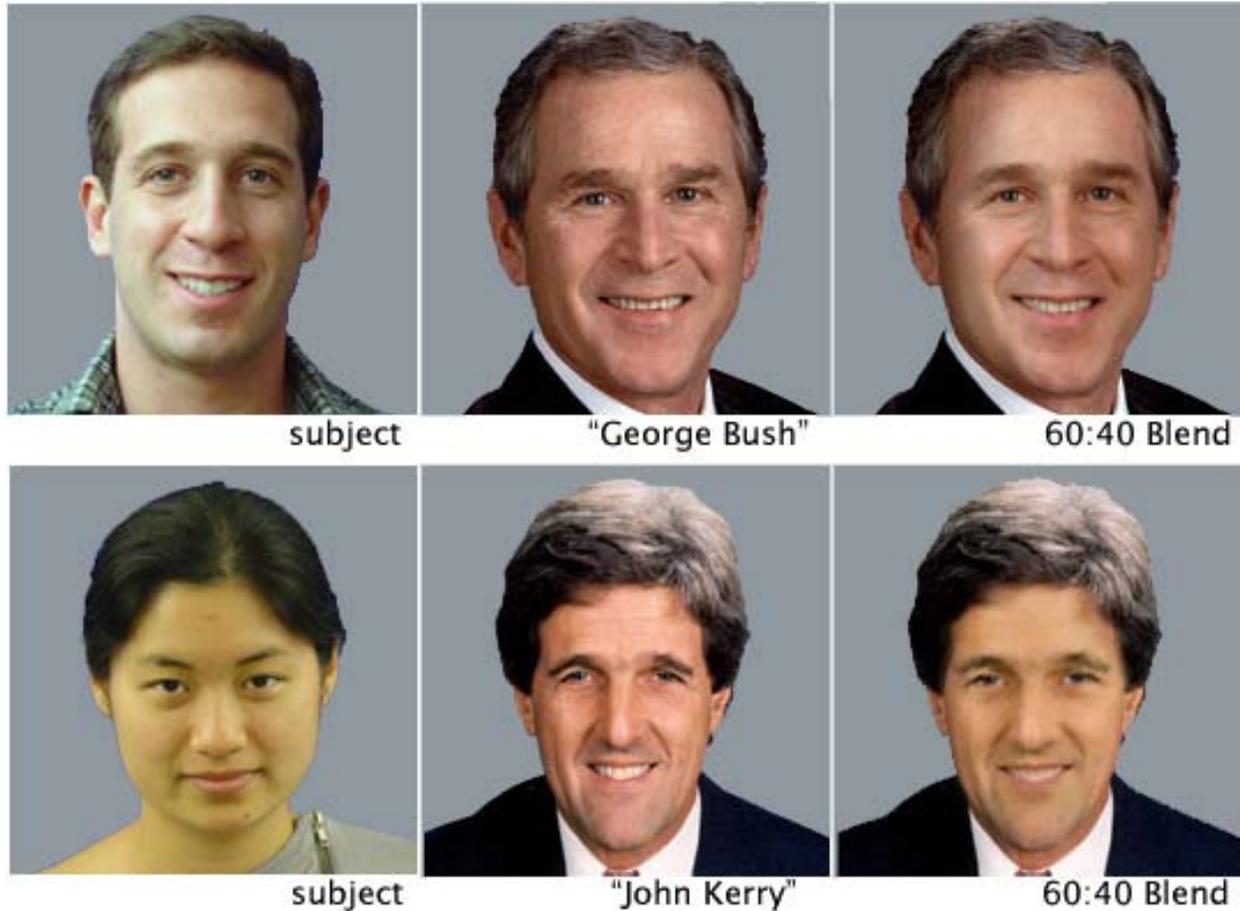
- David Mechanic, James P. Smith, Beth J. Soldo, Robert Wallace, Robert J. Willis, David Wise and Zeng Yi. 2001. *Preparing for an Aging World: The Case for Cross-national Research*. Washington, DC: National Academy Press.
- Kahn, Kim F., and Patrick J. Kenney. 1999. "Do Negative Campaigns Mobilize or Suppress Turnout? Clarifying the Relationship between Negativity and Participation." *American Political Science Review* 93: 877-90.
- Kinder, Donald R., and Thomas R. Palfrey. 1993. *Experimental Foundations of Political Science*. Ann Arbor: University of Michigan Press.
- Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Assessment." *American Political Science Review* 93: 851-75.
- Malhotra, Neil, and Jon A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Non-probability Samples." *Political Analysis* 15: 286-323.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Nelson, Charles A. 2001. "The Development of Neural Bases of Face Recognition." *Infant and Child Development* 10: 3-18.
- Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17: 776-83.
- Pierce, John C., and Nicholas P. Lovrich. 1982. "Survey Measurement of Political Participation: Selective Effects of Recall in Petition Signing." *Social Science Quarterly* 63: 164-71.
- Price, Vincent, and John R. Zaller. 1993. "Who Gets the News? Alternative Measures of News Reception and Their Implications for Research." *Public Opinion Quarterly* 57: 133-64.
- Prior, Markus. 2003. "Any Good News in Soft News? The Impact of Soft News Preference on Political Knowledge." *Political Communication* 20: 149-72.
- Prior, Markus. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. New York: Cambridge University Press.
- Riker, William H. 1967. "Bargaining in a Three-Person Game." *American Political Science Review* 61: 642-656.
- Rivers, Douglas R. 2006. "Sample Matching: Representative Sampling from Internet Panels." Retrieved from http://www.polimetrix.com/documents/Polimetrix_Whitepaper_Sample_Matching.pdf.

- Rivers, Douglas R., and Delia Bailey. 2009. "Inferences from Matched-Samples in the 2008 U.S. National Elections." Proceedings of the Survey Research Methods Section of the American Statistical Association: 627-39.
- Roskos-Ewoldsen, David, Beverly Roskos-Ewoldsen, and Francesca R. Carpentier. 2005. "Media priming: A Synthesis." In *Media Effects: Advances in Theory and Research*, eds. Jennings Bryant and Dolph Zillmann. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Scioli, Frank. 2009. Personal communication.
- Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on the Social Psychology View of Human Nature." *Journal of Personality and Social Psychology* 51: 515-30.
- Vavreck, Lynn. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 325-43.
- Wattenberg, Martin P., and Craig L. Brians. 1999. "Negative Campaign Advertising: Demobilizer or Mobilizer?" *American Political Science Review* 93: 891-900.
- Zajonc, Robert B. 2001. "Mere Exposure: A Gateway to the Subliminal." *Current Directions in Psychological Science* 10: 224-28.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Figure 6-1. Race of Suspect Manipulation



Figure 6-2. The Facial Similarity Manipulation



ⁱ An important impetus to the development of political psychology was provided by the Psychology and Politics Program at Yale University. Developed by Robert Lane, the program provided formal training in psychology to political science graduate students and also hosted postdoctoral fellows interested in pursuing interdisciplinary research. Later directors of this training program included John McConahay and Donald Kinder.

ⁱⁱ Scholars who played important editorial roles at *ESP* included Marilyn Dantico (who took over as coeditor of the journal when Scioli moved to NSF), Richard Brody, Gerald Wright, Heinz Eulau, James Stimson, Steven Brown and Norman Luttbeg.

ⁱⁱⁱ The social psychology laboratories included rooms with transparent mirrors and advanced video and sound editing systems.

^{iv} The extent of the Stony Brook political science department's commitment to interdisciplinary research was apparent in the department's hiring of several newly-minted social psychologists. The psychologists recruited out of graduate school -- none of whom fully understood, at least during their job interview, why a political science department would see fit to hire them -- included John Herrstein, George Quattrone, Kathleen McGraw and Victor Otatti. Of course, the psychologists were subjected to intense questioning by the political science faculty over the relevance and generalizability of their research. In one particularly memorable encounter, following a job talk on the beneficial impact of physical arousal on information processing and judgment, an expert on voting behavior asked the candidate whether he would suggest requiring voters to exercise prior to voting.

^v In a meta-analysis of political advertising research, Lau et al. concluded that experimental studies were not more likely to elicit evidence of significant effects (Lau et al. 1999). The meta-analysis, however, combines experiments that utilize a variety of designs most of which fail to isolate the negativity of advertising.

^{vi} Of course, the use of deception in experimental research necessitates full debriefing of participants at the conclusion of the study. Typically, participants are provided with a relatively detailed account of the experiment and are given the opportunity to receive any papers based on the study data. In recent years, experimental procedures have become highly regulated by university review boards in order to maximize the principle of informed consent and to preclude any lingering effects of deception. Most informed consent forms, for instance, alert participants to the use of deception in experimental research.

^{vii} Of course, this approach assumes a one-sided distribution of policy preferences and that the tone manipulation would be reversed for experimental participants who actually favored off shore drilling.

^{viii} We settled on the 60:40 ratio after a pretest study indicated that this level of blending was insufficient for participants to detect traces of themselves in the morph, but sufficient to move evaluations of the target candidate.

^{ix} Facial similarity is necessarily confounded with familiarity – people are familiar with their own faces. There is considerable evidence (see Zajonc 2001) that people prefer familiar to unfamiliar stimuli. An alternative interpretation of these results, accordingly, is that participants were more inclined to support the more familiar-looking candidate.

^x In the early days of the campaign advertising research, the experimental lab included a remote control device placed above the television set. This proved to be excessively realistic as some subjects chose to fast forward the videotape during the ad breaks. The device was removed.

^{xi} For further discussion of the subject recruitment issue and implications for external validity, see Druckman and Kam's chapter in this volume.

^{xii} The fact that the Polimetrix online samples can be matched according to a set of demographic characteristics does *not* imply that the samples are unbiased. All sampling modes are characterized by different forms of bias and opt-in web panels are no exception. In the US, systematic comparisons of the Polimetrix online samples with random digit dial (telephone) samples and face-to face interviews indicate trivial differences between the telephone and online modes, but substantial divergences from the face-to-face mode (see Hill, Lo, Vavreck, and Zaller 2007; Malhotra and Krosnick 2007). In general, online samples appear biased in the direction of politically engaged and attentive voters.

^{xiii} Indeed, comparativists are fond of pointing out the inherently noncomparative and hence pre-scientific nature of research in American politics.

^{xiv} A useful compilation of online experimental labs can be retrieved at <http://psych.hanover.edu/research/exponnet.html>

7. Experiments and Game Theory's Value to Political Science

John H. Aldrich and Arthur Lupia

In recent decades, formal models have become common means of drawing important inferences in political science. Best-practice formal models feature explicitly stated premises, explicitly stated conclusions, and proofs that are used to support claims about focal relationships between these premises and conclusions. When best practices are followed, transparency, replicability, and logical coherence are the hallmarks of the formal theoretic enterprise.

Formal models have affected a broad range of scholarly debates in political science – from individual-level inquiries about why people vote as they do, to large-scale studies of civil wars and international negotiations. The method's contributions come when answers to normative and substantive questions require precise understandings about the conditions under which a given political outcome is, or is not, consistent with a set of clearly stated assumptions about relevant perceptions, motives, feelings, and contexts. Indeed, many formal modelers use mathematics to sort intricate and detailed statements about political cause and effect by the extent to which they can be reconciled logically with basic premises about the people and places involved.

While formal models in political science have been influential, they have also been controversial. Although no one contends that explicitly stated assumptions or attention to logical consistency are anything other than good components of scientific practice, controversy often comes from the content of formal models themselves. Many formal models contain descriptions of political perceptions, opinions, and behaviors that are unrealistic. Some scholars, therefore, conclude that formal models, generally considered, are of little value to political science. Yet, if

we can offer a set of premises that constitutes a suitable analogy for what key political actors want, know, and believe, then we can use formal models to clarify conditions under which these actors will, and will not, take particular actions.

In what follows, we address two questions that are particularly relevant to debates about formal models' substantive relevance. We will present each question in turn and, in so doing, explain how experiments affect the value of formal modeling to political science.

One question is, "Will people who are in the situations you describe in your model act as you predict?" This question is about the internal validity of a model. Experiments permit the creation of a specialized setting in which a model's premises can be emulated, with the test being whether the experiment's subjects behave as the model predicts (akin to physicists studying the action of falling objects in a laboratory-created vacuum, thus as free from air resistance as possible). Lupia and McCubbins (1998), for example, devote an entire chapter of their book (Chapter 6) to pinpointing the correspondence between the experimental settings and the models that the experiments were designed to evaluate. Furthermore, if a modeler wants to claim that a particular factor is a unique cause of an important behavior, he or she can design various treatments that vary the presence of the presumably unique causal factor. If the focal behavior is observed only when the presumed factor is present, she will have greater evidence for her claim. Generally speaking, this way of thinking about the role of experiments in formal theoretic political science is akin to Roth's (1995) description of experiments as a means of "speaking to theorists."

A second question is, "Are your theoretical predictions representative of how people will act in more realistic circumstances?" This question speaks to the ecological validity of models and is akin to Roth's (1995) description of experiments as "whispering in the ears of princes."

Experimental designs that address these questions should incorporate elements that audiences would see as essential to proffering a substantive explanation, but that are not necessarily included in the model. Cross-national experiments are an example of designs that address such concerns. As Wilson and Eckel's chapter in this volume explains, for models whose predictions are not culture-specific, it is important to evaluate whether experimental subjects in different regions or countries will really play given games in identical ways. When experiments reveal cross-cultural differences, then theorists who desire broadly-applicable conclusions can use these findings to better integrate cultural factors into their explanations.

Indeed, a key similarity between what formal modelers and experimentalists do is that neither simply observes the environments they wish to explain. Instead, both seek to create settings that emulate the environments. When used in tandem, formal models can help experimentalists determine which settings are most critical to a particular causal hypothesis and experimenters can inform formal modelers by evaluating their theoretical predictions' performance in relevant environs.

Given the emphasis on logical precision in formal modeling, it is also worth noting that many model-related experiments follow practices in experimental economics, which includes paying subjects for their participation as a means of aligning their incentives with those of analogous actors in the formal models. As Palfrey (2007a) explains

[R]esearchers who were trained primarily as [formal] theorists – but interested in learning whether the theories were reliable – turned to laboratory experiments to test their theories, because they felt that adequate field data were unavailable. These experiments had three key features. First, they required the construction of isolated (laboratory) environments that operated under specific, tightly controlled, well-defined institutional rules. Second, incentives were created for the participants in these environments in a way that matched incentives that existed for the imaginary agents in theoretical models. Third, the theoretical models to be studied had precise context-free implications about behavior in any such environment so defined, and these predictions were quantifiable and therefore directly testable in the laboratory (915).

For formal modelers, these attributes of experimentation are particularly important, as a comparative advantage of formal theoretic approaches is precision in causal language.

In the rest of this chapter, we highlight a number of ways in which experiments have affected the value of formal modeling in political science. Because the range of such activities is so broad, we focus our attention on a type of formal modeling called game theory. Game theory is a way of representing interpersonal interactions. Premises pertain to specific attributes of individual actors and the contexts in which people interact. Conclusions describe the aggregate consequences of what these actors do, or the properties of the individuals themselves, that result from interactions amongst the actors.

The next two substantive sections of this chapter pertain to the two main types of game theory, respectively. Section 1 focuses on experiments in the domain of cooperative game theory. Political scientists have used cooperative game theory to address a number of key theoretical and normative debates about preference aggregation and properties of common decision rules. As the first game theoretic experiments in political science tested results from cooperative game theory, many of the standard protocols of experimental game theory were developed in this context. Section 2 focuses on experiments in the domain of noncooperative game theory. Most game-theoretic treatments of political science topics today use some form of noncooperative game theory. Noncooperative game theory can clarify how actors pursue their goals when they and those around them have the ability to perceive and adapt to important attributes of their environment. Influential models of this kind have clarified how institutions affect individual choices and collective outcomes, and how strategic uses of communication affect a range of important political outcomes.

In the conclusion, we speak briefly about how experiments may affect future relationships between formal modeling and political science. We argue that, while the psychological realism of many current models can be questioned, research agendas that integrate experimental and formal modeling pursuits provide a portal for more effective interdisciplinary work and can improve the applicability and relevance of formal models to a wide range of important substantive questions in political science.

1. Cooperative Game Theory and Experiments

Game theory is often divided into cooperative and noncooperative game theory, and it is true that most results can be fit into one or the other division. However, the formal definitions are both at the extreme points on a continuum, and thus virtually the entire continuum fits in neither category very precisely. So saying, the basic distinction is that in cooperative game theory, coalitions may be assumed to form, whereas in noncooperative game theory, any coalitions must be deduced from the model itself rather than assumed a priori. In the early days of game theory, which we can mark from publication of Von Neumann and Morgenstern (1944), to approximately the late 1970s when the Nash-Harsanyi-Selten revolution in noncooperative game theory carried the day (and won them a Nobel prize in 1994), this now-critical distinction was much less important. Theoretical results, for example, were not often identified as one or the other and theorists (including Nash himself, 1997) moved back and forth across this division easily.

This early blurring of the distinction between the two is relevant here because experimental game theory began very early in the history of game theory and experiments would at times include and intermingle results from cooperative and noncooperative game designs. This is perhaps most evident in the truly vast experimental literature on prisoner's dilemma games.

The PD, as it is sometimes known, played an important role in game theory from the beginning, because it was immediately evident that the very strong prediction of rational players both (or all) defecting with or without communication being possible was simply empirically false. Whether presented as a teaching device in an introductory undergraduate course or tested in the most sophisticated experimental setting, people simply do not follow the predictions of game theory (particularly noncooperative game theory) in this regard. Therefore, game theorists naturally turned to experimentation to study play by actual players to seek theoretical insight (see, for example, Rapoport and Chammah 1965 for a review of many early PD game experiments). This case is thus very similar to the experimental work on the “centipede game,” which will be discussed in Section 2, in that both endeavors led scholars to question key assumptions and to develop more effective modeling approaches.

Coalition formation

As noted, game theorists conducted experiments from the earliest days of game theory (see for example, Kalish et al. 1954 on which John Nash was a coauthor). A common application was to the question of coalitions. Substantively, Riker (1962) had argued for the centrality of coalitions for understanding politics. Theoretically, cooperative game theory was unusually fecund, with a diverse set of n-person games and many different solution concepts whose purposes were to characterize the set of coalitions that might form.

This set of solution concepts has three notable attributes. First, one concept, called the core, had the normatively attractive property that outcomes within it did not stray too far from the preferences of any single player. In other words, the core was a set of outcomes that were preferred by majorities of voters and not easily overturned by attempts to manipulate voting agendas. Unfortunately, such core outcomes have the annoying property of not existing for a

very large class of political contexts. As Miller's chapter in this volume details, the general nonexistence of the core raises thorny normative questions about the meaning and legitimacy of majority decision making. If there is no structure between the preferences of individuals and the choices that majority coalitions make, then it becomes difficult to argue that preference-outcome links (e.g., the will of the majority) legitimate majority decision making.

To address these and other normative concerns, scholars sought other solution concepts that not only described coalitional choices in cases where the core did not exist, but also retained some of the core's attractive properties. Second, theorists often developed these alternate concepts by adding assumptions that did not flow from the basic concepts of game theory as it was known at the time. It is fair to say, from a game theoretic perspective, that many of the new assumptions needed to characterize coalition behavior were simply arbitrary. Third, many of these new ways of characterizing coalition behaviors offered vague (as in many outcomes predicted) or unhelpful (as in, nonexistent) conclusions. Moreover, the diverse set of solution concepts often had overlapping predictions, making it often difficult to distinguish between them in observational data.

So while the multiplicity of solution concepts was part of an effort to clarify important attributes of coalition behavior, an aggregate consequence of such efforts was more confusion. Given his interest in the topic, and his substantive claim of the centrality of coalitions to politics, it is not surprising that the first published attempt to simulate a game theoretic context in a laboratory setting was by William Riker (1967; Riker and Zavoina 1970). These simulations were important for several reasons. First, they established what became a standard protocol for conducting game theoretic experiments. Preferences were typically induced by money. Communications between players were either controlled as carefully as possible by the

experimenter or else were carefully and, as far as technology permitted, fully recorded. Assumptions were built into the research design to mimic the assumptions of the solution concept being examined. Behaviors were closely observed, with the observational emphasis being on whether the coalitions chosen were consistent with the concept or concepts being tested. Lessons learned were also reported in the text. For example, Riker discovered that students learned that they could, in effect, deceive the experimenter by agreeing outside (and in advance) of the simulation to exchange their university's food cards as a way of reaching binding agreements. This outcome taught the lesson that not only is the repeated use of the same subjects potentially problematic, but also that subjects – and therefore, at least potentially, people in real situations – will devise strategies to make binding commitments even when the situation precludes them formally.

The proliferation of solution concepts, such as those we have described, motivated many interesting and important extensions to the described setting. With so many overlapping predictions, observational data were often of little help in selecting among competing accounts. For example, where Riker developed a theory of minimal winning coalitions (1962), virtually every other cooperative solution concept that focused on coalition formation was also consistent with the observation of minimal winning coalitions. This overlap made it difficult to distinguish the power of various solution concepts. Riker's research designs varied the construction of subject preferences – preferences that one could induce with money – to distinguish competing causal mechanisms. These distinctions, in turn, clarified which solution concepts were and were not viable in important cases. By this careful construction, Riker could distinguish through lab design what was very difficult to distinguish in real-world settings.

McKelvey and Ordeshook's research also captured theorists' attention. They developed one of the last cooperative game theoretic solution concepts, called the “competitive solution” (McKelvey and Ordeshook 1978, 1979, 1980, 1983; McKelvey, Ordeshook, and Winer 1978; Ordeshook 2007). They were very interested in evaluating their concept experimentally vis-à-vis other solution concepts. Their 1979 paper reports on a series of experiments that was able to establish predictive differences among nine solution concepts (some of which had noncooperative game theoretic elements). In particular, they used the data to proffer statistical estimates of likelihoods that revealed support for two of the concepts (theirs and the minimax set of Ferejohn, Fiorina, and Packel 1980, while effectively rejecting the other seven.

McKelvey and Ordeshook ended their 1979 paper with a conclusion that would turn out to be prescient. They write, “Finally, we cannot reject the hypothesis that [the competitive solution] succeeds here for the same reason that [other solutions] receive support in earlier experiments entailing transferable utility – namely that [the competitive solution’s] predictions correspond fortuitously to some more general solution notion” (165). To see why this statement is prescient, it is important to note that the experimental protocol was developed to allow researchers to evaluate various solution concepts on the basis of coalition-level outcomes. The solution concepts they evaluated, however, derived their conclusions about coalition-level outcomes from specific assumptions regarding how players would actually bargain. McKelvey and Ordeshook soon began to use their experimental protocol to evaluate the status of these assumptions (i.e., the experiments allowed them to observe how subjects actually bargained with one another). In their 1983 paper, they reported on experiments that once again produced the results predicted by their competitive solution. However, they also observed players bargaining in ways that appeared to violate the assumptions of the competitive solution. Such observations

ultimately led them to abandon the competitive solution research agenda for many years (Ordeshook 2007) and to devote their attention to other explanations of collective behavior (see Section 2 of this chapter, and Morton and Williams's chapter in this volume for descriptions of that work).

This idea of using the lab setting as a way of sorting among solution concepts reached an apogee in Fiorina and Plott (1978), in which they develop sixteen different sets of theoretical predictions. Some are from cooperative game theory, some from noncooperative game theory. Some focus on voting, some on agenda control. Some are not even based in rational actor theories. The beauty of their use of the lab setting (and their own creativity) is that they are able to design experiments that allow for competing tests between many of the pairs of theories, and sometimes tests that uniquely discriminate one account from virtually all others. In addition to differentiating outcomes in this way, they examined the effect of varying treatments – in particular, whether the payoffs were relatively high or low to the players and whether there was open exchange of communication or no communication at all among the players.

They found that when there is a core (a majority-preferred outcome that is not easily undone by agenda manipulation), it is almost always chosen. High payoffs yielded outcomes even closer to those predicted points than lower payoffs, and, to a much lesser extent, communication facilitated that outcome. Comparing these outcomes to those where no core exists allows the authors to conclude more sharply that it is indeed the existence of a core that drives the results. On the other hand, the absence of a core also yielded an apparent structure to the set of coalition-level outcomes rather than an apparently unpredictable set of results, as might have been expected by some readings of McKelvey's (1976) and Schofield's (1978) theoretical results about the unlimited range of possible outcomes of majority decision making in the

absence of a core. Rather (as, indeed, McKelvey argued in the original 1976 article), there seems to be structure to what coalitions do even in the absence of a core. Hence, the correspondence between individual intentions and coalition-level outcomes in the absence of a core is not totally chaotic or unstable, and this correspondence provides a basis for making normatively appealing claims about the legitimacy of majority rule (see Frohlich and Oppenheimer 1992 for a broader development of links among models, experiments, and important normative considerations). However, these structures, while observed by Fiorina and Plott, were not the result of any theory then established. Hence, one of the lasting contributions of the work by Fiorina and Plott, as is true of the other work described in this section, is that it set the stage for other experiments on political decision making, such as those discussed in the next section and by Morton and Williams's chapter in this volume.

2. Noncooperative Game Theory and Experiments

Noncooperative game theory is a method of formal modeling that allows researchers to draw logically transparent conclusions about how individuals adapt and react to the anticipated strategic moves of others. It uses the Nash Equilibrium concept, or well-known refinements of the concept, as a criterion for identifying behavioral predictions. In recent decades, noncooperative games using the extensive form have been formal modelers' primary instrument in attempting to make contributions to political science. The extensive form outlines, in order, the decisions to be reached, actor by actor, from the opening move to the final outcome. It thus offers a rich perspective for analyzing strategic decision making as well as the roles of beliefs and communication in decision making. These games have informed our discipline's attempts to clarify the relationship among political institutions, individual choices, and collective outcomes. They have also been the means by which scholars have examined positive and normative

implications of the strategic use of information (see Austen-Smith and Lupia 2007 for a recent review). As the field evolves, these games increasingly serve as a portal through which implications of substantive premises from fields such as economics and psychology can become better understood in political contexts.

Key moments in the evolution of such understandings have been experimental evaluations of these games. In this section, we review three examples of where the combination of noncooperative game theoretic models and experiments has produced new insights about important social scientific matters. The examples are: voter competence, jury decision making, and the centipede game.

Voter Competence

A conventional wisdom about mass politics is that candidates and their handlers seek to manipulate a gullible public and that the public makes inferior decisions as a result (see, e.g., Converse 1964). In recent decades, scholars have used formal models and experiments in tandem to examine when seemingly uninformed voters do – and do not – make inferior decisions. In this section, we will review two examples of such work. In each case, scholars use formal models to understand whether claims about the manipulability of voters are, and are not, consistent with clearly stated assumptions about voters' and candidates' incentives and knowledge. Experiments then clarify the extent to which subjects will act in accordance with focal model predictions when they are placed in decision-making environments that are similar to the ones described in the models.

McKelvey and Ordeshook (1990) focus on a spatial voting model in which two candidates compete for votes by taking policy positions on a unidimensional policy space. Voters have spatial preferences, which is to say that they have an ideal point that represents the policy

outcome they most prefer. A voter in these models obtains higher utility when the candidate whose policy preference is closest to their ideal point is elected.

If the game were one of complete information, the outcome would be that both candidates adopt the median voter's ideal point as their policy preference and the median voter's ideal point becomes the policy outcome (Black 1948). The focal research question for McKelvey and Ordeshook is how such outcomes change when voters know less. To address these questions, McKelvey and Ordeshook develop a model with informed and uninformed voters. Informed voters know the policy positions of two candidates. Uninformed voters do not, but they can observe poll results or interest group endorsements. McKelvey and Ordeshook examine when uninformed voters can use the polls and endorsements to cast the same votes they would have cast if completely informed.

In the model's equilibrium, voters make inferences about candidate locations by using poll results to learn how informed voters are voting. Uninformed voters come to correctly infer the candidates' positions from insights such as "if that many voters are voting for the [rightist candidate], he can't be too liberal." McKelvey and Ordeshook prove that the greater the percentage of informed voters represented in such polls, the quicker uninformed voters come to the correct conclusion about which candidate is closest to their interests.

McKelvey and Ordeshook evaluate key aspects of their theoretical work experimentally. As Palfrey (2007a, caveats in brackets inserted by us) reports,

Perhaps the most striking experiment...used a single policy dimension, but candidates had no information about voters and only a few of the voters in the experiments knew where the candidates located. The key information transmission devices explored were polls and interest group endorsements. In a theoretical model of information aggregation, adapted from the rational expectations theory of markets, they proved that this information alone [along with the assumption that voters know approximately where they stand relative to the rest of the electorate on a left-right scale] is sufficient to reveal

enough to voters that even uninformed voters behave optimally – i.e., as if they were fully informed (923).

Of course, uninformed voters in the McKelvey-Ordeshook model do not cast informed votes in all circumstances. The caveats inserted into the Palfrey quote highlight key assumptions that contribute to the stated result. However, it is important to remember that the conventional wisdom at the time was that uninformed voters could seldom, if ever, cast competent votes -- where competence refers to whether or not a voter casts the same vote that she would have cast if she possessed full information about all matters in the model that are pertinent to her choice (e.g., candidate policy positions). The breadth of conditions under which McKelvey and Ordeshook proved that a) uninformed voters vote competently and b) election outcomes are identical to what they would have been if all voters were informed prompted a reconsideration of the conditions under which limited information made voters incompetent.

Lupia and McCubbins (1998) pursue these conditions further. They examine multiple ways in which voters can be uninformed and incorporate focal insights from the psychological study of persuasion. By using formal models and experiments, they could clarify how conditional relationships among psychological, institutional, and other factors affect competence and persuasion in ways that the dominant approach to studying voter competence – conventional survey based analyses – had not.

The starting point for Lupia and McCubbins is that citizens must make decisions about things that they cannot experience directly. For voters, the task is to choose candidates whose future actions in office cannot be experienced in advance of the election. Relying on others for information in such circumstances can be an efficient way to acquire knowledge. However, many people who provide political information (e.g., campaign organizations) do so out of self-interest, and some may have an incentive to mislead. For voters who rely on others for

information, competence depends on whom they choose to believe. If they believe people who provide accurate information and ignore people who do otherwise, they are more likely to be competent.

A key move in the development of the Lupia-McCubbins model is to follow the arguments of empirical scholars of voting behavior and public opinion who linked this question to the social psychological study of persuasion. O'Keefe (1990) defines persuasion as “a successful intentional effort at influencing another’s mental state through communication in a circumstance in which the persuadee has some measure of freedom” (17). Seen in this way, the outcomes of many political interactions hinge on who can persuade whom. Social psychologists have generated important data on the successes and failures of persuasive attempts (see, e.g., McGuire 1985; Petty and Cacioppo 1986).

While psychological studies distinguish factors that can be antecedents of persuasion from factors that cannot, they are typically formulated in a way that limits their applicability to questions of voting behavior. The typical social psychological study of persuasion is a laboratory experiment that examines how a single variation in a single factor corresponds to a single attribute of persuasiveness. Such studies are designed to answer questions about the conditions under which some attributes will be more important than others in affecting the persuasive power of a particular presentation. In a formal model, it is possible to conduct an analysis of the conditions under which a range of factors has differential and conditional effects on whether persuasion occurs. Lupia and McCubbins do just that, examining the logical consequences of mixing a range of assumptions about beliefs and incentives to generate precise conclusions about the conditions under which a) one person can persuade another and b) persuasive attempts make voters competent -- that is, helps them choose as they would if fully informed.

Their models and experiments also show that any attribute causes persuasion only if it informs a receiver's perceptions of a speaker's knowledge or interests. Otherwise, the attribute cannot (and experimentally does not) affect persuasion, even if it actually affects the speaker's choice of words. Experiments on this topic clarified how environmental, contextual, and institutional variables (such as those that make certain kinds of statements costly for a speaker to utter) make learning from others easier in some cases and difficult in others.

These and other subsequent experiments demonstrate that the knowledge threshold for voting competently is lower than the normative and survey-based literatures at the time had conjectured (see Boudreau and Lupia's chapter in this volume for more examples of such experiments). Instead of being required to have detailed information about the utility consequences of all electoral alternatives, it can be sufficient for the voter to know enough to make good choices about whom to believe. So when information about endorsers is easier to acquire than information about policies, voters who appear to be uninformed can cast the same votes they would have cast if they knew more. In sum, there appears to be logic to how uninformed voters use information. Formal models have provided a basis for discovering it, and experimentation offers one important way for testing it.

Jury Decision Making

Experiments have also clarified implications of a visible game theoretic claim about jury decision making. Many courts require a unanimous vote of a jury in order to convict a defendant. A common rationale for this requirement is that unanimity minimizes the probability of convicting the innocent.

Feddersen and Pesendorfer (1998) identify an equilibrium in which unanimity produces more false convictions than was previously believed. The logic underlying their result is as

follows. Suppose that all jurors are motivated to reach the correct verdict. Suppose further it is common knowledge that every juror receives a signal about a defendant's status (i.e., courtroom testimony and/or jury room deliberation) that is true with a known probability.

In this case, a juror is either not pivotal (i.e., her vote cannot affect the outcome) or is pivotal (i.e., her vote does affect the outcome). Under unanimity, if at least one other juror is voting to acquit, then a juror is not pivotal, the defendant will be found not guilty no matter how this juror decides. Likewise, a juror is pivotal under unanimity rule only if every other juror is voting to convict. Hence, a juror can infer that either her vote makes no difference to the outcome or that all other jurors are voting to convict. Feddersen and Pesendorfer examine how such reasoning affects the jurors' assessment of the defendant's guilt. They identify conditions in which the weight of each juror's conjecture about what other jurors are doing leads every juror to vote to convict -- even if every single juror, acting solely on the basis of the signal they received, would have voted innocent. False convictions come from such calculations and are further fueled by jury size (as n increases, so does the informational power of the conjecture that "if I am pivotal, then it must be the case that every other juror is voting to convict.") Feddersen and Pesendorfer use these results to call into question claims about unanimity's convictions of the innocent.

A number of scholars raised questions about whether making more realistic assumptions about jurors could yield different results. Some scholars pursued the question experimentally. Guernaschelli, McKelvey, and Palfrey (2000) examine student juries of different sizes ($n=3$ and $n=6$) that were otherwise in the type of decision environment described by Feddersen and Pesendorfer. Guernaschelli, McKelvey, and Palfrey report that where: "Feddersen and Pesendorfer (1998) imply that large unanimous juries will convict innocent defendants with

fairly high probability... this did not happen in our experiment” (416). In fact, and contrary to Feddersen and Pesendorfer’s claims, this occurrence happened less frequently as jury size increased.

Experimental results such as these imply that the frequency at which unanimity rule convicts the innocent requires additional knowledge of how jurors think. These experiments helped to motivate subsequent modeling that further clarified when strategic voting of the kind identified by Feddersen and Pesendorfer cause unanimity requirements to produce false convictions. With a model whose assumptions are built from empirical studies of juries by Pennington and Hastie (1990, 1993) and psychological experiments on need for cognition by Cacioppo and Petty (1982), Lupia, Levine, and Zharinova (2010) prove that it is not strategic voting per se that generates Feddersen and Pesendorfer’s high rate of false convictions. Instead, driving the increase in false convictions is the assumption that all jurors conjecture that all other jurors are thinking in the same manner as they are. Lupia, Levine, and Zharinova (2010) show that strategic voting under different, and more empirically common beliefs, can cause far fewer false convictions. Collectively, experiments and subsequent models show that using more realistic assumptions about jurors generate equilibria with many fewer false convictions.

More generally, we believe that the pairing of formal models and experiments can be valuable in improving political scientists’ efforts to pursue psychological explanations of behavior. Models can help scholars determine whether claims being made about citizen psychology must be true given a set of clearly stated assumptions, or whether the claim is possibly true given those foundations. These types of questions are now being asked with increasing directness (e.g., Lupia and Menning 2009) and are serving as the foundations for several exciting new research agendas, such as Dickson’s chapter in this volume describes.

Contributions to Other Fields

Political scientists have also used combinations of game theory and experiments to make contributions whose relevance extends well beyond political science. Eckel and Wilson's discussion of trust (chapter in this volume) and Coleman and Ostrom's discussion of collective action (chapter in this volume) provide prominent examples. Other such examples are experiments by McKelvey and Palfrey (1992, 1995, 1998). Their experimental and theoretical efforts provide a focal moment in the emergence of behavioral economics.

Behavioral economics is a movement that seeks to derive economically relevant conclusions from premises with increased psychological realism. The evolution and gradual acceptance in economics of a behavioral approach was motivated by an important set of experiments. These experiments revealed systematic divergence between the predictions of several well-known game theoretic models and the behavior of laboratory subjects in arguably similar decision contexts.

One such model is called the centipede game. In a centipede game, two players decide how to divide an object of value (say, ten dollars). One player can take a very unequal share of the object for herself (say, "I get seven dollars and you get three dollars") or the player can pass on that opportunity. If she takes the larger share, the game ends and players are paid accordingly. If she passes, the object doubles in value and the other player can take the larger share (say, "I get fourteen dollars and you get six dollars"). An important part of the game is that payoffs are arranged so that a player gets slightly more from taking in the current round (seven dollars) than they will if the other player takes in the next round (six dollars). The game continues for as long players pass. In every subsequent round, the object continues to double in value after each pass and players alternate in their ability to take.

It is easy to imagine that both players could earn very high payoffs by passing for a while to let the object grow in value. The game, however, has a unique Nash equilibrium (Rosenthal 1981) and it does not involve such behavior. Instead, it predicts that players will take at the first opportunity. A number of scholars raised questions about the applicability of this prediction. Two political scientists, McKelvey and Palfrey, ran experiments to address these questions. As Palfrey (2007b) describes of their experimental efforts,

[W]e designed and conducted an experiment, not to test any particular theory (as both of us had been accustomed to doing), but simply to find out what would happen. However, after looking at the data, there was a problem. Everything happened! Some players passed all the time, some grabbed the big pile at their first opportunity, and others seemed to be unpredictable, almost random. But there were clear patterns in the average behavior, the main pattern being that the probability of taking increased as the piles grew (426).

Their efforts to explain such behavior evolved into the development of a new equilibrium concept, Quantal Response Equilibrium (QRE). QRE is a variant of the Nash Equilibrium concept that allows modelers to account for a much wider set of assumptions about what players believe about one another than did traditional concepts. As applied to the centipede game, the concept allowed players to adjust their strategies to varying beliefs that the other players would (or would not) play the strategies named in the game's unique Nash equilibrium. This concept allowed McKelvey and Palfrey to explain patterns of play in the centipede game far more effectively than other approaches.

Both McKelvey and Palfrey's experimental documentation of the problems with the applicability of the centipede game's Nash Equilibrium and their use of these results to develop an alternate equilibrium concept now serve as models for why a more behavioral approach to economic topics is needed and how more realistic psychological content can begin to be incorporated.

3. Conclusion

An important attribute of formal models is that they allow scholars to analyze, with precision and transparency, complex conditional relationships among multiple factors. As such, scholars can use formal models to evaluate the conditions under which various kinds of causal relationships are logically consistent with clearly stated assumptions about actors and institutions. The distinguishing characteristic of formal models is their ability to facilitate constructive and precise conversations about what-is-related-to-what in domains of political choice.

In some cases, however, scholars raise reasonable questions about whether the logic of a particular formal model is relevant to a particular set of real-world circumstances. At such moments, empirical demonstrations can be valuable. They can demonstrate that the model does in fact explain relevant behaviors well, or they can show that the model requires serious revision.

In many cases, however, nature does not provide the kinds of data scholars would need to answer such questions. Moreover, if the claims in question pertain to how certain actors would react under a wide range of currently hypothetical circumstances, or if the controversy pertains to whether a particular claim is true given a different set of underlying counterfactuals, then there may be no observational approach that will provide sufficient data. In such cases, experiments can help us evaluate the relevance and applicability of focal model attributes to important political phenomena.

This chapter has described a few instances in which experiments played an important role in the development and evaluation of game theoretic political science. Several chapters in this volume review other interesting examples. Morton and Williams, for example, detail how a number of clever experimental agendas clarify how various institutional rules affect electoral behavior and outcomes. Coleman and Ostrom review how experiments of many different kinds

have helped scholars from multiple disciplines better understand the prerequisites for effective collective action.

Diermeier's chapter in this volume highlights experimental evaluations of the Baron-Ferejohn model of coalition bargaining. He describes how his early experiments consistently demonstrate that the person with the ability to propose coalition agreements to other actors consistently takes less power than the model predicts. Drawing from psychology, he argues that other factors relevant to sustained human interaction could induce a bargainer to offer potential partners more than the absolute minimum amounts that they would accept. His later experiments incorporate these factors and show promise as a foundation of more effective explanations of coalition behavior.

Wilson and Eckel's chapter in this volume shows how experiments have clarified many questions about the relevance and applicability of formal models of trust. They begin by describing the Nash equilibrium of the "investment game." It is a game where players can benefit by trusting one another to contribute some of their resources to a common pool. But the unique Nash equilibrium of this particular game is that players will not trust one another enough to realize these gains. They then review a series of experiments that examine many psychological and biological factors relevant to trust in a range of cultural and institutional settings around the world. Through these efforts, theories and experimental designs build off of one another and what results is clarification about when we can expect trust in a set of critical social relationships.

Collectively, these chapters reveal both the challenges inherent in using formal models and experiments to provide substantive insight to political science, and the ways in which experiments help formal modelers and scholars with more substantive interests communicate

more effectively. Given the increasing number of political scientists who are interested in experiments, we believe that the examples described in this chapter are merely the tip of the iceberg relative to the relationship among formal models, experiments, and political science. For as long as scholars who are knowledgeable about political contexts want models to be closer to facts or built from premises with more psychological or sociological realism, there will be demand for bridges between the logic of the models and the world in which we live. Experiments are uniquely positioned to serve as the foundations of those bridges.

References

- Austen-Smith, David, and Arthur Lupia. 2007. "Information in Elections." In *Positive Changes in Political Science: The Legacy of Richard McKelvey's Most Influential Writings*, eds. John H. Aldrich, James E. Alt, and Arthur Lupia. Ann Arbor, MI: University of Michigan Press.
- Black, Duncan. 1948. "On the Rationale of Group Decision-making." *Journal of Political Economy* 56: 23-34.
- Cacioppo, John T., and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42: 116-31.
- Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and its Discontents*, ed. David E. Apter. New York, NY: The Free Press of Glencoe.
- Feddersen, Timothy, and Wolfgang Pesendorfer. 1998. "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting." *American Political Science Review* 92: 23-35.
- Ferejohn, John A., Morris P. Fiorina, and Edward Packel. 1980. "Non-equilibrium Solutions for Legislative Systems." *Behavioral Science* 25: 140-48.
- Fiorina, Morris P., and Charles R. Plott. 1978. "Committee Decisions under Majority Rule: An Experimental Study." *American Political Science Review* 72: 575-598.
- Frohlich, Norman, and Joe A. Oppenheimer. 1992. *Choosing Justice: An Experimental Approach to Ethical Theory*. Berkeley, CA: University of California Press.
- Guernaschelli, Serena, Richard D. McKelvey, and Thomas R. Palfrey. 2000. "An Experimental Study of Jury Decision Rules." *American Political Science Review* 94: 407-23.

- Kalish, G.K., J.W. Milnor, J. Nash, and E.D. Nehrig. 1954. "Some Experimental n-Person Games." In *Decision Processes*, eds. Robert M. Thrall, Clyde H. Coombs, and Robert L. Davis. New York, NY: John Wiley.
- Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* New York: Cambridge University Press.
- Lupia, Arthur, and Jesse O. Menning. 2009. "When Can Politicians Scare Citizens Into Supporting Bad Policies?" *American Journal of Political Science* 53: 90-106.
- Lupia, Arthur, Adam Seth Levine, and Natasha Zharinova. 2010. "Should Political Scientists Use the Self-Confirming Equilibrium Concept? Benefits, Costs and an Application to Jury Theorems." *Political Analysis* 18: 103-23.
- McGuire, William J. 1985. "Attitudes and Attitude Change." In *Handbook of Social Psychology*, Volume II (3rd Edition), eds. Gardner Lindzey, and Elliot Aronson. New York: Random House.
- McKelvey, Richard D. 1976. "Intransitives in Multi dimensional Voting Models and Some Implications for Agenda Control." *Journal of Economic Theory* 18: 472-82.
- McKelvey, Richard D., and Peter C. Ordeshook. 1978. "Competitive Coalition Theory." In *Game Theory and Political Science*, ed. Peter C. Ordeshook. New York: New York University Press.
- McKelvey, Richard D., and Peter C. Ordeshook. 1979. "An Experimental Test of Several Theories of Committee Decision Making Under Majority Rule." In *Applied Game Theory*, eds. Steven J. Brams, A. Schotter, and G. Schwodiauer. Wurzburg, Ger: Springer-Verlag.
- McKelvey, Richard D., and Peter C. Ordeshook. 1980. "Vote Trading: An Experimental Study." *Public Choice* 35: 151-84.
- McKelvey, Richard D., and Peter C. Ordeshook. 1983. "Some Experimental Results that Fail to Support the Competitive Solution." *Public Choice* 40: 281-91.
- McKelvey, Richard D., and Peter C. Ordeshook. 1990. "A Decade of Experimental Research on Spatial Models of Elections and Committees." In *Government, Democracy, and Social Choice*, eds. Melvin J. Hinich, and James Enelow. Cambridge: Cambridge University Press.
- McKelvey, Richard D., Peter C. Ordeshook, and Mark D. Winer. 1978. "The Competitive Solution for N-Person Games Without Transferable Utility, with an Application to Committee Games." *American Political Science Review* 72: 599-615.
- McKelvey, Richard D., and Thomas R. Palfrey. 1992. "An Experimental Study of the Centipede Game." *Econometrica* 60: 803-36.

- McKelvey, Richard D., and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10: 6-38.
- McKelvey, Richard D., and Thomas R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics* 1: 9-41.
- Nash, John F. 1997. *Essays on Game Theory*. New York: Edward Elgar.
- O'Keefe, Daniel J. 1990. *Persuasion: Theory and Research*. Newbury Park, CA: Sage Publications.
- Ordeshook, Peter C. 2007. "The Competitive Solution Revisited." In *Positive Changes in Political Science: The Legacy of Richard D. McKelvey's Most Influential Writings*, eds. John H. Aldrich, James E. Alt, and Arthur Lupia. Ann Arbor, MI: University of Michigan Press.
- Palfrey, Thomas R. 2007a. "Laboratory Experiments." In *The Oxford Handbook of Political Economy*, eds. Barry R. Weingast, and Donald Wittman. New York: Oxford University Press.
- Palfrey, Thomas R. 2007b. "McKelvey and Quantal Response Equilibrium." In *Positive Changes in Political Science: The Legacy of Richard D. McKelvey's Most Influential Writings*, eds. John H. Aldrich, James E. Alt, and Arthur Lupia. Ann Arbor, MI: University of Michigan Press.
- Pennington, Nancy, and Reid Hastie. 1990. "Practical Implications of Psychological Research on Juror and Jury Decision Making." *Personality and Social Psychology Bulletin* 16: 90-105.
- Pennington, Nancy, and Reid Hastie. 1993. "Reasoning in Explanation-based Decision Making." *Cognition* 49: 123-63.
- Petty, Richard E., and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.
- Rapoport, Anatol, and Albert M. Chammah. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, MI: University of Michigan Press.
- Riker, William H. 1962. *The Theory of Political Coalitions*. New Haven, CT: Yale University Press.
- Riker, William H. 1967. "Bargaining in a Three-Person Game." *American Political Science Review* 61: 642-56.
- Riker, William H., and William James Zavoina. 1970. "Rational Behavior in Politics: Evidence from a Three Person Game." *American Political Science Review* 64: 48-60.

Rosenthal, Robert. 1981. "Games of Perfect Information, Predatory Pricing, and the Chain Store." *Journal of Economic Theory* 25: 92-100.

Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, eds. John H. Kagel, and Alvin E. Roth. Princeton, NJ: Princeton University Press.

Schofield, Norman. 1978. "Instability of Simple Dynamic Games." *Review of Economic Studies* 45: 575-94.

Von Neumann, John, and Oscar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

8. The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation

Paul M. Sniderman

First, a confession. The title promises a chapter about methods. But here, as everywhere, my concerns are substantive, not methodological. Still, what one wants to learn and how one ought to go about learning it are intertwined. So I propose to bring out the logic of the survey experiment by presenting a classification of survey experiment designs. Specifically, I will distinguish three designs: manipulative, permissive, and facilitative. The distinctions among the three designs turn on the hypotheses being tested, not the operations performed, and, above all, on the role of predispositions. The first design aims to get people to do what they are not predisposed to do; the second to allow them to do what they are predisposed to do, without encouraging them to do it; and the third to provide them with a relevant reason to do what they already are predisposed to do. Against the background of this three-fold classification, I want to comment briefly on some issues of causal inference and external validity, then conclude by offering my own view on the reasons for the explosive growth in survey experiments in the study of public opinion.

A personal story comes first, though. The modern survey experiment is the biggest change in survey research in a half century. There is some interest in how it came about, I am told. So I shall begin by telling how I got the idea of computer-assisted survey experiments. I excuse this personal note partly on the basis that the editors requested it but more importantly on the grounds that writing it allows me to acknowledge publicly the contributions of others.

1. The Logic of Discovery

Experiments have been part of public opinion surveys for a good many years. But they took the form of the so-called split ballot: The questionnaire would be printed in two versions: test questions would appear in each, identical in every respect but one. A more procrustean design is difficult to imagine. So it is the more compelling testimony to the ingenuity of researchers that they managed to learn a good deal all the same.¹ But what they learned, though of advantage for the applied side of public opinion research, was of less value for the academic side. Indeed, if one wanted to be waspish, one could argue that this first generation of survey experiments did more harm than good. By appearing to show that even trivial changes in question wording could produce profound changes in responses, they contributed to a Zeitgeist that presumed that citizens did not really have genuine attitudes and beliefs. And, by the sheer repetitiveness of the split-ballot design, they reinforced in the minds of several generations of subsequent researchers that survey experiments had to fit the straight jacket of the two -- and only two -- conditions.

Partly for those reasons, but partly also because my interest is substantive rather than methodological, the idea that survey experiments could be a useful tool did not enter my head. The idea came to me by a much more circuitous route. The coffee machine at the Survey Research Center (SRC) at the University of California Berkeley was located on the second floor. Since there was only one machine, whoever wanted coffee had to go there. One day, in 1983, a cup of coffee was just what I wanted. The line was long and directly behind me was Merrill Shanks. He was pumped up. So I asked him what was happening. Merrill is basketball tall; I less so, it would be fair to say. So it must have been quite a sight, Merrill towering over me, gesticulating with excitement, explaining that he had succeeded in writing a general purpose computer-assisted interviewing program, and illustrating with (physical, as well as mental) gusto

the measure of the breakthrough.ⁱⁱ I was thrilled for my friend's achievement, even if quite uninterested in the achievement itself. Computer-assisted interviewing passed right through — and out of — my mind.

A year later, we took our children to spend a year in Toronto, living in my in-laws' house, so that they would know their grandparents, and their grandparents would know them, not as children parachuted in from California for a brief stay, with their grandmother placing vats of candy by their bedsides, but as a family living together. It seemed like a good idea and once again I learned the danger of good ideas. Our children were heart-broken at returning to California. That was the downside.

There was also an upside. Living with one's in-laws, however welcoming they are, is an out-of-equilibrium experience. I mention this only because it says something about the social psychology of discovery. I do not believe I would have had the break-through idea about computer-assisted survey experiments but for the sharp and long break with the everyday routine. Among other things, it allowed the past to catch up with the present.

As a child, I went to a progressive summer camp. After a day of games on land and water, we would be treated to a late afternoon lecture in the rec hall on issues of social importance. Discrimination and prejudice are quite different things, that was one of the lessons we were taught. Prejudice is how they feel about us (that is, Jews). Discrimination is how they treat us. Prejudice is a bad thing. But how they feel about us is nowhere near as important as how they treat us. Covenants against Jews buying property in 'protected' areas, bans on membership in clubs, quotas on university admission were the norm then.ⁱⁱⁱ That was then and now is now. But between then and now, the memory of the lecture on the difference between prejudice and discrimination would regularly recur, and I would just as regularly be struck by the frustrating

irony that I had enlisted in a vocation, survey research, that could study prejudice (attitudes) but could not study discrimination (action). That persisting frustration was why the idea that struck me on my walk struck me with such force, I believe.

Here was the idea. The first computer-assisted interviewing program was purpose-built for question sequencing. Depending on the answer that a respondent gave to the first question in a series, the interviewer's screen would automatically light up with the next question that it would be appropriate to ask; in turn, depending on the answer that he gave to the second question, the screen would light up with the next question that it would be appropriate to ask; and so on. The depth of Shanks' achievement, though, was that he had transformed the opinion questionnaire into a computer program, which could be put to many uses. His was question sequencing; mine was randomized experiments.

I saw that, with one change, inserting a random operator to read a computer clock, computer-assisted interviewing could provide a platform for randomized experiments. Depending on the value of the random operator, questions could be programmed to appear on an interviewer's monitor, varying their wording, formatting, and order. The procedure would be effortless for the interviewer. She need only ask the form of the question that appeared on her monitor. And it would be invisible to the respondent. The fact that there were multiple versions of the question, randomly administered, would be invisible to the interviewees, since they would be asked only one.^{iv}

Voila! There was the -- broad -- answer to the summer camp lecture on the distinction between prejudice and discrimination. Ask a randomly selected set of respondents: How much help should the government give a white American who has lost his job in finding another? Ask the others exactly the same question except that it is a black American who has been laid off. If

more white Americans back a claim to government assistance if the beneficiary is white than black, we are capturing not merely how they feel about black Americans but how they treat them.

That was the idea. I remember the street that I was on, the house that I was looking at, when I had it. And absolutely nothing would have come of it but for Tom Piazza. Tom and I had seen each other around the halls of the SRC for years, but the main thing we knew about each other is that we shared an interest in the analysis of racial attitudes (see Apostle et al. 1983). Blanche DuBois relied on the kindness of strangers. I have relied on their creativity and character. Tom was the one who made computer-assisted randomized experiments work. Every study that I have done since, we have done together, whether his name has appeared on the project or not.

Childhood memories, disruption of routines, social science as a collaborative enterprise, technology as door-opening, research centers as institutionalized sources of ecological serendipity — those are the themes of the first part of my story of the logic of discovery. The theme of the second part of my story is a variation on Robert Merton's classic characterization of the communist — his word, not mine -- character of science.^v

We (Tom and I) had a monopoly position. Rather than take advantage of Merrill's breakthrough, the Institute for Social Research (ISR) at the University of Michigan and the National Opinion Research Center (NORC) at the University of Chicago attempted — for years - - to write their own computer-assisted program. Bad decision for them. They failed. An ideal outcome for us, you might think. Only studies done through the Berkeley SRC could exploit the flexibility of computer-assisted interviewing in the design of randomized experiments; which meant that we would have no competition in conducting survey experiments for years into the

future. Merton was right about the communist character of science, however. We would succeed. But we would succeed alone. And if we succeeded alone, we would fail. If other researchers could not play in our sandbox, they would find another sandbox to play in, and our work would always be at the margins.

The — properly — communist character of survey research showed up in a second way. Public opinion surveys are expensive. So not many got a chance to do them. Then Warren Miller effected the biggest ever structural change in the study of public opinion and elections. Consistent with Merton's doctrine of communism, Miller made the data of the flagship voting studies shared scientific property. In my judgment, there cannot be enough parades in his honor. It was quantitative analysis that shot ahead, however, not the design of surveys.^{vi} The American National Election Studies (ANES) offered opportunity for some innovation in measurement. But its overriding obligation was to time series. Continuity of design was the primary value, innovation in design a secondary one.

I had had my chance to come to bat in designing a study, actually two studies,^{vii} before I began thinking, how could other political scientists with new ideas get their chance to swing at the ball? The first article that we succeeded in publishing using randomized experiments gave me the idea. The article was built on the analysis of two experiments. But each of the experiments was only a question, admittedly a question that came in many forms, but at the end of the day only one question; which is to say, it took only about 30 seconds to administer. It then came to me that an interview of standard length could be used as a platform for multiple investigators, with each having time for 2 to 4 experiments, each having access to a common pool of right-hand side variables. Each would be a principal investigator. If their experiments were a success, they would be a success. If not, at least they had a chance to swing at the ball.

This idea of a shared platform for independent studies was the second best design idea that I have had. It made it possible for investigators, in the early stages of their careers, to do original survey research without having to raise the money.^{viii} But how to identify who should have the chance? A large part of the motivation is that very few had had an opportunity to distinguish themselves through the design of original studies. My solution: I shamelessly solicited invitations to give talks at any university that would have me, in order to identify a pool of possible participants; I then invited them to write a proposal, on the understanding that their idea was theirs alone but the responsibility for making the case to the National Science Foundation was mine. My sales pitch: we'll do thirteen studies for the price of one. That was the birth of the Multi-Investigator Project. It is Karen Garret, who directed the Multi-Investigators, who deserves all the credit for making them a success.

I have one more personal note to add. The Multi-Investigator ran two waves. The day that I received the grant from NSF for the second wave, I made a decision. I should give up the Multi-Investigator. Gatekeepers should be changed, I had always believed, and that applied to me, too. Diana Mutz and Arthur Lupia were the obvious choices. As the heads of Time-Sharing Experiments in the Social Sciences (TESS), they transformed the Multi-Investigator. To get some order of the magnitude of the difference between the two platforms, think of the Multi-Investigator as a stage coach and TESS as a Mercedes Benz truck. Add the support of the National Science Foundation, particularly through the Political Science Program, the creativity of researchers, the radical lowering of costs to entry through cooperative election studies, and survey experiments have become a standard tool in the study of public opinion and voting surveys. There is not a medal big enough to award Lupia and Mutz that would do justice to their achievements.

What is good fortune? Seeing an idea of yours travel the full arc, from being viewed at the outset as ridiculous to becoming in the end common place.^{ix} And my sense of the idea has itself traveled an arc. Gradually, I came to view it is a tool to do another.

2. A Design Classification

To bring out the explanatory roles of survey experiments in the study of public opinion, I shall distinguish among three design templates for survey designs: manipulative, permissive, and facilitative.^x

Manipulative Designs

Standardly, the distinction between observational and experimental designs parallels the distinction between those that are representing and intervening (Hacking 1983). Interventions or manipulations are the natural way to think of the treatment condition in an experiment. How does one test a vaccine?^{xi} By intervening on a random basis, administering a vaccine to some patients and a placebo to others, and noting the difference in outcome between the two. Moreover, the equation of intervention and manipulation seemed all the more natural against the background understanding of public opinion a generation ago. Knowing and caring little about politics, the average citizen arranged her opinions higgledy-piggledy (the lack of constraint problem), even supposing that she had formed some in the first place (the nonattitudes problem), the *reductio* of this conception of public opinion being the claim that “most” people lacked attitudes on “most” issues, preferring instead “to make it up as they go along.”^{xii} What, then, was the role of survey experiments? To demonstrate how easily one could get respondents to do what they were not predisposed to do.

The first generation of “framing” experiments are a poster child example of a *manipulative* design (e.g., Zaller 1992; Nelson and Kinder 1996). In one condition, a policy was framed in a

way to evoke a positive response; in the other, the same policy was framed in a way to evoke a negative response. And, would you believe, the policy enjoyed more support in the positive framing condition and evoked more opposition in the negative one. The substantive conclusion that was drawn was that the public was a marionette, its strings could be pulled for or against a policy by controlling the frame. But this is to tell a story about politics with the politics left out. The parties and candidates battle over how policies should be framed just as they battle over the positions that citizens should take on them.¹³ So Theriault and I carried out a pair of experiments replicating the positive and negative conditions of the first generation of framing experiments, but adding a third condition in which both frames were presented, and a fourth in which neither appeared (Sniderman and Theriault 2004). The first two conditions replicated the findings of the first generation of framing experiments. But the third led to a quite different conclusion. Confronted with both frames in the experiment (as they typically would be in real life, if not simultaneously, then in close succession), rather than being confused and thrown off the tracks, respondents are better able to pick the policy alternative closest to their general view of the matter. Druckman has pried this small opening into a seminal series of studies on framing (e.g., Druckman 2001a, b, c; Druckman 2004; Chong and Druckman 2007a, b; Druckman et al. 2010). In the areas in which I have research expertise, I am hard pressed to think of another who has, step by step, progressively deepened our understanding of a focal problem.

Survey experiments employing a manipulative design can be of value. But my own approach to the logic and, therefore, the design of survey experiments, travels in the opposite direction. To overstate, my premise is that you can get people to do in a survey experiment mainly what they already are willing to do – which is fortunate, since this what we want to learn after all. This premise is the rationale for the next two designs: permissive and facilitative.

Permissive Designs

A manipulative design aims to get respondents to do what they are not predisposed to do. In contrast, a *permissive design* aims to allow respondents to do what they are predisposed to do without encouraging them to do it. The strategy is to remove, rather than apply, pressure to favor one response alternative over another. Think of this as experimental design in the service of unobtrusive measurement (Webb et al. 1996).

A showpiece example of a permissive design in survey experiments is the List Experiment.¹⁴ The measurement problem is this: Can one create a set of circumstances, in which a person being interviewed can express a potentially objectionable sentiment without the interviewer being aware that she has expressed it?¹⁵ Kuklinski's creative insight: to devise a question format that leads respondents to infer, correctly, that the interviewer cannot tell which responses they have made, but the data analyst can determine ex post the proportion of them making a particular response. To give a hyper-simplified description of the procedure, in the baseline condition, the interviewer begins by saying: "I am going to read you a list of some things that make some people angry. I want you to tell me how many make you angry. Don't tell me which items make you angry. Just how many." The interviewer then reads a list of, say, four items. In the test condition, everything is exactly the same, except that the list now has one more item, say, affirmative action for blacks. To determine the proportion of respondents angry over affirmative action, it is necessary only to subtract the mean angry responses in the baseline condition from the mean angry responses in the test condition, then multiply by 100. Characteristics of respondents that increase (or decrease) the hit rate can be identified iteratively.

This type of design I baptize permissive because it allows respondents to make a response without encouraging, or inducing, or exerting pressure on them to do so. So it is with the List

Experiment. Why do some respondents respond with a higher number in the treatment condition than in the baseline condition? Because they were predisposed to do so. They are they are angry over affirmative action, and found themselves with the opportunity to express their anger, believing (correctly) that the interviewer had no way to know that they had done so, without realizing that a data analyst could deduce the proportion expressing anger ex post.

How should we conceive of the logic of a permissive design like the List Experiment? Baseline and test conditions were the words I used to refer to the two conditions in our hyper-simplified example of the List Experiment.¹⁶ The baseline condition corresponds to the natural understanding of the control condition. But in what sense is the “test” condition a “treatment” condition? It entails exposure to a stimulus – affirmative action. Affirmative action is a provocative stimulus, one could argue. But to make this argument would be to miss the point. If a person is indifferent to affirmative action, or sympathetic to it, or simply ignorant of it, the mere mention of affirmative action will not evoke an angry response. To evoke an angry response, it is necessary that he already is angry about it.

A second example of a permissive design comes from a celebrated series of studies on risk aversion by Tversky and Kahneman. They have demonstrated that people have strikingly different preferences on two logically equivalent choices, depending on whether the choice is framed in terms of gains or losses. Their Asian Flu Experiment is a paradigmatic example. People are far more likely to favor exactly the same course of action if the choice alternatives are posed in terms of lives saved as opposed to lives lost. This result, labeled risk aversion, is highly robust. With an ingenious design, Druckman carried out an experiment that had two arms: one matched the Kahneman-Tversky design, the other added credible advice, in the form of endorsements of a course of action by political parties (Druckman 2001a). The key finding:

partisans take their cue from party endorsements, so much so that the gain-loss framing effect virtually disappears. I want to make two points with this example. First, framing effects are robustly found between choices that are logically equivalent, depending on whether the choices are framed in terms of gains or losses, in the absence of other information to exploit. Second, the observed effect is not a function of an experimental intervention in the form of an application of pressure on a respondent to respond in a particular direction. It is instead a matter of allowing people to respond as they are predisposed to respond without encouraging them to do so.

Facilitative Designs

The third type of design for survey experiments I shall christen *facilitative*. Permissive designs aim to allow respondents to do what they are predisposed to do without encouraging them to do it. Manipulative designs aim to get people to do what they are not predisposed to do. Like permissive designs but unlike manipulative ones, facilitative designs do not involve the use of coercive or impelling force. Unlike permissive designs, facilitative designs involve a directional force. Unlike manipulative designs, facilitative ones involve a directional force, in the form of a relevant reason to do what people already are predisposed to do.

This notion of a relevant reason is a tip-off to a primary use of survey experiments for the study of public opinion, I have become persuaded. Let me illustrate what I mean by the notion of a relevant reason with an experiment designed by Laura Stoker (Stoker 1998). The aim of this experiment is to determine the connection between support for a policy and the justification provided for it. Stoker picks affirmative action in its most provocative form – mandatory job quotas.

This in-your-face formulation policy frame should trigger the emotional logic that Converse (1964) argued underlies ‘reasoning’ about racial policies in general. How one feels

about blacks, he hypothesized, is the key to understanding why whites tend to line up on one or the other side of racial policies across the board. Feel negatively about blacks, and you will oppose policies to help them; feel positively, and you will support them. Stoker's experiment opens a new door on policy reasoning, though. It investigates the persuasive weight of two different reasons for mandatory quotas. Stoker's results show that one reason, the underrepresentation of blacks, counts as no reason at all – that is, there is no difference between deploying it as a justification and not deploying a justification at all. In contrast, the other reason, a finding of discrimination, counts as a relevant reason indeed – that is, it markedly increases support for affirmative action even framed in its most provocative form. Stoker's discovery is not the common sense idea that policy justifications can make a difference. It is rather the differentiation of justifications that makes a difference. There is a world of difference between declaiming that fairness matters and specifying what counts as fairness.

As a second example of facilitation, consider the counter-argument technique. The counter-argument technique was introduced in Sniderman and Piazza 1993 and explored further in Sniderman et al. 1996. Gibson has made it a central technique in the survey researchers' toolkit, deploying it in a remarkably ambitious series of survey settings.¹⁷ The first generation of counter-arguments only comprises quasi-experiment, however. The counter-argument presented to reconsider support for a policy is (naturally enough) different from the one presented to reconsider opposition to it. Hence the relevance of the second generation of the counter-argument experiments (Jackman and Sniderman 2006). Respondents take a position on an issue, then are presented with a reason to reconsider. What should count as a reason to reconsider, one may reasonably ask, and what more exactly are people doing when they are reconsidering their initial position? Two content-laden counter-arguments are administered. One presents a

substantive reason for respondents who have supported more government help to renounce this position, while the other provides a substantive reason for respondents who have opposed it to renounce their position (Jackman, Simon, and Sniderman 2006). In addition, a content-free counter-argument – that is, an objection to the position that respondents have taken that has the form of an argument but not the specific substantive content of one¹⁸ – also is administered. Thus, one half of the respondents initially supporting the policy get a content-laden counter-argument, one half a content-free one. Ditto for respondents initially opposing the policy.

There are two points I would make. The first is that respondents at all levels of political sophistication discriminate between a genuine reason (that is, an argument that provides a substantive argument to reconsider) and a pseudo reason (that is, an argument that merely points to the uncertainty of taking any position). Twice as many report changing their minds in response to a content-laden, rather than a content-free, counter-argument. There is, in short, a difference between getting an argument and getting argued with. The second point is that the bulk of those changing in the face of a content-laden counter-argument had taken a position at odds with their general view of the matter. What work, then, was the content-laden counter-argument doing? Most who change their initial position in response to a content-laden counter-argument had good reason to change. The side of the issue they had initially chosen was inconsistent with their general view of the matter.¹⁹ They were rethinking their initial position by dint of a reason that, from their point of view, should count as a reason to reconsider their position. In reconsidering, they were not changing their mind; they were correcting a misstep. What, then, was the experimental intervention accomplishing? It was facilitating their reconsideration of the position they had taken, in light of a consideration that counted as a relevant reason for reconsideration, given their own general view of the matter.

3. Experimental Treatments and Political Predispositions

In a pioneering analysis of the logic of survey experiments, Gaines, Kuklinski, and Quirk (2007) bring to the foreground a neglected consideration. Respondents do not enter public opinion interviews as blank slates. They bring with them the effects of previous experiences. Gaines et al. refer to the enduring effects of previous experience as pre-treatment. In their view, understanding how pre-treatments condition experimental responses is a precondition of understanding the logic of survey experiments. This is a dead-on-target insight. In my view, it is an understatement. The purpose of survey experiments in the study of public opinion is precisely to understand pre-treatment – or, as I think of it, previous conditioning

The Null Hypothesis

What is the null hypothesis in a survey experiment? It sounds odd to ask this, I acknowledge. Textbooks drill into us a uniform understanding of the null: the absence of a difference between responses in the treatment and control conditions. From which it follows that the purpose of an experimental treatment is to produce a difference in the treatment condition; and if it fails to do this, the experiment has failed. So it is commonly – and wrongly -- supposed.²⁰

To bring out the logic of the problem, I enlist the SAT Experiment (Sniderman and Piazza 2002). African Americans have their own culture, it is claimed (Dawson 2001). There is a positive sense in which this claim may be true. But there is negative sense in which it is false. The values of the American culture are as much the values of African Americans as of white Americans.²¹

To test this hypothesis of shared values, respondents, all of whom are black, are told of two young men, one black and the other white. Only one of the two can be admitted. The young white man's college entrance exam score is always 80; the young black man's exam score is

(randomly) 55, 60, 65, 70, and 75.²² Respondents are asked which of the two young men should be admitted, if the college can admit only one.

Our hypothesis was that African Americans share the core values of the common culture. So far as they do, they should choose the young white man, since he always has the higher exam score. On the other hand, it surely is a reasonable expectation that African Americans will take account the continuing burden of discrimination. The question then is, how small does the difference in scores between the two young men need to be in order to be regarded as negligible for African Americans to give the nod to the black candidate on other grounds – for example, the fact that they have to overcome obstacles that whites do not. We worked to establish feet-in-cement expectations, recruiting a sample of experts to pick the point at which a majority of African Americans would favor the black candidate. Seventy-five percent of our experts picked a difference of just ten points to be so small as to wave away against the historic and continuing injustices done to blacks. And 100 percent of them predicted that a difference of only five points would be judged as insignificant. In fact, even when the difference in scores is smallest, the overwhelming number of African Americans picked the white candidate.²³

The hypothesis is that African Americans share the core values of the American culture. If true, they should overwhelmingly favor the candidate with the higher exam score -- even though that always means favoring a white candidate over a black candidate. The null hypothesis, then, is that responses in the treatment and the control conditions should not differ. In fact, whether the difference between candidates' SAT scores was large or small, they were equally likely to favor the high scorer – even though the high scorer in the experiment always was the white student. It is difficult for us to conceive of a more compelling demonstration of the commitment of African

Americans to the value of achievement.²⁴ Nor, at a lower rhetorical register, to imagine a better example of an absence of a treatment effect being evidence for a substantive hypothesis.

Interactions

God made the world additive, a simplifying assumption I recommend for theoretical self-discipline. But like all simplifying assumptions, it oversimplifies. Consider one of the first survey experiments that Tom Piazza and I conducted, the Laid-off Worker Experiment (Sniderman and Piazza 1993).

The question that the Laid-off Worker Experiment was designed to investigate was whether political conservatives discriminate against African Americans. Are they as willing to honor a claim for government assistance made by a white American than by a black American? But framing the question broadly obscures the real question, we reasoned. Supposing that being black made a difference to conservatives, what is it about being black that makes a difference? Three stigmatizing characterizations of blacks stood out: “lazy” blacks; unmarried black mothers; young black (stereotypically aggressive) males. Accordingly, in the Laid-off Worker Experiment, respondents are told about a person who has lost his or her job and asked how much help the government should give them in finding another. The race of the person who has been laid off randomly varied, naturally. But so too is the gender, age, marital-parental status, and work history (dependable versus undependable).

And when the data were analyzed, what should pop up but the finding that political conservatives are more, not less, likely to favor government assistance for a black worker who has lost his or her job than a white worker. “Pop up” is not a scientific term, I recognize. But it would be a scam to imply that we had anticipated that conservatives would go all-out for out-of-work blacks. We had a reasonable expectation that conservatives would be harder on blacks than

on whites. We had never expected to find that they would respond with more sympathy and more support for a black who had lost his job than for a white who similarly found himself on the street. Nor had anyone else. The result would discredit the whole idea of using randomized experiments in public opinion, I feared. Days of frenzied analysis followed. On the fourth day, Tom Piazza and I solved the puzzle. It was not blacks in general that evoked an especially supportive response from political conservatives: it was hard-working blacks distinctively. And why did conservatives respond to a hard-working black? Precisely because, for them, a hard-working black was the exception. So they wanted to make an exception for them, and have the government help them find another job. So we argued in our initial study, and so we cross-validated in a follow-up.²⁵

From this experience, I draw two methodological lessons. We had designed the experiment to test the hypothesis that conservatives racially discriminate (and would have had a blessed-on-all-sides career had the Laid-off Worker Experiment done the job that we thought it would do.) But the result was nothing like we had anticipated. And that is the first methodological lesson. Surprise is a cognitive emotion. And just because the design of experiments requires a definition of expectations, experiments can surprise in a way that observational analysis cannot. Hypotheses precede experiments rather than the other way about: that is the reason that each is designed the way it is and not some other. The second methodological point I would make is that the expression “split-half” should be banished. The presumption that survey experiments can have only two conditions has handcuffed survey experimenters. Complexity is not a value in and of itself. To say that an experiment has the right design is to say that it is set up in the right way to answer the question it is designed to answer. And computer-assisted surveys are a breakthrough, in among other respects, because of the plasticity of the designs that they permit.

Survey Experiments and Counterfactual Conditionals: Majorities and Counter-Majorities under the Same Equilibrium Conditions

The principal business of survey experiments is to reveal what people already are predisposed to do, I have argued. Ironically, this means that they can put us in a position to explore possible worlds. An example of possible worlds will make clear what I have in mind.

With Ted Carmines, I investigated a hypothesis about the potential for a breakthrough in public support for policies to assist blacks. Researchers of symbolic racism maintain that racial prejudice has a death grip on the American mind. In their view, for the grip of racism to weaken, nothing less than a change in the hearts and minds of white Americans was necessary. In contrast, we thought there was a political opening. Revive the moral universalism of the civil rights movement, we reasoned, and a winning coalition of whites and blacks could be brought into existence.

To test this conjecture, we carried out a pair of experiments, the Regardless of Race Experiment and the Color Blind Experiment (Sniderman and Carmines 1997).²⁶ Both experiments showed that support for policies that would help blacks is markedly higher if the arguments made on their behalf were morally universalistic, rather than racially particularistic. To be sure, conservatives are no more likely to support the policy when a universalistic appeal is made on its behalf than when a particularistic one is. But then again, why should they? They are being asked to support a liberal policy. Consistent with our hypothesis, moderates are markedly more likely to support the policy in the face of a universalistic, rather than a particularistic, appeal. Still more telling, so, too, are liberals.

This result illustrates a general point about politics and a specific one about racial politics. The general point is this: in politics, more than one winning coalition can exist under the same

equilibrium conditions. There is the majority that one observes, conditional on the available political alternatives. But there are the counter-majorities that one would observe, conditional on different alternatives or different reasons for choosing between the same alternatives. This claim of multiple majorities under the same equilibrium conditions goes further than the standard interpretation of Riker's heresthetics (Riker 1996). His claim is that bringing about a new winning coalition requires bringing a new dimension of cleavage to the fore. Thanks to experiments opening up exploration of possible worlds, one can see how a new winning coalition can be brought about without bringing a new dimension of cleavage to the fore.

The second point has to do with the politics of race. Many race specialists in political science have nailed their flag to the claim that a change in the politics of race requires a change in the core values of Americans, in order to establish a new majority on the issue of race. By contrast, our claim was that it was not necessary first to change the hearts and minds of white Americans in order to change the politics of race. A counter-majority ready to support a politics of race that was morally universalistic was in existence and already in position. It would be brought to the surface when a politician was ambitious and clever enough to mobilize it. It would go too far to say that our analysis predicted the Obama victory.²⁷ It does not go too far to say that that it is the only analysis of race and American politics that is consistent with it.

4. A Final View

The experimental method has made inroads on many fronts in political science, but why have survey experiments met with earlier and broader acceptance? Part of the answer to this question is straightforward. Survey experiments (and, when I say survey experiments, I include the whole family of interviewing modes, from face-to-face to telephone to web-based) have a lower hurdle to jump in meeting requirements of external validity. Lower does not mean low, I

would hastily add. A second part of the answer for the explosive growth in survey experiments is similarly straightforward. Research areas flourish in inverse proportion to barriers to entry. With the introduction of the Multi-Investigator studies, then the enormous advance of TESS as a platform for survey experiments, the marginal cost of conducting survey experiments plummeted. Cooperative election studies, providing teams of investigators the time to carry out autonomously designed studies, have become the third stage of this cost revolution.

The importance of both of these factors should not be underestimated. But a third factor is even more important than the first two, in my opinion. When it comes to survey experiments as a method for the study of politics, the ‘what’ that is being studied has driven the ‘how’ it is studied, rather than the other way round. It is the power of the ideas of generations of researchers in the study of public opinion and voting, incorporating theoretical frameworks from the social psychological to the rational, that has provided the propulsive force in the use of survey experiments in the study of mass politics.

References

- Apostle, Richard A., Charles Y. Glock, Thomas Piazza, and Marijane Suelze. 1983. *The Anatomy of Racial Attitudes*. Berkeley, CA: University of California Press.
- Bartels, Larry, and Henry E. Brady .1993. “The State of Quantitative Political Methodology.” In *The State of the Discipline II*, ed. Ada Finifter. American Political Science Association.
- Chong, Dennis, and James N. Druckman. 2007a. “Framing Public Opinion in Competitive Democracies.” *American Political Science Review* 101: 637-55.
- Chong, Dennis, and James N. Druckman. 2007b. “Framing Theory.” *Annual Review of Political Science* 10: 103-26.
- Converse, Philip E. 1964. “The Nature of Belief Systems in Mass Publics.” In *Ideology and Discontent*. ed., David E. Apter. New York: Free Press.
- Dawson, Michael C. 2001. *Black Visions*. Chicago: University of Chicago Press.

- Druckman, James N. 2001a. "Using Credible Advice to Overcome Framing Effects." *The Journal of Law, Economics, & Organization* 17: 62-82.
- Druckman, James N. 2001b. "On The Limits Of Framing Effects." *Journal of Politics* 63: 1041-66.
- Druckman, James N. 2001c. "The Implications of Framing Effects for Citizen Competence." *Political Behavior* 23: 225-256.
- Druckman, James N. 2004. "Political Preference Formation." *American Political Science Review* 98: 671-86.
- Druckman, James N., Cari Lynn Hennessy, Kristi St. Charles, and Jonathan Weber. 2010. "Competing Rhetoric Over Time: Frames Versus Cues." *Journal of Politics* 72: 136-48.
- Freedman, David A. 2010. *Statistical Models and Causal Inference*, eds. David Collier, Jasjeet Sekhon, and Philip B. Stark. New York: Cambridge University Press.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.
- Gibson, James L., and Amanda Gouws. 2003. *Overcoming Intolerance in South Africa: Experiments in Democratic Persuasion*. New York: Cambridge University Press.
- Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. New York: Cambridge University Press.
- Hurwitz, Jon, and Mark Peffley. 1998. *Perception and Prejudice: Race and Politics in the United States*. New Haven, CT: Yale University Press.
- Jackman, Simon, and Paul M. Sniderman. 2006. "The Limits of Deliberative Discussion: A Model of Everyday Political Arguments." *Journal of Politics* 68: 272-83.
- Merton, Robert K. 1973. *The Sociology of Science*. Chicago, IL: University of Chicago Press.
- Nelson, Thomas E., and Donald R. Kinder. 1996. "Issue Frames and Group-Centrism in American Public Opinion." *Journal of Politics* 58: 1055-78.
- Riker, William H. 1996. *The Strategy of Rhetoric*. New Haven, CT: Yale University Press.
- Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.
- Sniderman, Paul M., and Edward G. Carmines. 1997. *Reaching Beyond Race*. Cambridge, MA: Harvard University Press.
- Sniderman, Paul M., Joseph F. Fletcher, Peter Russell, and Philip E. Tetlock. 1996. *The Clash of Rights: Liberty, Equality, and Legitimacy in Pluralist Democracies*. New Haven, CT: Yale University Press.

Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: Harvard University Press.

Sniderman, Paul M., and Thomas Piazza. 2002. *Black Pride and Black Prejudice*. Princeton, NJ: Princeton University Press.

Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion*, eds. Willem Saris, and Paul M. Sniderman. Princeton, NJ: Princeton University Press.

Stoker, Laura. 1998. "Understanding Whites' Resistance to Affirmative Action: The Role of Principled Commitments and Racial Prejudice." In *Perception and Prejudice: Race and Politics in the United States*, eds. Jon Hurwitz, and Mark Peffley. New Haven, CT: Yale University Press.

Webb, Eugene, J. Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 1996. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. New York: Rand McNally.

Zaller, John. R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

ⁱ Schuman and Presser (1981) is the seminal work.

ⁱⁱ It was an exceptional achievement. ISR at Michigan and NORC in Chicago, the two heavyweight champions of academic survey research, gave years and a treasure chest of man hours attempting to match the programming achievement of Merrill and his colleagues, only to fail.

ⁱⁱⁱ My father and father-in-law were among the first Jews permitted to attend the University of Toronto Medical School. My wife was a member of the first class of the University of Toronto Medical School in which the Jewish Quota was lifted.

^{iv} The contrast is with the then-common practice of asking a series of items, varying the beneficiary of a policy; i.e., Would you favor the program if it benefited a white American?, If it benefited a black American?, etc. I am also presuming, when I speak of the procedure being invisible to the respondent, the artful writing of an item.

^v When referring to communism, Merton meant the principle of common ownership of scientific discoveries. We do not have a right to the means to make scientific discoveries. But we have a right to share in them. And those who make them have a corresponding duty to allow us to share in them. See Merton (1973).

^{vi} For an overview of how much progress was made on how many fronts, see Bartels and Brady (1993).

^{vii} The first was the Bay Area Survey with Thomas Piazza, which led to Sniderman and Piazza (1993). The second was The Charter of Rights Study, which led to Sniderman et al. (1996). The third was the National Race and Politics Study (RAP), which led to, among many other publications, Sniderman and Carmines (1997) and Hurwitz and Peffley (1998). RAP was the trial run for the Multi-Investigator, and involved eight co-principal investigators.

^{viii} One of the benefits I did not anticipate was that, even if their first try had not succeeded, they had a leg up in writing a proposal for a full-scale study.

^{ix} My first proposal to the National Science Foundation to do survey experiments was judged by two of the reviewers to be farcical undertaking, one of whom took eight pages to make sure that his opinion of the project was clear.

^x The classification hinges on the aims of experiments. Since I know the hypotheses that experiments I have designed were designed to test, I shall (over) illustrate the principles using examples of experiments that my colleagues and I have conducted.

^{xi} See Freedman (2010), who offers the Salk vaccine test as a paradigm example of randomized experiment.

^{xii} For a detailed critique of this view of public opinion, see Sniderman (2001).

¹³ The idea of dual frames — or, to use Chong and Druckman’s term, competitive frames (Chong and Druckman 2007a) -- came to me while watching on television a Democratic campaign ad framing an issue to its advantage, immediately followed by a Republican ad framing the same issue to its advantage.

¹⁴ Here is one of the few times when I know for certain where an idea came from. I myself was a witness at the creation of List Experiment. During a planning session for the 1990 Race and Politics Study at the Circle 7 Ranch, I took Jim Kuklinski for a Jeep ride in the meadow. Suddenly, by the front gate, he stood up, exclaimed the equivalent of “Eureka,” and outlined the design of the List Experiment. I mention this for two reasons. First, to put on record that Kuklinski devised the List Experiment, easily the widely used survey experiment design; and second, to offer an historical example of the creativity of multi-investigator studies: the National Race and Politics Study had nine coprincipal investigators, and contributed more innovations than any previous study because of their power for innovation.

¹⁵ There is another possibility, and a more likely one in my view. They do not wish to say openly that affirmative action makes them angry because doing so conflicts with their sense of themselves and their political principles; it violates a principle or image of themselves that they value (see Sniderman and Carmines 1997).

¹⁶ Again, by way of underlining the decisive difference between the straight-jacket or the split-ballot design and the plasticity of the computer-assisted interviewing, I would underline that the actual design of List Experiments tends to involve a number of test conditions, allowing for the comparison and contrast of, say, responses to African Americans becoming neighbors and asking for affirmative action.

¹⁷ For an especially fascinating example, see Gibson and Gouws (2003).

¹⁸ The wording of the content-free counter-argument in this study is: “However, if one thinks of all the problems this is going to create ...”

¹⁹ Treatment and control groups were thus identically positioned. Analysis searching for asymmetrical effects conditional on being pro or con the policy failed to detect any.

²⁰ This is a costly view. Among other things, it produces a publication bias of experiments being regarded as succeeding when they produce differences and failing when they don’t.

²¹ This is an example of a descriptive as opposed to a causal hypothesis, though it should not be assigned second-class status on this account. The former is capable of being as enlightening as the latter, and better grounded by far.

²² Their social class (in the form of their father’s occupation) also is randomly varied.

²³ As a test of social desirability, we examined separately respondents interviewed by black interviewers, and they were even more likely to hew to the value of achievement than those interviewed by white interviewers.

²⁴ I am curious how many would like to bet that white Americans would show a similar measure of commitment to the value of achievement in an equivalent situation.

²⁵ See the Helping Hand Experiment (Sniderman and Carmines 1997).

²⁶ Designing experiments in pairs provides invaluable opportunities for replication in the same study.

²⁷ We had in mind an ambitious and gifted politician like President Clinton but alas, Monica Lewinsky prevented a test of our hypothesis. It never entered our heads that the country had so progressed that an African American could do so.

9. Field Experiments in Political Science

Alan S. Gerberⁱ

After a period of near total absence in political science, field experimentation is now a common research design. In this essay, I discuss some of the reasons for the increasing use of field experiments. Several chapters in this volume provide comprehensive introductions to specific experimental techniques and detailed reviews of the now extensive field experimental literatures in a variety of areas. This chapter will not duplicate these contributions, but instead provide background, arguments, opinions, and speculations. I begin by defining field experiments in Section 1. In Section 2, I discuss the intellectual context for the emergence of field experimentation in political science, beginning with the recent revival of field experimentation in studies of voter turnout. In Section 3, I discuss how field experiments address many of the common methodological deficiencies identified in earlier observational research on this topic. Section 4 reviews the range of applications of field experimentation. In Section 5, I answer several frequently asked questions about the limitations and weaknesses of field experimentation. In Section 6, I briefly discuss some issues that field experimentation faces as it continues to develop into one of the common methodological approaches in political science. This includes a discussion of the external validity of field experimental results and consideration of how difficulties related to replication and bias in experimental reporting might affect the development of field experiment literatures.

1. Definition

In social science experiments, units of observation are randomly assigned to groups and treatment effects are measured by comparing outcomes across groups.ⁱⁱ Random assignment

permits unbiased comparisons because randomization produces groups that, prior to the experimental intervention, differ with respect to both observable and unobservable attributes only due to chance. Field experiments seek to combine the internal validity of randomized experiments with increased external validity, or generalizability, gained through conducting the experiment in real-world settings. Field experiments aim to reproduce the environment in which the phenomenon of interest naturally occurs and thereby enhance the external validity of the experiment.

Experiments have many dimensions, including the type of subjects, the experimental environment, the treatments, the outcome measurements, and subject awareness of the experiment. The degree to which each of these dimensions parallels the real-world phenomenon of interest may vary, leading to a blurring of the distinction between what is and is not a field experiment. The economists Harrison and List propose a system for classifying studies according to their varying degrees of naturalism (Harrison and List 2004). According to their taxonomy, the least naturalistic experimental study is the *conventional lab experiment*. This type of study is the familiar laboratory experiment that involves an abstract task, such as playing a standard game (e.g., prisoner's dilemma, dictator game, etc.) and employs the typical student subject pool. The *artefactual field experiment* is a conventional laboratory experiment with a nonstandard subject pool. Examples of this work include Habyarimana and colleagues, who (among other things) investigate ethnic cooperation through an exploration of the degree of altruism displayed in dictator games. This study, which was conducted in Africa, drew its subjects from various ethnic groups in Kampala, Uganda (Habyarimana et al. 2007). The *framed field experiment* is the same as the artefactual field experiment, except the task is more naturalistic. An example is Chin, Bond, and Geva's study of the effect of money on access to members of Congress through an

experiment in which congressional staffers make scheduling decisions after being told whether or not the meeting is sought by a Political Action Committee (PAC) representative or a constituent (Chin, Bond, and Geva 2000). The *natural field experiment*, which is the design commonly referred to in political science as a “field experiment,” is the same as the framed field experiment except it involves subjects who naturally undertake the task of interest, in the natural environment for the task, and who are unaware they are participating in an experiment. Research in which political campaigns randomly assign households to receive different campaign mailings to test the effect of alternative communications on voter turnout is one example of a natural field experiment.

This chapter focuses on natural field experiments. Although the degree of naturalism in field experiments is the distinctive strength of the method, it is important to keep in mind that the goal of most experimental interventions is to estimate a causal effect, not to achieve realism. It might appear from the classification system that the movement from conventional lab experiment to natural field experiment is similar to “the ascent of man,” but that would be incorrect. The importance of naturalism along the various dimensions of the experimental design will depend on the research objectives and whether there is concern about the assumptions required for generalization. Consider the issue of experimental subjects. If the researcher aims to capture basic psychological processes that may safely be assumed to be invariant across populations, experimental contexts, or subject awareness of the experiment, then nothing is lost by using a conventional lab experiment. That said, understanding behavior of typical populations in natural environments is frequently the ultimate goal of social science research and it is a considerable, if not impossible, challenge to even recognize the full set of threats to external validity present in

artificial contexts, let alone to adjust the measured experimental effects and uncertainty to account for these threats (Gerber, Green, and Kaplan 2004).

2. Intellectual Context for Emergence of Field Experiments

General Intellectual Environment

The success of randomized clinical trials in medicine provided a general impetus for exploring the application of similar methods to social science questions. The first large-scale randomized experiment in medicine – the landmark study of the effectiveness of streptomycin in treating tuberculosis (Medical Research Council 1948) – appeared shortly after World War Two. In the years since, the use of randomized trials in clinical research has grown to the point where this method now plays a central role in the evaluation of medical treatments.ⁱⁱⁱ The prominence of randomized trials in medicine led to widespread familiarity with the method and appreciation of the benefits of the use of random assignment to measure the effectiveness of interventions.

With some important exceptions, such as the negative income tax experiments of the late 1960s and 1970s, there were relatively few social science field experiments prior to the 1990s. The increased use of field experimentation in the social sciences emerged from an intellectual climate of growing concern about the validity of the key assumptions supporting observational research designs and increased emphasis on research designs in which exogeneity assumptions were more plausible. During the mid-1980s, there was increasing appreciation in the social sciences, especially in economics, of the extreme difficulty in estimating causal effects from standard observational data (e.g., Lalonde 1986). Particularly in the field of labor economics, leading researchers began searching for natural experiments to overcome the difficulties posed by unobservable factors that might bias regression estimates. The result was a surge in studies

that investigated naturally occurring randomizations or near-randomized application of a “treatment”. Examples of this work include Angrist’s study of the effect of serving in the Vietnam War on earnings, where a lottery draw altered the likelihood of service (Angrist 1990), and Angrist and Krueger’s use of birthdates and minimum age requirements for school attendance to estimate the effect of educational attainment on wages (Angrist and Krueger 1991).

The Development of Field Experimentation in Political Science

The earliest field experiments in political science were performed in the 1920s by Harold Gosnell, who investigated the effect of get out the vote (GOTV) mailings in the 1924 presidential election and 1925 Chicago mayoral election (Gosnell 1927).^{iv} In the 1950s, Eldersveld conducted a randomized field experiment to measure the effects of mail, phone, and canvassing on voter turnout in Ann Arbor, Michigan (Eldersveld 1956). These pioneering experiments had only a limited effect on the trajectory of subsequent research. Field experimentation was a novelty and, when considered at all, was dismissed as impractical or of limited application (Gerber and Green 2008). The method was almost never used; there was no field experiment published in a major political science journal in the 1990s.

The recent revival of field experiments in political science began with a series of experimental studies of campaign activity (Gerber and Green 2000; Green and Gerber 2004). The renewed attention to field experimentation can be traced to persistent methodological and substantive concerns regarding important political behavior literatures. To explore the intellectual context for the revival of field experimentation in political science, I will briefly review the state of the literature on campaign effects at the time of the Gerber and Green New Haven experiment. This literature, in my view, includes some of the very best empirical political science studies of their time. However, although the research designs used to study campaign

spending effects and voter mobilization were often ingenious, these extensive literatures suffered from important methodological weaknesses and conflicting findings. Many of the methodological difficulties are successfully addressed through the use of field experimentation.

Consider first the work on the effect of campaign spending on election outcomes circa 1998, the date of the first modern voter mobilization experiment (Gerber and Green 2000). This literature did not examine the effects of specific campaign activities, but rather the relationship between overall Federal Election Commission (FEC) reported spending levels and candidate vote shares.^v There were three main approaches to estimating the effect of campaign spending on candidate vote shares. In the earliest work, Jacobson and others estimated spending effects using ordinary least squares regressions of vote shares on incumbent and challenger spending levels (e.g., Abramowitz 1988; Jacobson 1978, 1985, 1990, 1998). This strategy assumes that spending levels are independent of omitted variables that also affect vote share. Concern that this assumption was incorrect was heightened by the frequently observed *negative* correlation between incumbent spending and incumbent vote share. In response to this potential difficulty, there were two main alternative strategies. First, some scholars proposed instrumental variables for candidate spending levels (e.g., Green and Krasno 1988; Ansolabehere and Snyder 1996; Gerber 1998).^{vi} Second, Levitt examined the performance of pairs of candidates who faced each other more than once. The change in vote shares between the initial contest and rematch were compared to the changes in candidate spending between the initial contest and rematch, a strategy which serves to difference away difficult to measure district- or candidate-level variables that might be lurking in the error term (Levitt 1994).

Unfortunately, the alternative research designs produce dramatically different results. Table 9-1 reports, in dollar per vote terms, the cost per additional vote implied by the alternative

approaches. The dollar figures listed are the cost of changing the vote margin by one vote.^{vii} Table 9-1 illustrates the dramatic differences in the implications of the alternative models and underscores how crucial modeling assumptions are in this line of research. Depending on the research design, it is estimated to cost as much as 500 dollars or as little as 20 dollars to improve the vote margin by a single vote (Gerber 2004). However, it is not clear which estimates are most reliable, as each methodological approach relies on assumptions that are vulnerable to serious critiques.^{viii} The striking diversity of results in the campaign spending literature, and the sensitivity of the results to statistical assumptions, suggested the potential usefulness of a fresh approach to measuring the effect of campaign activity on voter behavior.

[Table 9-1 about here]

One feature of the campaign spending literature is that it attempts to draw conclusions about campaign effects using overall campaign spending as the independent variable. Overall campaign spending is an amalgamation of spending for particular purposes, and so insight into the effectiveness of campaign spending overall can be gained by learning the effect of particular campaign activities, such as voter mobilization efforts. This suggests the value of obtaining a reasonable dollar per vote estimate for the cost of inducing a supporter to vote. Indeed, as the campaign spending literature progressed, a parallel and independent literature on the effects of campaign mobilization on voter turnout was developing. What did these observational and experimental studies say about the effectiveness of voter mobilization efforts?

As previously mentioned, at the time of the 1998 New Haven study there was already a small field experimental literature on the effect of campaigns on voter turnout. Table 9-2 summarizes the field experiment literature prior to the 1998 New Haven experiment. By far, the largest previous study was Gosnell's (1927), which measured the effect of nonpartisan mail on

turnout in Chicago. In this pioneering research, eight thousand voters were divided by street into treatment (GOTV mailings) and control group. Three decades later, Eldersveld conducted a randomized intervention during a local charter reform vote to measure the effectiveness of alternative campaign tactics. He later analyzed the effect of a drive to mobilize apathetic voters in an Ann Arbor municipal election (Eldersveld and Dodge 1954; Eldersveld 1956). These experiments measured the turnout effects of a variety of different modes of communications. In the years following these studies, only a handful of scholars performed similar research. Miller, Bositis, and Baer (1981) examined the effects of a letter sent to residents of a precinct in Carbondale, Illinois prior to the 1980 general election. Adams and Smith (1980) conducted an investigation of the effect of a single thirty-second persuasion call on turnout and candidate choice in a special election for a Washington D.C. city council seat. In sum, prior to 1998, only a few field experiments on mobilization – spread across a range of political contexts and over many decades – had been conducted. Nevertheless, these studies formed a literature that might be taken to support several very tentative conclusions. First, the effects of voter contacts appeared to be extremely large. Treatment effects of twenty percentage points or more appear common in these papers. Thus, we might conclude that voters can be mobilized quite easily, and since mobilizing supporters is a key task, by implication even modest campaign resource disparities will play an important role in election results. Second, there is no evidence that the effect of contacts decreased over time – the effectiveness of mailings in the 1980s was as great as what had been found in earlier decades.

[Table 9-2 about here]

In addition to these early field experiments, another important line of work on campaign effects used laboratory experiments to investigate how political communications affect voter

turnout. A leading example is Ansolabehere and Iyengar (1996), which finds that exposure to negative campaign advertisements embedded in mock news broadcasts reduced subjects' reported intention to vote, particularly among independent voters. As with field experiments, these laboratory studies use random assignment to measure the causal effects of the treatments. However, integrating the results of these important and innovative laboratory studies into estimates of mobilization effects is challenging. Although the internal validity of such studies is impressive, the magnitude of the laboratory effects may not provide a clear indication of the magnitude of treatment effects in naturalistic contexts. More generally, though it is often remarked that a laboratory experiment will reliably indicate the *direction* though not the magnitude of the effect that would be observed in a natural setting, to my knowledge this has not been demonstrated and it is not obviously correct in general or in specific cases.^{ix} Further, despite the efforts of the researchers to simulate a typical living room for conducting the experiment, the natural environment differs from the laboratory environment in many obvious and possibly important ways, including the subject's awareness of being monitored.

In contrast to the relatively sparse experimental evidence, there is a large amount of observational research on campaigns and voter turnout. As of 2000, the most influential work on turnout were survey-based analyses of the causes of participation. Rosenstone and Hansen's book (1993) is a good example of the state of the art circa 1998 (see also Verba, Schlozman, and Brady 1995). This careful study is an excellent resource that is consulted and cited (according to Google Scholar, as of June 30, 2010 over 1500 times) by nearly everyone who writes about turnout, and the style of analysis employed is still common in current research. Rosenstone and Hansen use the American National Election Studies (ANES) to measure the effect of campaign contacts (among other things) on various measures of participation. They assess the contribution

of many different causes of participation in presidential and midterm years (see Tables 5.1 and 5.2 in Rosenstone and Hansen 1993) using estimates from a pooled cross-sectional analysis of ANES survey data. The estimated effect of campaign contact on reported voter turnout is approximately a ten percentage point boost in turnout probability.

This sizable turnout effect from campaign contact is of similar magnitude to many of the effects measured in the field experiments from the 1920s through 1980s. The sample size used in the Rosenstone and Hansen study is impressive, giving the estimation results the appearance of great precision. However, there are several methodological and substantive reasons why the findings might be viewed as unreliable. First, the results from the survey-based voter mobilization research appear to be in tension with at least some of the aggregate campaign spending results. If voters can be easily mobilized by a party contact, then it is difficult to understand why a campaign would have to spend so much to gain a single vote (see Table 9-1). Rather, modest amounts of spending should yield large returns. This tension could perhaps be resolved if there are large differences between average and marginal returns to mobilization expenditures or if campaign spending is highly inefficient. Nevertheless, taking the survey evidence as well as the early field experiments on voter mobilization seriously, if a campaign contact in a presidential year boosts turnout by ten percentage points and a large share of partisans in the ANES report not being contacted, then it is hard to simultaneously believe both the mobilization estimates and also the findings (summarized in Table 9-1) suggesting that campaigns must spend many hundreds of dollars per vote.^x More importantly, the survey work on turnout effects is vulnerable to a number of methodological criticisms. The key problem in the survey-based observational work is the possibility that those who report campaign contact are different from those who do not report contact in ways that are not adequately captured by the

available control variables. The Gerber and Green studies were in many ways an attempt to address this and other possible weaknesses of the earlier work.

3. How do experiments address the problems in the prior research?

In this section, I present a framework for analyzing empirical results and apply the framework to describe how field experiments eliminate some of the sources of bias in observational studies. For concreteness, I will use the Rosenstone and Hansen study as a running example. In Rosenstone and Hansen's participation study, some respondents are contacted by campaigns and others are not. In the language of experiments, some subjects are "treated" (contacted) and others are "untreated" (not contacted). The key challenge in estimating the causal treatment effect is that the analyst must somehow use the available data to construct an estimate of a counterfactual: what outcome would have been observed for the treated subjects had they not been treated? The idea that for each subject there is a potential outcome in the treated and the untreated state is expressed using the notational system termed the "Rubin Causal Model" (RCM) after Rubin (1978, 1990).^{xi} To focus on the main ideas, I initially ignore covariates. For each individual i let Y_{i0} be the outcome if i does not receive the treatment (in this example, contact by the mobilization effort), and Y_{i1} be the outcome if i receives the treatment. The treatment effect for individual i is defined as:

$$(1) \quad \tau_i = Y_{i1} - Y_{i0}.$$

The treatment effect for individual i is the difference between the outcomes for i in two possible though mutually exclusive states of the world, one in which i receives the treatment, and another in which i does not. Moving from a single individual to the average for a set of individuals, the average treatment effect for the treated (ATT) is defined as:

$$(2) \quad ATT = E(\tau_i | T_i=1) = E(Y_{i1} | T_i=1) - E(Y_{i0} | T_i=1),$$

where E stands for a group average and $T_i=1$ when a person is treated. In words, $Y_{i1}|T_i=1$ is the post-treatment outcome among those who are actually treated, and $Y_{i0}|T_i=1$ is the outcome for i that would have been observed if those who are treated had not been treated. Equation 2 suggests why it is difficult to estimate a causal effect. Because each individual is either treated or not, for each individual we observe either Y_1 or Y_0 . However, to calculate (2) requires *both* of these quantities for each treated individual. In a dataset the values of Y_1 are observed for those who are treated, but the causal effect of the treatment cannot be measured without an estimate of what the average Y would have been for these individuals had they not been treated. Experimental and observational research designs employ different strategies for producing this counterfactual. Observational data analysis forms a comparison group using those who remain untreated. This approach generates selection bias in the event that the outcomes in the untreated state for those who are untreated are different from the outcomes in the untreated state among those who are treated. In other words, selection bias occurs if the differences between those who are treated and those who are not extend beyond exposure to the treatment. Stated formally, the observational comparison of the treated and the untreated estimates:

$$(3) E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=0) = [E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=1)] + [E(Y_{i0}|T_i=1) - E(Y_{i0}|T_i=0)] = ATT + Selection\ Bias.$$

A comparison of the average outcomes for the treated and the untreated equals the average treatment effect plus a selection bias term. The selection bias is due to the difference in the outcomes in the untreated state for those treated and those untreated. This selection bias problem is a critical issue addressed by experimental methods. Random assignment forms groups without reference to either observed or unobserved attributes of the subjects and consequently creates groups of individuals that are similar prior to application of the treatment. When groups

are formed through random assignment, the group randomly labeled the control group has the same expected outcome in the untreated state as the set of subjects designated at random to receive the treatment. Due to the independence of the group assignment and the potential outcomes, the randomly assigned control group can be used to measure what the outcome would have been for the treatment group, had the treatment group remained untreated.

The critical assumption for observational work is that, controlling for covariates (whether through regression or through matching), $E(Y_{i0}|T_i=1) = E(Y_{i0}|T_i=0)$ – i.e., apart from their exposure to the treatment, the treated and untreated group outcomes are on average the same in the untreated state. Subject to sampling variability, this will be true by design for groups formed at random. In contrast, observational research uses the observables to adjust the observed outcomes and thereby produce a proxy for the treated subject's potential outcomes in the untreated state. If this effort is successful, then there is no selection bias. Unfortunately, without a clear rationale based on detailed knowledge of why some observations are selected for treatment and others are not, this assumption is rarely convincing. Consider the case of estimating the effect of campaign contact on voter turnout. First, there are likely to be important omitted variables correlated with campaign contact that are not explained by the included variables. Campaigns are strategic and commonly use voter files to plan which households to contact. A key variable in many campaign targeting plans is the household's history of participation, and households that do not vote tend to be ignored. The set of control variables available in the ANES data, or other survey datasets, does not commonly include vote history or other variables that might be available to the campaign for its strategic planning. Second, past turnout is highly correlated with current turnout. Therefore, $E(Y_{i0}|T_i=1)$ may be substantially higher than $E(Y_{i0}|T_i=0)$. Moreover, while it may be possible to make a reasonable guess at the

direction of selection bias, analysts rarely have a clear notion of the magnitude of selection bias in particular applications and so it is uncertain how estimates may be corrected.^{xii}

In addition to selection bias, field experiments address a number of other common methodological difficulties in observational work, many of these concerns related to measurement. In field experiments, the analyst controls the treatment assignment and so there is no error in measuring who is targeted for treatment. Although observational studies could, in principle, also measure the treatment assignment accurately, in practice analysis is frequently based on survey data, which relies on self reports. Again, consider the case of the voter mobilization work. Contact is self-reported (and, for the most part, so is the outcome, voter turnout). When there is misreporting, the collection of those individuals who report receiving the treatment are in fact a mixture of treated and untreated individuals. By placing untreated individuals in the treated group and treated individuals in the untreated, random misclassification will tend to attenuate the estimated treatment effects. In the extreme case, where the survey report of contact is unrelated to actual treatment status or individual characteristics, the difference in outcomes for those reporting treatment and those not reporting treatment will vanish. In contrast, systematic measurement error could lead to exaggeration of treatment effects. In the case of survey-based voter mobilization research, there is empirical support for concern that misreporting of treatment status leads to overestimation of treatment effects. Research has demonstrated both large amounts of misreporting and also a positive correlation between misreporting having been contacted and misreporting having voted (Vavreck 2007; Gerber and Doherty 2009).

There are some further difficulties with survey-based observational research that are addressed by field experiments. In addition to the uncertainty regarding who was assigned the

treatment, it is sometimes unclear what the treatment was, as survey measures are sometimes not sufficiently precise. For example, the ANES item used for campaign contact in the Rosenstone and Hansen study asks respondents: “Did anyone from one of the political parties call you up or come around and talk to you about the campaign?” This question ignores nonpartisan contact, conflates very different modes of communication, grouping together face-to-face canvassing, volunteer calls, and commercial calls (while omitting important activities such as campaign mailings), and does not measure the frequency or timing of contact.

In addition to the biases discussed thus far, another potential source of difference between the observational and experimental estimates is that those who are treated outside of the experimental context may not be the same people who are treated in an experiment. If those who are more likely to be treated in the real world (perhaps because they are likely to be targeted by political campaigns) have especially large (or small) treatment effects, then an experiment which studies a random sample of registered voters will underestimate (or overestimate) the ATT of what may often be the true population of interest – those individuals who are most likely to be treated in typical campaigns. A partial corrective for this is weighting the result to form population proportions similar to the treated population in natural settings, though this would fail to account for differences in treatment effects between those who are actually treated in real world settings and those who “look” like them but are not treated.

Finally, although this discussion has focused on the advantages of randomized experiments over observational studies, in estimating campaign effects field experimentation has some advantages over conventional laboratory experimentation. Briefly, field experiments of campaign communications typically study the population of registered voters (rather than a student population or other volunteers), measure behavior in the natural context (versus a

university laboratory or a “simulated” natural environment; also, subjects are typically unaware of the field experiment), and typically estimate the effect of treatments on the actual turnout (rather than on a surrogate measure such as stated vote intention or political interest).

4. The Development and Diffusion of Field experiments in Political Science

The details of the 1998 New Haven study are reported in Gerber and Green (2000). Since the 1998 New Haven study, which assessed the mobilization effects of nonpartisan canvassing, phone calls, and mailings, over 100 field experiments have measured the effects of political communications on voter turnout. The number of such studies is growing quickly (more than linearly) and dozens of researchers have conducted voter mobilization experiments. Some studies essentially replicate the New Haven study and consider the effect of face-to-face canvassing, phone, or mail in new political contexts, including other countries (e.g., Guan and Green 2006; Gerber and Yamada 2008; John and Brannan 2008). Other work looks at new modes of communication or variations on the simple programs used in New Haven, including analysis of the effect of phone calls or contacts by communicators matched to the ethnicity of the household (e.g., Michelson 2003), repeat phone calls (Michelson, Garcia Bedolla, and McConnell 2009), television and radio (Panagopoulos and Green 2008; Gerber et al. 2009) and new technologies, such as email and text messaging (Dale and Strauss 2007). Field experiments have also measured the effect of novel approaches to mobilization, such as Election Day parties at the polling place (Addonizio, Green, and Glaser 2007).

The results of these studies are compiled in a quadrennial review of the literature, *Get out the Vote!*, the latest version of which was published in 2008 (Green and Gerber 2008). A detailed review of the literature over the past ten years is also contained in Nickerson and Michelson’s chapter in this volume. A meta-analysis of the results of dozens of canvassing, mail, and phone

studies shows that the results from the initial New Haven study have held up fairly well.

Canvassing has a much larger effect than does the less personal modes of communication, such as phone and mail. The marginal effect of brief commercial calls, such as those studied in New Haven, and nonpartisan mailings appear to be less than one percentage point, while canvassing boosts turnout by about seven percentage points in a typical electoral context.^{xiii}

In recent years, field experimentation has moved well beyond the measurement of voter mobilization strategies and has now been applied to a broad array of questions. Although the first papers were almost entirely by American politics specialists, comparative politics and international relations scholars are now producing some of the most exciting work. Moreover, the breadth of topics in American politics that researchers have addressed using field experiments has grown immensely. A sense of the range of applications can be gained by considering the topics addressed in a sampling of recent studies using field experiments:

Effect of partisanship on political attitudes: Gerber, Huber, and Washington (2010) study the effect of mailings informing unaffiliated, registered voters of the need to affiliate with a party to participate in the upcoming closed primary. They find that the mailings increase formal party affiliation and, using a post-treatment survey, they find a shift in partisan identification, as well as a shift in political attitudes.

Influence of the media on politics: Gerber, Karlan, and Bergan (2009) randomly provide *Washington Post* and *Washington Times* newspaper subscriptions to respondents prior to a gubernatorial election and they examine the effect of media slant on voting behavior. They find that the newspapers increase voter participation and also shift voter preference toward the Democratic candidate.

Effect of interpersonal influence: Nickerson (2008) analyzes the effect of a canvassing effort on members of the household that are not directly contacted by the canvasser. He finds that spouses and roommates of those who are contacted are also more likely to vote following the canvassing treatment.

Effect of mass media campaigns: Gerber et al. (2009) analyze the effect of a multi-million dollar partisan television advertising campaign. Using tracking polls to measure voter preferences each day, they find a strong but short-lived boost in the sponsor's vote share.

Effect of candidate name recognition: Panagopoulos and Green (2008) measure the effect of radio ads that boost name recognition in low salience elections. They find that ads that provide equal time to both the incumbent's name and challenger's name have the effect of boosting the (relatively unknown) challenger's vote performance.

Effect of partisan political campaigns: Wantchekon (2003) compares broad policy versus narrow clientelistic campaign messages in a 2001 Benin election. Gerber (2004) reports the results of a 1999 partisan campaign.

Effect of political institutions and policy outcomes and legitimacy: Olken (2010) compares the performance of alternative institutions for the selection of a public good in Indonesia. He finds that, although more participatory institutions do not change the set of projects approved, participants are more satisfied with the decision-making process.

Effect of Election Day institutions on election administration: Hyde (2010) studies the effect of election monitors on vote fraud levels.

Effect of lobbying on legislative behavior: Bergan (2009) examines the effect of a lobbying effort on a bill in the New Hampshire legislature. An email from an interest group causes a statistically significant increase in roll call voting for the sponsor's measure.

Effect of constituency opinion on legislator behavior: Butler and Nickerson (2009) examine the effect of constituency opinion on legislative voting. They find that mailing legislators polling information about an upcoming legislative measure results in changes in the pattern of roll call support for the measure.

Effect of voter knowledge on legislative behavior: Humphreys and Weinstein (2007) examine the effect of legislative performance report cards on representatives' attendance records in Uganda. They find that showing legislators' attendance records to constituents results in higher rates of parliamentary attendance.

Effect of social pressure on political participation: Gerber, Green, and Larimer (2008) investigate the effect of alternative mailings, which exert varying degrees of social pressure. They find that a pre-election mailing listing the recipient's own voting record and a mailing listing the voting record of the recipient and their neighbors caused a dramatic increase in turnout.

Media and interethnic tension/prejudice reduction: Paluck and Green (2009) conduct a field experiment in post-genocide Rwanda. They randomly assign some communities to a condition where they are provided with a radio program designed to encourage people to be less deferential to authorities. The findings demonstrate that listening to the program makes listeners more willing to express dissent.

Mickelson and Nickerson's chapter and Wantchekon's chapter in this volume provide many further examples. In addition to addressing important substantive questions, field experiments can make methodological contributions, such as assessing the performance of standard observational estimation methods. In this line of research data from experimental studies are reanalyzed using observational techniques. The performance of the observational estimation method is evaluated by comparing the estimation results from an application of the observational method with the unbiased experimental estimates. Arceneaux, Gerber, and Green (2006) conduct one such comparison by assessing the performance of regression and matching estimators in measuring the effects of experimental voter mobilization phone calls. The study compares the experimental estimates of the effect of a phone call (based on a comparison of treatment and control group) and the estimates that would have been obtained had the experimental dataset been analyzed using observational techniques (based on a comparison of those whom the researchers were able to successfully contact by phone and those not contacted). They find that exact matching and regression analysis overestimate the effectiveness of phone calls, producing treatment effect estimates several times larger than the experimental estimates.

Reviewing all of these contributions, both substantive and methodological, a collection that is only a part of the vast body of recent work, shows the depth and range of research in political science using field experimentation. The earliest studies have now been replicated many times, while new studies are branching into exciting and surprising areas. I doubt that ten years ago anyone could have predicted the creativity of recent studies and the range of experimental manipulations. From essentially zero studies just over a decade ago, field experimentation is now a huge enterprise. I draw several conclusions about recent developments. First, voter mobilization is still studied, but the research focus has shifted from simply measuring the

effectiveness of campaign communications to broader theoretical issues such as social influence, norm compliance, collective action, and interpersonal influence. Second, there has been a move from studying only political behavior to the study of political institutions as well. Third, field experimentation has spread from initial application in American politics to comparative politics and international relations. Fourth, field experiments are now used to study both common real-world phenomena (such as campaign television commercials or the effect of election monitors), as well as novel interventions for which there are no observational counterparts (unusual mailings or legislative report cards in developing countries). For these novel interventions, of course, no observational study is possible.

5. Frequently Asked Questions

Field experiments are not a panacea and there are often substantial challenges in the implementation, analysis, and interpretation of findings. For an informative recent discussion of some of the limitations of field experiments, see Humphreys and Weinstein (2009), and especially Deaton (2009); for a reply to Deaton, see Imbens (2009). Rather than compile and evaluate a comprehensive list of potential concerns and limitations, in this section I provide a somewhat informal account of how I address some of the questions I am frequently asked about field experiments. The issue of the external validity of field experiments is left for the concluding section.

Some field experiments have high levels of noncompliance due to the inability to treat all of those assigned to the treatment group (low contact rates). Other methods, such as lab experiments, seem to have perfect compliance. Does this mean field experiments are biased?

Given that one-sided noncompliance (i.e., the control group remains untreated, but some of those assigned to the treatment group are not treated) is by far the most common situation in

political science field experiments, the answer will address this case. If the researcher is willing to make some important technical assumptions (see Angrist, Imbens, and Rubin 1996 for a formal statement of the result) when there is failure to treat in a random experiment, a consistent (large sample unbiased) estimate of the average treatment effect on treated can be estimated by differencing the mean outcome for those assigned to the treatment and control group and dividing this difference by the proportion of the treatment group that is actually treated.

The consequences of failure to treat are illustrated in Figure 9-1, which depicts the population analogues for the quantities that are produced by an experiment with noncompliance.^{xiv} Figure 9-1 provides some important intuitions about the properties and limitations of the treatment effect estimate when some portion of the treatment group is not treated. The figure depicts a population where there are three types of people (a person's type is not directly observable to the experimenter), each with different values of $Y_i(0)$ and $Y_i(1)$, where $Y_i(X)$ is the potential outcome for a subject of type i when treated ($X=1$) or untreated ($X=0$). Individuals are arrayed by group, with the X-axis marking the population proportion of each group and the Y-axis indicating outcome levels. Panel A depicts the population when assigned to the treatment group, and Panel B shows the population when assigned to control group. Alternatively, the figure can be thought of as depicting the potential outcomes for a large sample from a population, with some subjects randomly assigned to the treatment group and others to the control group. In this case, the independence of treatment group assignment and potential outcomes ensures that for a large sample the proportions of each type of person are the same for the treatment and the control group, as are the $Y_i(X)$ levels.

[Figure 9-1 about here]

Panel A shows the case where two of the three types of people are actually treated when assigned to the treatment group and one type is not successfully treated when assigned to the treatment group (in this example, Type 3 people are “noncompliers”). The height of each of the three columns represents the average outcome for each of the three groups and their widths represent the proportion of the population in that group. Consider a simple comparison of the average outcome when an individual is assigned to the treatment group versus the control group, a.k.a. the intent-to-treat effect (ITT). The geometric analogue to this estimate is to calculate the difference in the total area of the shaded rectangles for the treatment and the control group. Visually it is clear that the difference between the total area in Panel A and Panel B is the area created by the change in Y in Panel A due to the application of the treatment to groups 1 and 2 (the striped rectangles). Algebraically, the difference between the treatment group average and the control group average, the ITT, is equal to $[Y_1(1)-Y_1(0)] p_1 + [Y_2(1) - Y_2(0)] p_2$. Dividing this quantity by the share of the treatment group actually treated, $(p_1 + p_2)$, produces the average of the treatment effect among those actually treated (ATT).^{xv} This is also called the complier average causal effect (CACE), highlighting the fact that the difference between the average outcomes when the group is assigned to the treatment versus the control condition is produced by the changing treatment status and subsequent difference in outcomes for the subset of the population that are compliers.

As Figure 9-1 suggests, one consequence of failure to treat all of those assigned to the treatment group is that the average treatment effect is estimated for the treated, not the entire subject population. The average treatment effect for the entire population, the ATE, equals $[Y_1(1)-Y_1(0)] p_1 + [Y_2(1) - Y_2(0)] p_2 + [Y_3(1) - Y_3(0)] p_3$. As the final term in the ATE expression is not observed, an implication of noncompliance is that the researcher is only able to

directly estimate treatment effects for the subset of the population that one is able to treat. The implications of measuring the ATT rather than the ATE depend on the research objectives and whether treatment effects vary across individuals. Sometimes the treatment effect among those who are treated is what the researcher is interested in, in which case failure to treat is a feature of the experiment, not a bug. For example, if a campaign is interested in the returns from a particular type of canvassing sweep through a neighborhood, the campaign wishes to know the response of the people whom the effort will likely reach, not the hypothetical responses of people who do not open the door to canvassers or who have moved away.

If treatment effects are homogeneous, the ATT and the ATE (the average treatment effect for the population, compliers as well as noncompliers) are the same, regardless of the contact rate. Demonstrating that those whom are treated in an experiment have pre-treatment observables that differ from the overall population mean is not sufficient to show that the ATT is different from the average population treatment effect, as what matters is the treatment effect (see equation 1) not the covariates or the level of $Y_i(0)$. Figure 9-1 could be adjusted (by making the size of the gap between $Y_i(0)$ and $Y_i(1)$ equal for all groups) so that all groups have different $Y_i(0)$ but the same values of $Y_i(1) - Y_i(0)$. Further, higher contact rates may be helpful at reducing any gap between ATT and ATE. As Figure 9-1 illustrates, if the Type 3 (untreated) share of the population approaches zero (the column narrows), the treatment effect for this type would have to be very different from the other subjects in order to produce enough “area” for this to lead to a large difference between the ATE and ATT. Although raising the share of the treatment group that is successfully treated typically reduces the difference between ATE and ATT, in a pathological case if the marginal treated individual has a more atypical treatment effect than the average of those “easily” treated, then the gap between ATT and ATE may grow as the

proportion treated increases. The ATE and ATT gap can be investigated empirically by observing treatment effects under light and heavy efforts to treat. This approach parallels the strategy of investigating the effects of survey nonresponse by using extra effort to interview and seeing if there are differences in the lower and higher response rate samples (Pew 1998).

Although the issue of partial treatment of the target population is very conspicuous in many field experiments, it is a common problem in laboratory experiments as well. Designs, such as typical laboratory experiments, which put off randomization until compliance is assured, will achieve a 100 percent treatment rate, but this does not “solve” the problem of measuring the treatment effect for a population (ATE) versus those who are treated (ATT). The estimand for a laboratory experiment is the ATE for the particular group of people who show up for the experiment. Unless this is also the ATE for the broader target population as well, failure to treat has entered at the subject recruitment stage.

A final note: nothing in this answer should be taken as asserting that a low contact rate does not matter. Non-compliance affects the precision of the experimental estimates. Intuitively, when there is nearly 100% failure to treat, it would be odd if meaningful experimental estimates of the CACE could be produced, since the amount of noise produced by random differences in Y due to sampling variability in the treatment and control groups would presumably swamp any of the difference between to the treatment and control groups that was generated by the treatment effect. Indeed, a low contact rate will lead to larger standard errors, and may leave the experimenter unable to produce useful estimates of the treatment effect for the compliers.

Do field experiments all assume homogeneous treatment effects?

No. See Figure 9-1, which depicts a population in which the compliers are divided into two sub-populations with different treatment effects. The ITT and the ATT both estimate the *average* treatment effects, which may vary across individuals.

Are field experiments ethical?

All activities, including research, raise ethical questions. It is surprising to read that certain physics experiments currently being conducted are understood by theoreticians to have a measurable (though very small) probability of condensing the planet Earth into a sphere 100 meters in diameter (Posner 2004). I am not aware of any field experiments in political science that pose a remotely similar level of threat. A full treatment of the subject of research ethics is well beyond the scope of a brief response and also not my area of expertise, but I will make several points that I feel are sometimes neglected.

First, advocates of randomized trials in medicine turn the standard ethical questions around and argue that those who treat patients in the absence of well-controlled studies should reflect on the ethics of using unproven methods and not performing the experiments necessary to determine whether the interventions they employ actually work. They argue that many established practices and policies are often merely society-wide experiments (and, as such, poorly designed experiments which lack a control group but somehow sidestep ethical scrutiny and bureaucratic review). They recount the tragedies that have followed when practices were adopted without the support of experimental evidence (Chalmers 2003). Taking this a step further, recent work has begun to quantify the lives lost due to delays imposed by Institutional Review Boards (IRB) (Whitney and Schneider 2010).

Second, questions are occasionally raised as to whether an experimental intervention might change a social outcome, such as affecting an election outcome by increasing turnout.

Setting aside the issue of whether changing an election outcome through increased participation or a more informed electorate (the most common mechanism for this hypothetical event, given current political science field experiments) is problematic or praiseworthy, in the highly unlikely event that an experiment did alter an election result, this would only occur for the small subset of elections where the outcome would have been tied or nearly tied in the absence of the experiment. In this case, there are countless other mundane and essentially arbitrary contributions to the outcome with electoral consequences that are orders of magnitude larger than the typical experimental intervention. A partial list includes: ballot order (Miller and Krosnick 1998), place of voting (Berger, Meredith, and Wheeler 2008), the number of polling places (Brady and McNulty 2004), use of optical scan versus punch card ballots (Ansolabehere and Stewart 2005), droughts, floods, or recent shark attacks (Achen and Bartels 2004), rain on election day (Knack 1994), and a win by the local football team on the weekend prior to the election (Healy, Malhotra, and Mo 2009). That numerous trivial or even ridiculous factors might swing an election seems at first galling, but note that these factors only matter when the electorate is very evenly divided. In this special case, however, regardless of the election outcome, an approximately equal number of citizens will be pleased and disappointed with the result. As long as there is no regular bias in which side gets the benefit of chance, there may be little reason for concern. Perhaps this is why we do not bankrupt the treasury to make sure our elections are entirely error free.

Field experiments do not control for background activity. Does this cause bias?

Background activity affects the interpretation of the experimental results but does not cause bias. Because treatment is randomly assigned, background conditions affect $Y_i(0)$ and $Y_i(1)$ similarly in both the treatment and control group. Treatment effects can be estimated in the

usual fashion. That is not to say that background conditions do not matter, as they may affect $Y(0)$ and $Y(1)$, and therefore the treatment effect $Y(1) - Y(0)$. If the treatment effect varies with background conditions, then background factors affect the generalizability of the results; the treatment effect that is estimated should be thought of as conditional on the background conditions.

Are field experiments too expensive?

Field experiments tend to be expensive but there are ways to reduce the cost, sometimes dramatically. Many recent field experiments were performed in cooperation with organizations that are interested in evaluating a program or communications effort. Fortunately, a growing proportion of foundations are requiring (and paying for) rigorous evaluation of the programs they support, which should provide a steady flow of projects looking for partners to assist in experimental evaluations.

What about treatment “spillover” effects?

Spillover effects occur when those who are treated in turn alter their behavior in a way that affects other subjects.^{xvi} Spillover is a potentially serious issue in field experiments and the importance of this concern will vary by case. It is also fair to note that spillover is typically *not* a problem in the controlled environment of laboratory experiments, as contact among subjects can be observed and regulated. In most cases, the presence of spillover effects in field experiments attenuate estimated treatment effects by causing the control group to be partially treated. If the researcher is concerned about mitigating the danger from spillover effects, reducing the density of treatment is one option, as this will likely reduce the share of the control group affected by spillover. Another perspective is to consider spillover effects as worth measuring in their own right; some experiments have been designed to measure spillover (Nickerson 2008). It is

sometimes forgotten that spillover is also an issue in observational research. In survey-based observational studies of party contact and candidate choice, for example, only those who report direct party contact are coded as contacted. If those who are contacted affect those who are not contacted, this will introduce bias into the observational treatment effect estimates for party contact, which are based on comparison of those coded treated and those coded untreated.

6. Further Issues

In this section, I sound some notes of caution regarding the development of field experiments in political science. I first discuss the question of how to interpret the results of field experiments. Field experiments to date have often focused on producing accurate measurement rather than illuminating broader theoretical issues. However, in the absence of some theoretical context, it may be difficult to judge what exactly is being measured. Second, I discuss some difficulties with the development of literatures based on field experiments. These issues relate to the difficulty of replication and the potential sources of bias in experimental reporting, especially when projects are undertaken with nonacademic partners.

Unbiased estimates... of what?

Field experimentation is a measurement technique. Many researchers who use field experiments are content to report treatment effects from an intervention and leave it at that, an empiricism that has led some observers to dismiss field experiments as mere program evaluations. This line of attack fails to appreciate the enormous importance of obtaining convincing causal estimates. Throughout the history of science, new measurement technologies (e.g., the microscope, spectography) and reliable causal estimates (controlled experiments) have been the crucial impetus to productive theorizing. There are also practical costs to ignoring solid empirical demonstrations because of concerns about theoretical mechanisms. To take one

example from the history of medicine, the major attack against the prescient findings of Semmelweis, who conducted a pioneering experiment in the 1860s demonstrating that washing hands in a disinfectant significantly reduced death from post-partum infection, was that his theory about how the intervention worked was flawed and incomplete (Loudon 2000). This justified critique of Semmelweis' theoretical arguments was taken as a license by the medical community to ignore his accurate empirical conclusions, resulting in countless unnecessary deaths over the next several decades.

Without gainsaying the value of measurement, what is lost if there is no clearly articulated theoretical context? There are implications for both external and internal validity. First, consider external validity. There is no theoretical basis for the external validity of field experiments comparable to the statistical basis for claims of internal validity and it is often very plausible that treatment effects might vary across contexts. The degree of uncertainty assigned when applying a treatment effect produced in one context (place, people, time, treatment details) to another context is typically based on reasonable conjecture. The appeals to reasonableness are, in the absence of evidence or clear theoretical guidance, disturbingly similar to the assumptions on which observational approaches often rest.

To be concrete, consider the case of canvassing to mobilize voters where the treatment effect is the effect of the intervention on the subject's turnout. In the most rudimentary framework, the size of this effect might depend on how the intervention affects his or her beliefs about the costs and benefits of voting in the upcoming election. Beliefs about the costs and benefits of participation may depend in turn on, among other things, how the intervention affects subject knowledge about or the salience of the upcoming election, expectations about the closeness of the election, beliefs about the importance of the election, and the perceived social

desirability of voting. The intervention's effect on these variables might depend on the political context, such as the political history or political norms of the place in which the experiment occurs. Additional factors affecting the size of the treatment effect might include which subset of the population is successfully treated and how near the subjects are to the threshold of participation. The treatment effect estimated by a given experiment might conceivably be a function of variables related to any and all of these considerations.^{xvii}

Understanding the mechanism by which the treatment is working may be critical for accurate predictions about how the treatment will perform outside of the initial experimental context. Consider the challenge of extrapolating the effectiveness of face-to-face canvassing. Alternative theories have very different implications. One way this intervention may increase participation is if contact by a canvasser increases the subject's perception of the importance of the election (changing the subject's beliefs about the benefits of participation). However, a subject's beliefs may be less affected by canvassing in a place where canvassing is routine than in a place where it occurs only under the most extreme political conditions. Turning to the long-term effectiveness of canvassing, if canvassing works by causing subjects to update their perceptions of the importance of voting, the link between canvassing and turnout effects may not be stable. If, following an intensive canvassing effort, the election turns out to be a landslide or the ballot has no important contests, a voter might ignore subsequent canvassing appeals as uninformative. In a similar vein, interventions may fail upon repetition if they work in part due to their novelty. Alternatively, if the effect of canvassing works through social reciprocity, where the canvasser exerts effort and the subject exchanges a pledge of reciprocal effort to vote, then the voter's experience at the polls may not alter the effectiveness of the intervention; the estimate of canvassing effects in today's election might apply well to subsequent interventions. What

matters is the voter's perception that the canvasser has exerted effort – perhaps canvassing in a snowstorm would be especially effective. This discussion of the effect of canvassing suggests the value of delineating and adjudicating among the various possible mechanisms. More generally, being more explicit about the theoretical basis for the observed result might inspire some caution and provide guidance when generalizing findings.

Reflecting on the theoretical context can also assist in establishing the internal validity of the experiment. For example, it might be useful to reflect on how the strategic incentives of political actors can alter treatment effects. Continuing with the example of a canvassing experiment, suppose some local organization is active in a place where a canvassing experiment is (independently) being conducted. Consider how the canvassing intervention might affect the behavior of such an independent group that expends a fixed amount of effort making calls to people and asking them if they intend to vote. Suppose that the group operates according to the rule: If the voter says he or she will vote, there is no further attempt to encourage them, while if the voter says no, the group expends substantial time and effort to encourage the subject to vote. If the canvassing treatment took place prior to the independent group's efforts and the canvassing was effective, this will result in a share of their limited mobilization resources being diverted from the treatment group to the control group (who are less likely to say they plan to vote), depressing the estimated treatment effect. Less subtle, if an experimental canvassing effort is observed, this might alter the behavior of other campaigns. More common violations of the requirement that treatment group assignment not affect potential outcomes may occur if a treated subject communicates directly with other subjects. The importance of these effects may vary with context and treatment. For example, if the treatment is highly novel, it is much more likely that subjects will remark upon the treatment to housemates or friends.

Finally, careful consideration of the complete set of behavioral changes that might follow an intervention may also suggest new outcome measures. Theorizing about how the intervention alters the incentives and capabilities of subjects may affect which outcome measures are monitored. It is common to measure the effect of a voter mobilization intervention on voter turnout. However, it is unclear how and whether political participation in elections is related to other forms of political involvement. If citizens feel that they have fulfilled their civic responsibility by voting, then voting may be a substitute for attending a Parent-Teacher Association (PTA) meeting or contributing to the Red Cross. Alternately, the anticipation of voting may lead to enhanced confidence in political competence and a stronger civic identity, which may then inspire other forms of political and community involvement or information acquisition.

Publication process: Publication bias, Proprietary research, Replication

One of the virtues of observational research is that it is based on public data. The ANES data are well known and relatively transparent. People would notice if, for some reason, the ANES data were not released. In contrast, experimental data are produced by the effort of scholars, unsupervised, who then must decide whether to write up results and present the findings to the scholarly community. These are very different situations and it is unclear what factors affect which results are shared when sharing depends on the choices of researchers and journal editors. The process by which experimental results are disclosed or not affects how much one's priors should move as a result of an experimental report. Under ideal circumstances, updating is a mundane matter of adjusting priors using the new reported effect sizes and standard errors. However, the uncharted path from execution of the experiment to publication adds an additional source of uncertainty, as both the direction and magnitude of any bias incorporated

through this process are unknown.^{xviii} Although any given experiment is unbiased, the experimental literature may nevertheless be biased if the literature is not a representative sample of studies. This issue is especially vexing in the case of proprietary research. A significant amount of experimental research on campaign effects is now being conducted by private organizations such as campaigns, unions, or interest groups. This type of work has the potential to be of immense benefit, as the results are of interest, the studies numerous and conducted in varying contexts, and the cost of the research is borne by the sponsoring organization. This benefit might be offset, however, if only a biased subset of experiments are deemed fit for public release.

It is often suggested that the scholarly publication process may be biased in favor of publicizing arresting results. However, anomalous reports have only a limited effect when there is a substantial body of theory and frequent replication. The case of the “discovery” of cold fusion illustrates how theory and replication work to correct error. When it was announced that nuclear fusion could be achieved at relatively low temperatures using equipment not far beyond that found in a well-equipped high school lab, some thought that this technology would be the solution to the world’s energy problems. Physicists were quite skeptical about this claim from the outset, because theoretical models suggested it was not very plausible. Well-established models of how atoms interact under pressure imply that the distance between the atoms in the cold fusion experiments would be billions of times greater than what is necessary to cause the fusion effects claimed.

Physics theory also made another contribution to the study of cold fusion. The famous cold fusion experiment result was that the experimental cell produced heat in excess of the heat input to the system. Extra heat was, in fact, the bottom line measurement of greatest practical

importance, as it suggested that cold fusion could be an energy source. Theoretical work on fusion, however, pointed to a number of other outputs that could be used to determine if fusion was really occurring. These included gamma rays, neutrons, and tritium. These fusion byproducts are easier to measure than is excess heat, which requires a careful accounting of all heat inputs and outputs to accurately calculate the net change. It was the absence of these byproduct measurements (in both the original and replication studies), in addition to the theoretical implausibility of the claim, that led many physicists and chemists to doubt the experimental success from the outset. Replication studies began within twenty-four hours of the announcement. In a relatively short period of time, the cold fusion claim was demolished.^{xix}

Compare this experience to how a similar drama would unfold in political science field experimentation. The correctives of strong theory and frequent replication are not available in the case of field experimental findings. Unfortunately, field experiments tend to be expensive and time consuming. There would likely be no theory with precise predictions to cast doubt on the experimental result or provide easy to measure byproducts of the experimental intervention to lend credence to the experimental claims. The lack of theoretically induced priors is a problem for all research, but it is especially significant when replication is not easy. How long would it take political science to refute “cold fusion” results? If the answer is “until a series of new field experiments refute the initial finding,” then it might take many years.

7. Conclusion

After a generation in which non-experimental survey research dominated the study of political behavior, we may now be entering the age of experimentation. There was not a single field experiment published in a major political science journal in the 1990s, while in the past decade scholars have published dozens of such papers. The widespread adoption of field

experimentation is a striking development. The results accumulated to date have had a substantial effect on what we know about politics and have altered both the methods used to study key topics and the questions that are being asked. Further, field experimentation in political science has moved well beyond the initial studies of voter mobilization to consider the effects of campaign communications on candidate choice, the political effects of television, newspapers, and radio, the effects of deliberation, tests of social psychology theories, the effects of political institutions, and measurement of social diffusion. The full impact of the increased use of field experiments may be difficult to measure. It is impossible to know for sure, but some of the recent attention to causal identification in observational research in political science may have been encouraged by the implicit contrast between the opaque and often implausible identification assumptions used in observational research and the more straightforward identification enjoyed by randomized interventions.

Although the past decade has seen many exciting findings and innovations, there are important areas for growth and improvement. Perhaps most critically, field experimentation has not provoked the healthy back and forth between theory and empirical findings that is typical in the natural sciences. Ideally, experiments would generate robust empirical findings and then theorists would attempt to apply or produce (initially) simple models that predict the experimental results and, critically, make new experimentally testable predictions. For the most part, the field experiment literature in political science has advanced by producing measurements in new domains of inquiry rather than by addressing theoretical puzzles raised by initial results. This may be in part due to the relatively light theorizing in most political science field experiments to date. This is a missed opportunity for intellectual progress. However, the responsibility for this may be shared, as the relatively slight role of theory in the evolution of the

political science field experiment literature so far may in part reflect a lack of engagement by our more theoretically inclined colleagues.

References

- Abramowitz, Alan I. 1988. "Explaining Senate Election Outcomes." *American Political Science Review* 82: 385-403.
- Achen, Christopher H., and Larry M. Bartels. 2004. "Blind Retrospection: Electoral Responses to Droughts, Flu, and Shark Attacks." Estudio/Working Paper 2004/199.
- Adams, William C., and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44: 389-95.
- Addonizio, Elizabeth, Donald Green, and James M. Glaser. 2007. "Putting the Party Back into Politics: An Experiment Testing Whether Election Day Festivals Increase Voter Turnout." *PS: Political Science & Politics* 40: 721-27.
- Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review* 80: 313-36.
- Angrist, Joshua D., Guido Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-72.
- Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106: 979-1014.
- Ansola-behere, Stephen D., and Alan S. Gerber. 1994. "The Mismeasure of Campaign Spending: Evidence from the 1990 U.S. House Elections." *Journal of Politics* 56: 1106-18.
- Ansola-behere, Stephen D., and Shanto Iyengar. 1996. *Going Negative: How Political Advertising Divides and Shrinks the American Electorate*. New York: The Free Press.
- Ansola-behere, Stephen D., and James M. Snyder. 1996. "Money, Elections, and Candidate Quality." MIT, Typescript.
- Ansola-behere, Stephen D., and Charles Stewart III. 2005. "Residual Votes Attributable to Technology." *Journal of Politics* 67: 365-89.
- Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 1-36.

- Arceneaux, Kevin, and David Nickerson. 2009. "Who is Mobilized to Vote? A Re-Analysis of Eleven Randomized Field Experiments." *American Journal of Political Science* 53: 1-16.
- Bergan, Daniel E. 2009. "Does Grassroots Lobbying Work?: A Field Experiment Measuring the Effects of an e-Mail Lobbying Campaign on Legislative Behavior." *American Politics Research* 37: 327-52.
- Berger, Jonah, Marc Meredith, and S. Christian Wheeler. 2008. "Contextual Priming: Where People Vote Affects How They Vote." *Proceedings of the National Academy of Sciences* 105: 8846-49.
- Brady, Henry E., and John E. McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Presented at the annual meeting of the Society for Political Methodology. Palo Alto, CA.
- Butler, Daniel M., and David W. Nickerson. 2009. "Are Legislators Responsive to Public Opinion? Results from a Field Experiment." Unpublished paper, Yale University.
- Chalmers, Iain. 2003. "Trying to do more Good than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-Date Evaluations." *The ANNALS of the American Academy of Political and Social Science* 589: 22-40.
- Chin, Michelle L., Jon R. Bond, and Nehemia Geva. 2000. "A Foot in the Door: An Experimental Study of PAC and Constituency Effects on Access." *Journal of Politics* 62: 534-49.
- Dale, Allison, and Aaron Strauss. 2007. "Mobilizing the Mobiles: How Text Messaging Can Boost Youth Voter Turnout." Working Paper, University of Michigan.
- Davenport, Tiffany C., Alan S. Gerber, and Donald P. Green. 2010. "Field Experiments and the Study of Political Behavior." In *The Oxford Handbook of American Elections and Political Behavior*, ed. Jan E. Leighley. New York: Oxford University Press.
- Deaton, Angus S. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." NBER Working Paper No. 14690.
- Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50: 154-65.
- Eldersveld, Samuel J., and Dodge, Richard W. 1954. "Personal Contact or Mail Propaganda? An Experiment in Voting Turnout and Attitude Change." In *Public Opinion and Propaganda*, ed. Daniel Katz. New York: Holt, Rinehart and Winston.
- Erikson, Robert S., and Thomas R. Palfrey. 2000. "Equilibria in Campaign Spending Games: Theory and Data." *American Political Science Review* 94: 595-609.
- Gerber, Alan S. 1998. "Estimating the Effect of Campaign Spending on Senate Election Outcomes Using Instrumental Variables." *American Political Science Review* 92: 401-11.

- Gerber, Alan S. 2004. "Does Campaign Spending Work?: Field Experiments Provide Evidence and Suggest New Theory." *American Behavioral Scientist* 47: 541-74.
- Gerber, Alan S. Forthcoming. "New Directions in the Study of Voter Mobilization: Combining Psychology and Field Experimentation."
- Gerber, Alan S., and David Doherty. 2009. "Can Campaign Effects Be Accurately Measured Using Surveys?: Evidence From a Field Experiment." Yale University, Typescript.
- Gerber, Alan S., David Doherty, and Conor M. Dowling. 2009. "Developing a Checklist for Reporting the Design and Results of Social Science Experiments." Typescript, Yale University.
- Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2009. "The Size and Duration of Campaign Television Advertising Effects: Results from a Large-Scale Randomized Experiment." Working Paper, ISPS Yale University.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653-63.
- Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In *Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press.
- Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*, eds. Ian Shapiro, Rogers Smith, and Tarek Massoud. New York: Cambridge University Press.
- Gerber, Alan S., and Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.
- Gerber, Alan S., Donald P. Green, and David W. Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9: 385-392.
- Gerber, Alan S., Gregory A. Huber, and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." Working Paper, ISPS Yale University.
- Gerber, Alan S., Dean Karlan, and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1: 35-52.
- Gerber, Alan S., and Kyohei Yamada. 2008. "Field Experiment, Politics, and Culture: Testing Social Psychological Theories Regarding Social Norms Using a Field Experiment in Japan." Working Paper, ISPS Yale University.

- Green, Donald P., and Alan S. Gerber. 2004 *Get Out The Vote: How to Increase Voter Turnout*. Washington, DC: Brookings Institution Press.
- Green, Donald P., and Alan S. Gerber. 2008. *Get Out The Vote: How to Increase Voter Turnout* (Second Edition). Washington, DC: Brookings Institution Press.
- Green Donald P., and Jonathan S. Krasno. 1988. "Salvation for the Spendthrift Incumbent: Reestimating the Effects of Campaign Spending in House Elections." *American Journal of Political Science* 32: 884–907.
- Gosnell, Harold F. 1927. *Getting-out-the-vote: An Experiment in the Stimulation of Voting*. Chicago: University of Chicago Press.
- Guan, Mei, and Donald P. Green. 2006. "Non-Coercive Mobilization in State-Controlled Elections: An Experimental Study in Beijing." *Comparative Political Studies* 39: 1175-93.
- Habyarimana, James, Macartan Humphreys, Dan Posner, and Jeremy Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision? An Experimental Approach." *American Political Science Review* 101: 709-25.
- Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *The Journal of Economic Literature* 42: 1009-55.
- Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2009. "Do Irrelevant Events Affect Voters' Decisions? Implications for Retrospective Voting." Stanford Graduate School of Business Working Paper No. 2034.
- Humphreys, Macartan, and Jeremy M. Weinstein. 2007. "Policing Politicians: Citizen Empowerment and Political Accountability in Africa." Presented at the annual meeting of the American Political Science Association, Chicago, IL.
- Humphreys, Macartan, and Jeremy Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12: 367-78.
- Hyde, Susan D. 2010. "Experimenting in Democracy Promotion: International Observers and the 2004 Presidential Elections in Indonesia." *Perspectives on Politics*, Forthcoming.
- Imbens, Guido W. 2009. "Better Late than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." National Bureau of Economic Research Working Paper No. 14896.
- Jacobson, Gary C. 1978. "The Effects of Campaign Spending in Congressional Elections." *American Political Science Review* 72: 469-91.
- Jacobson, Gary C. 1985. "Money and Votes Reconsidered: Congressional Elections, 1972-1982." *Public Choice* 47: 7-62.

- Jacobson, Gary C. 1990. "The Effects of Campaign Spending in House Elections: New Evidence for Old Arguments." *American Journal of Political Science* 34: 334-62.
- Jacobson, Gary C. 1998. *The Politics of Congressional Elections*. New York: Longman.
- John, Peter, and Tessa Brannan. 2008. "How Different Are Telephoning and Canvassing? Results from a 'Get Out the Vote' Field Experiment in the British 2005 General Election." *British Journal of Political Science* 38: 565-74.
- Knack, Steve. 1994. "Does Rain Help the Republicans? Theory and Evidence on Turnout and the Vote." *Public Choice* 79: 187-209.
- LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76: 604-20.
- Levitt, Steven D. 1994. "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House." *Journal of Political Economy* 102: 777-98.
- Loudon, Irvine. 2000. *The Tragedy of Childbed Fever*. Oxford: Oxford University Press.
- Medical Research Council, Streptomycin in Tuberculosis Trials Committee. 1948. "Streptomycin Treatment for Pulmonary Tuberculosis." *British Medical Journal* 2: 769-82.
- Michelson, Melissa R. 2003. "Getting Out the Latino Vote: How Door-to-door Canvassing Influences Voter Turnout in Rural Central California." *Political Behavior* 25: 247-63.
- Michelson, Melissa R., Lisa García Bedolla, and Margaret A. McConnell. 2009. "Heeding the Call: The Effect of Targeted Two-Round Phonebanks on Voter Turnout." *Journal of Politics* 71: 1549-63.
- Miller, Roy E., David A. Bositis, and Denise L. Baer. 1981. "Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman." *International Political Science Review* 2: 445-60.
- Miller, Joanne M., and Jon A. Krosnick. 1998. "The Impact of Candidate Name Order on Election Outcomes." *Public Opinion Quarterly* 62: 291-330.
- Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49-57.
- Olken, Benjamin. 2010. "Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia." *American Political Science Review* 104: 243-67.
- Paluck, Elizabeth Levy, and Donald P. Green. 2009. "Deference, Dissent, and Dispute Resolution: An Experimental Intervention using Mass Media to Change Norms and Behavior in Rwanda." *American Political Science Review* 103: 622-44.

- Panagopoulos, Costas, and Donald P. Green. 2008. "Field Experiments Testing the Impact of Radio Advertisements on Electoral Competition." *American Journal of Political Science* 52: 156-68.
- Pew Research Center for the People & the Press. 1998. "Possible Consequences of Non-Response for Pre-Election Surveys: Race and Reluctant Respondents." Survey Report, May 16. Retrieved from <http://people-press.org/report/89/possible-consequences-of-non-response-for-pre-election-surveys>.
- Posner, Richard A. 2004. *Catastrophe: Risk and Response*. Oxford: Oxford University Press.
- Rosenstone, Steven J., and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: MacMillan.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6: 34-58.
- Rubin, Donald B. 1990. "Formal Modes of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25: 279-92.
- Taubes, Gary. 1993. *Bad Science: The Short Life and Weird Times of Cold Fusion*. New York: Random House.
- Vavreck, Lynn. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 287-305.
- Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.
- Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior. Evidence from a Field Experiment in Benin." *World Politics* 55: 399-422.
- Whitney, Simon N., and Carl E. Schneider. 2010. "A Method to Estimate the Cost in Lives of Ethics Board Review of Biomedical Research." Presented at the 'Is Medical Ethics Really in the Best Interest of the Patient?' conference, Uppsala, Sweden.

Table 9-1. Approximate Cost of Adding One Vote to Candidate Vote Margin

	Incumbent	Challenger
Jacobson (1985)	\$250/vote	\$16/vote
Green and Krasno (1988)	\$20/vote	\$17/vote
Levitt (1994)	\$488/vote	\$146/vote
Erikson and Palfrey (2000)	\$61/vote	\$32/vote

NOTE: 2008 dollars. Calculations are based on 190,000 votes cast in a typical House district. For House elections, this implies that a 1% boost in the incumbent's share of the vote increases the incumbent's vote margin by 3,800 votes. Adapted from Gerber (2004).

ⁱ This review draws on previous literature reviews I have authored or coauthored, including Gerber and Green (2008), Davenport, Gerber, and Green (2010), Gerber (forthcoming), and Gerber (2004). The author thanks Jamie Druckman, John Bullock, David Doherty, Conor Dowling, and Eric Oliver for helpful comments.

ⁱⁱ The discussion in this section draws on Gerber and Green (2008).

ⁱⁱⁱ For a comparison of medical research and social science research, see Gerber, Doherty, and Dowling (2009).

^{iv} Gosnell assembled a collection of matched pairs of streets and selected one of the pair to get the treatment, but it is not entirely clear that Gosnell used random assignment to decide which was to be treated. Given this ambiguity, it might be more appropriate to use some term other than experiment to describe the Gosnell studies; perhaps “controlled intervention.”

^v There were some exceptions, e.g., Ansolabehere and Gerber (1994).

^{vi} A closely related approach was taken by Erikson and Palfrey (2000), who use a theoretical model to deduce conditions under which candidate spending levels could be treated as exogenous.

^{vii} If a campaign activity causes a supporter who would otherwise have stayed home on Election Day to vote, this changes the vote margin by one vote. If a campaign activity causes a voter to switch candidates, this would change the vote margin by two votes. For further details about these calculations, see Gerber (2004).

^{viii} The OLS estimate relies on the questionable assumption that spending levels are uncorrelated with omitted variables. The instrumental variables approach relies on untestable assumptions about the validity of the instruments. Levitt's study uses a small nonrandom subset of election contests, and so there is a risk that these elections are atypical. Further, restricting the sample to repeat elections may reduce, but not fully eliminate, the biases due to omitted variables because changes in spending levels between the initial election and the rematch (which is held at least two, and sometimes more years later) may be correlated with unobservable changes in variables correlated with vote share changes.

^{ix} For instance, the natural environment will provide the subject more behavioral latitude, which might reverse the lab findings. For example, exposure to negative campaigning might create an aversion to political engagement in a lab context, but negative information may pique curiosity about the advertising claims which, outside the lab, could lead to increased information search and gossiping about politics, and in turn *greater* interest in the campaign.

^x Compounding the confusion, the results presented in the early field experimental literature may overstate the mobilization effects. The pattern of results in those studies suggests the possibility that the effect sizes are exaggerated due to publication-related biases. There is a strong negative relationship between estimates and sample size, a pattern consistent with inflated reports due to file drawer problems or publication based on achieving conventional levels of statistical significance (Gerber, Green, and Nickerson 2001).

^{xi} This section draws heavily on and extends the discussion in Gerber and Green (2008).

^{xii} Note that this uncertainty is not contained in the reported standard errors and, unlike sampling variability, it remains undiminished as the sample size increases (Gerber, Green, and Kaplan 2004). The conventional measures of coefficient uncertainty in observational research thereby underestimate the true level of uncertainty, especially in cases where the sample size is large.

^{xiii} For details, see appendix A, B, and C of *Get Out The Vote: How to Increase Voter Turnout* (Green and Gerber 2008).

^{xiv} Figure 9-1 and subsequent discussion incorporates several important assumptions. Writing the potential outcomes as a function of the individuals own treatment assignment and compliance rather than the treatment assignment and compliance of all subjects employs the stable unit treatment value assumption (SUTVA). Depicting the potential outcomes as independent of treatment group assignment given the actual treatment or not of the subjects employs the exclusion restriction. See Angrist, Imbens, and Rubin (1996).

^{xv} This estimand is also known as the complier average causal effect, or CACE, since it is the treatment effect for the subset of the population that are "compliers." Compliers are subjects who are treated when assigned to the treatment group and remain untreated when assigned to the control group.

^{xvi} See Sinclair's chapter in this volume for further discussion.

^{xvii} The most striking difference across contexts demonstrated to date is that mobilization effects appear strongest for those voters predicted to be about 50 percent likely to vote, a result that follows theoretically from a latent variable model (Arceneaux and Nickerson 2009).

^{xviii} For a related point, where the target is observational research, see Gerber, Green, and Kaplan (2004).

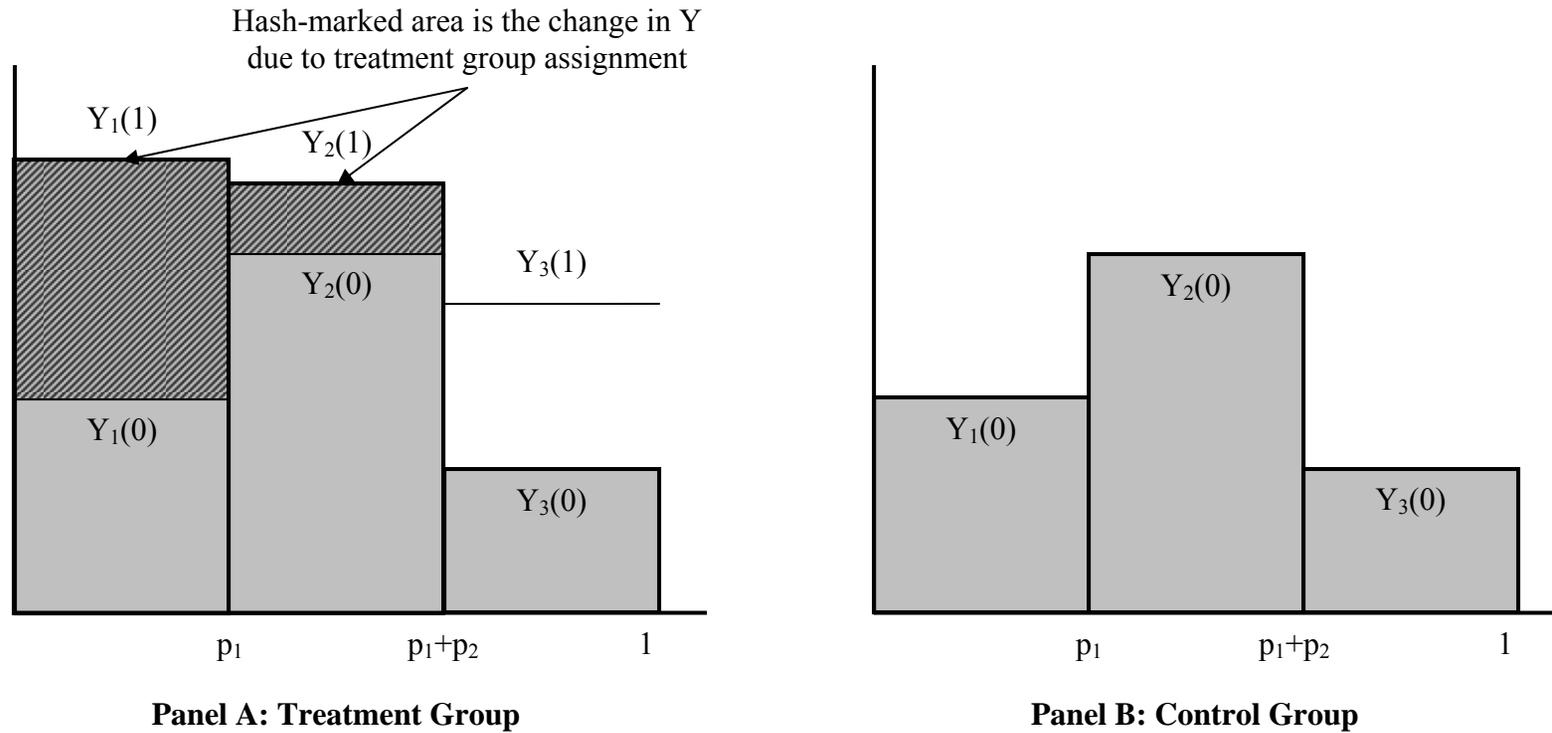
^{xix} This account draws on Gary Taubes' *Bad Science* (1993).

Table 9-2. Voter Mobilization Experiments Prior to 1998 New Haven Experiment

Study	Date	Election	Place	N of subjects (including control group)	Treatment	Effects on Turnout *
Gosnell (1927)	1924	Presidential	Chicago	3,969 registered voters	Mail	+1%
Gosnell (1927)	1925	Mayoral	Chicago	3,676 registered voters	Mail	+9%
Eldersveld (1956)	1953	Municipal	Ann Arbor	41 registered voters	Canvass	+42%
				43 registered voters	Mail	+26%
Eldersveld (1956)	1954	Municipal	Ann Arbor	276 registered voters	Canvass	+20%
				268 registered voters	Mail	+4%
				220 registered voters	Phone	+18%
Miller et al. (1981)	1980	Primary	Carbondale, IL	79 registered voters	Canvass	+21%
				80 registered voters	Mail	+19%
				81 registered voters	Phone	+15%
Adams and Smith (1980)	1979	Special city council	Washington, DC	2,650 registered voters	Phone	+9%

* These are the effects reported in the tables of these research reports. They have not been adjusted for contact rates. In Eldersveld's 1953 experiment, subjects were those who opposed or had no opinion about charter reform. In 1954, subjects were those who had voted in national but not local elections. Note that this table includes only studies that use random experimental design (or near-random, in the case of Gosnell [1927]). Adapted from Gerber, Green, and Nickerson (2001).

Figure 9-1. Graphical Representation of Treatment Effects with Noncompliance



Note: $\bar{Y}_T = p_1 Y_1(1) + p_2 Y_2(1) + (1-p_1-p_2) Y_3(0)$
 $\bar{Y}_C = p_1 Y_1(0) + p_2 Y_2(0) + (1-p_1-p_2) Y_3(0)$

$Y_i(X)$ = Potential outcome for type i when treated status is X ($X=0$ is untreated, $X=1$ is treated).
 The Y-axis measures the outcome, the X-axis measures the proportion of the subjects of each type.

III. Decision Making

10. Attitude Change Experiments in Political Science

Allyson L. Holbrook

The importance of attitudes and the processes by which they are formed and changed is ubiquitous throughout political science. Perhaps the most obvious example is research exploring citizens' attitudes towards candidates, how these attitudes are influenced by political advertising and other persuasive messages, and how these attitudes influence decisions and behavior (see McGraw's chapter in this volume). Attitudes toward candidates are fundamental to the democratic process because they help voters make vote choices, perhaps the most basic way in which citizens can express their opinions and influence government (e.g., Rosenstone and Hansen 1993). Other key attitudes in the political domain include attitudes toward specific policies which also help voters make important decisions about voting, vote choice, and activism (e.g., Rosenstone and Hansen 1993). Attitudes toward institutions such as political parties and government entities also influence people's view of government. Finally, attitudes toward other groups in society (e.g., African-Americans or women) may help determine support for specific policies (e.g., Transue 2007). Thus, attitudes play a central role in many of the democratic processes studied by political scientists reviewed in this volume (e.g., Gadarian and Lau; Hutchings; Lodge and Tabor; McGraw; Nelson; Wilson and Eckel).

Defining Attitudes

Different definitions of attitudes have been proposed by psychologists (e.g., Thurstone 1931; Allport 1935; Bem 1970). Perhaps the most widely accepted modern definition was proposed by Eagly and Chaiken (1993, p. 1), who defined an attitude as: "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or

disfavor.” One key feature of attitudes is that an attitude is directed toward a specific attitude object. This attitude object can be almost anything: a person, place, idea, inanimate object, experience, behavior, or any other object. A second feature of attitudes is that they are evaluative and reflect the extent of positivity or negativity a person has toward the attitude object.¹ Most attitude change research in political science assesses change in explicit attitudes, but recent interest in psychology has also focused on implicit attitudes (see Lodge and Taber’s chapter in this volume).

Early definitions of attitudes also defined them as being stable over time (Cantril 1934) and influencing behavior and thought (Allport 1935). However, more recent definitions conceive of attitudes as not only having a valence, but also strength. Strong attitudes are stable, resist change, and/or influence behavior and thought and weak attitudes do not (for a review, see Krosnick and Petty 1995). Very weak attitudes that are based on little information and which people may construct on the spot when asked to report their attitudes are similar to what some have labeled as non-attitudes (e.g., Converse 1964, 1970). Recent evidence suggests, however, that even the weakest attitudes may not truly be non-attitudes, but logical constructions based on whatever information people have (Krosnick et al. 2002).

Attitude Change

The key dependent variable of interest in this chapter is attitude change. For the purposes of this chapter, attitude changes includes processes of attitude formation (i.e., a change from having no attitude toward an attitude object to having an attitude toward the object) as well as change in an existing attitude (i.e., an existing attitude becoming more or less positive or negative). Generally, attitudes researchers in psychology have conceived of attitude formation as a special case of attitude change and have argued that the processes involved in the former are

often similar to those involved in the latter (Eagly and Chaiken 1993). Attitude change is sometimes directly measured and at other times indirectly inferred. For example, one could measure a group of people's attitudes toward Barack Obama, have them watch several of his recent speeches, and then measure attitudes again. Alternatively, one could compare the attitudes of people who have been exposed to different information. If attitudes differed as a function of the information to which people were exposed, one can infer that differential attitude change or persuasion occurred.

1. Measuring Attitudes

Attitudes are an inherently subjective construct and there is no current measure of attitudes that is without some error. Researchers in political science have typically used three types of measures to assess attitudes in both observational and experimental studies. One of the most commonly used measure involves asking people self-report questions inquiring whether they like or dislike (or favor or oppose) a political candidate, policy, or other attitude object. In most cases, these questions not only capture the direction of the attitude (e.g., like or dislike), but also some measure of extremity (e.g., like a great deal, like somewhat, or like a little). These measures of attitudes are perhaps the most direct, but they rely on the assumption that respondents are willing and able to report their attitudes, which may not always be true. Many such attitude reports may be subject to social desirability response bias, whereby respondents are motivated to report attitudes that are more socially desirable and avoid reporting those that might make others look at them less favorably (e.g., attitudes toward African-Americans or towards legalizing same-sex marriage; Warner 1965; Himmelfarb and Lickteig 1982). In addition, direct attitude questions may be affected by response biases such as extreme response style and acquiescence response bias (e.g., Baumgartner and Steenkamp 2001).

A second set of measures asks respondents about their preferences regarding, for example, political candidates or policies. For example, to assess policy preferences regarding immigration, respondents might be asked: “Under current law, immigrants who come from other countries to the United States legally are entitled, from the very beginning, to government assistance such as Medicaid, food stamps, or welfare on the same basis as citizens. But some people say they should not be eligible until they have lived here for a year or more. Which do you think? Do you think that immigrants who are here legally should be eligible for such services as soon as they come, or should they not be eligible?” (Davis, Smith, and Marsden 2007). Similarly, a respondent might be asked to report which candidate he or she preferred in an upcoming election.

These questions assess attitudes indirectly through preferences or choices (in which a study participant is asked to choose among a list of options or rank order them). Preferences have been defined as “a comparative evaluation of (i.e., ranking over) a set of objects” (Druckman and Lupia 2000, 2). Attitudes are distinct from both preferences and choices, but attitudes are one influence on preferences. Preferences may allow researchers to assess attitudes in ways that may be less affected by social desirability bias than more direct measures (e.g., Henry and Sears 2002; although see Berinsky 2002) and perhaps less affected by some response effects than more direct attitude questions (e.g., acquiescence response bias). On the other hand, particular policy preference questions usually frame a policy decision in terms of two (or more) choices, forcing respondents to choose the response that *most closely* matches their preference. Thus, preference measures may not be as precise or sensitive as more direct measures of attitudes. A preference for one candidate over another does not indicate whether a person is positive or negative toward the preferred candidate. In addition, policy and candidate preferences may not only be influenced

by attitudes toward the various policies, but also potentially by strategic concerns and beliefs about the effectiveness of policies and the role of government. For example, a respondent may feel very positive about reducing global warming, but not support a government policy requiring businesses to reduce carbon emissions, either because he or she does not believe the policy would be effective at reducing future global warming or because he or she does not believe that it is the government's responsibility to pass and enforce such regulations.

A third type of attitudinal measure uses attitude-expressive behaviors (e.g., financially supporting a particular candidate or organization) as indicators of attitudes. These behaviors have been assessed through self-report behavioral intention questions, retrospective self-reports of past behaviors, and observing behaviors directly. Self-report measures of behavioral intentions assume that respondents can accurately predict their own behavior and are willing to do so, but such reports may be inaccurate, both because respondents may not be able to accurately predict their own behavior under some conditions (e.g., Wolosin, Sherman, and Cann 1975) and because reports of behaviors may be influenced by socially desirability response bias (e.g., Warner 1965). Retrospective reports of past behaviors assume that respondents' memory for past behaviors is accurate and that respondents are honestly reporting these behaviors, although this may not always be the case (e.g., Belli et al. 1999). Finally, direct measures of behaviors can be cumbersome and may be very limited. For example, willingness to act to express one's opinion could be measured by giving study participants the opportunity to sign either a pro-choice or pro-life petition. However, this measure is very specific and may be influenced by factors other than participants' attitudes.

Measures of behaviors only indirectly assess attitudes and may be influenced by other factors as well. There is a long history of research in psychology examining whether attitudes

predict behavior and the conditions under which they do and do not (Ajzen and Fishbein 2005; also, for a review, see Eagly and Chaiken 1993). There are many reasons why behaviors and attitudes may not be entirely consistent. For example, behaviors and behavioral intentions may be influenced by factors other than attitudes such as social norms and a person's perceived control of the behavior (Ajzen and Fishbein 2005) and the resources required to engage in the behavior (e.g., Miller and Krosnick 2004).

Behaviors and behavioral intentions may also be influenced by strategic concerns. For example, consider the case of the primary for the 2008 presidential election. In this election, John McCain acquired enough delegates to become the presumptive Republican nominee in early March of 2008, well before all the Republican primaries were complete. However, the choice of Barack Obama for the Democratic nominee was not determined until June of 2008. A person living in a state that allows its citizens to choose which party's primary in which to vote and that held its primary after John McCain had been determined to be the presumptive Republican nominee, but before the Democratic candidate had been determined, might strategically choose to vote in the Democratic primary and vote for the candidate she believed would be least likely to beat the Republican in the general election. In fact, during this primary, there was speculation in the media that Republicans may have voted for Hillary Clinton in Democratic primaries (in states where this was allowed) for just this reason (and to disrupt and draw out the process of deciding the Democratic nominee.ⁱⁱ Thus, measures of behaviors are indirect measures of attitudes that may contain error when used as measures of attitudes, although they can sometimes be directly observed in ways that attitudes cannot.

Experiments and Attitude Measurement

Experiments have been widely used in measuring attitudes and in improving attitude measurement. In particular, researchers have used survey experiments for these purposes (see Sniderman's chapter in this volume). For example, researchers using the list technique use a survey experiment to assess sensitive attitudes, beliefs, or behaviors (e.g., Kuklinski and Cobb 1998). This technique involves, for example, randomly assigning half of respondents a list of nonsensitive behaviors and asking how many they have done. The other half of respondents are randomly assigned to be given the same list of nonsensitive behaviors plus one sensitive behavior and asked to report how many they have done. The proportion of respondents who have done the sensitive behavior can be calculated as the difference in the means between the two groups without any respondents ever having to directly report that they performed the behavior.

Other researchers have used experiments to assess the effects of question wording or order on the quality of attitude measurement. For example, Krosnick et al. (2002) examined whether explicitly including or omitting a no opinion or don't know response option in telephone administered attitude survey questions improved or decreased data quality. Their findings show that including an explicit no opinion response option reduces data quality and gives respondents a cue to avoid going through the cognitive steps necessary to optimally answer survey questions and provide a practical guide to researchers about how best to ask attitude questions in surveys.

2. Observational research designs

Many researchers have used observational or non-experimental research designs to examine attitude change using a variety of approaches. One of the most basic approaches has been to look at the associations between one or more hypothesized causes of attitudes and attitudes themselves in cross-sectional data. For example, one might look at the association between the frequency with which respondents report having read a daily newspaper and

attitudes toward the war in Iraq. If a negative association is found, one might conclude that greater exposure to newspaper coverage led to less positive attitudes toward the war in Iraq.

The assumption underlying this kind of research design is that an association between attitudes and a hypothesized ingredient of such attitudes suggests that the latter influenced the former (for a review, see Kinder 1998) and that the size of the association (often assessed via multiple linear regression) reflects the size of the impact of each hypothesized ingredient on attitudes (see Rahn, Krosnick, and Breuning 1994). This type of approach has been used to provide support for a long list of factors that influence attitudes toward candidates, including party affiliations and policy positions (Campbell et al. 1960), prospective judgments of candidates' likely performance in office (Fiorina 1981), perceptions of candidates' personalities, emotional responses to candidates (Abelson et al. 1982), and retrospective assessments of the national economy (Kinder and Kiewiet 1979). The primary problem with this assumption is the old caveat that "correlation is not causation." Finding an association between two variables does not rule out the possibility that the hypothesized dependent variable (attitudes, in this case) is actually the cause of the hypothesized independent variable (newspaper reading) or that a third variable (e.g., perhaps political knowledge or interest) is the cause of both the independent and dependent variable of interest. Thus, the internal validity, or the confidence with which one can conclude from these studies that the independent variable caused the dependent variable, is very low. For example, although many researchers have tested predictors of candidate evaluations via this approach, analyses of longitudinal data suggests that people may form candidate evaluations or preferences first and that many of the "predictors" of these evaluations or preferences (including the reasons respondents list for voting for or against the candidate when asked

directly) are in fact rationalized from the evaluations rather than candidate evaluations being derived from these ingredients (e.g., Rahn et al. 1994).

A second, less common, observational design used as an alternative to experimentation is a repeated cross-sectional design. In this design, attitudes are measured before and after a naturally occurring event. For example, Krosnick, Holbrook, and Visser (2000) reported the results of just such a study to examine the effects of the Clinton administration's efforts in 1997 to bring attention to the issue of global warming. These efforts led to a debate in the media in which Clinton and other Democrats argued that global warming was happening, whereas Republican leaders argued that there was little or no evidence that global warming was happening and that any fluctuation in earth temperatures was due to natural fluctuations in climate, not to human actions. To study the effect of the initial efforts of the Clinton administration and the debate that followed, a nationally representative sample of respondents was surveyed about their attitudes and beliefs toward global warming before the media coverage and debate. A second nationally representative sample of adults was interviewed after the media coverage and debate had taken place. The effects of the media coverage and debate were assessed by comparing attitudes and beliefs measured before the debate to those measured after.

This kind of design more effectively isolates the causal influence of an event than a simple association between hypothesized cause and effect, but it is not without difficulties. The first is that any changes in attitudes in the samples over time are attributed to a specific event occurring between data collections. As a result, conclusions about causality are threatened by history, or the possibility that an event other than the one of interest to the researcher might have caused the attitude change. Furthermore, there are practical issues with this design. Conducting this kind of study is more expensive than a cross-sectional design because it involves multiple

data collections and one has to be aware that a naturally occurring event will occur in advance (so that attitudes can be measured before the event).

Another observational design is used when some people are exposed to a naturally occurring event and others are not but the experimenter does not control who is in each group (e.g., Huber and Arceneaux 2007). Inferences about causality from this design are threatened by self-selection because in these types of designs respondents are not randomly assigned to the two conditions. Therefore, respondents in the two groups may have differed before the event of interest. For example, Huber and Arceneaux (2007) studied the effect of campaign advertisements on attitudes toward political candidates by comparing the attitudes of residents in non-contested states who resided in an area near an adjoining highly contested state to those who resided in an area in an uncontested state that was not near an adjoining highly contested state. Thus, respondents were not randomly assigned to condition, but self selected into the two conditions by virtue of where they lived. In many cases, researchers argue that respondents in different conditions in these types of studies are comparable along many dimensions, but because random assignment is not used to assign respondents, one cannot be confident that the two groups are comparable on all important variables that might affect results.

A final common observational design involves longitudinal data collection in which the same people are interviewed or assessed at multiple points in time. This approach can be used to assess attitudes before and after a naturally occurring event as with the multiple cross-sectional data collection approach described above. The repeated measures approach has some advantages, specifically that one can examine individual-level correlates of attitude change between time 1 and time 2. However, it introduces an additional possible threat of testing to internal validity (see

Campbell and Stanley 1963). That is to say that being interviewed at time 1 could change attitudes measured at time 2.

A second difficulty with this design is the problem of attrition, whereby it may be difficult to re-interview all people interviewed at time 1 and time 2. This can be particularly problematic if the people interviewed only at time 1 differ from those interviewed at both times, what is sometimes called “selective nonresponse” (e.g., Miller and Wright 1995; see also Taris 2000). Selective nonresponse can threaten both the external validity of a study’s findings (e.g., if the final sample used in the analyses is not representative) and the internal validity (e.g., if attrition across panel waves is nonequivalent across groups).

An alternative approach to analyzing longitudinal data is to examine the effects of predictors measured at time 1 on attitudes measured at time 2 controlling for attitudes at time 1. Because causes occur temporally before their consequences, this procedure helps to establish causality and results in greater confidence in internal validity than any of the other observational designs reviewed thus far. This design faces quite a few practical concerns, including attrition and added costs. Although this design is one of the best observational designs for establishing causality, it is also one of the least commonly used because of the practical difficulties with data collection.ⁱⁱⁱ

The most common threat to these nonexperimental designs is that they tend to have low internal validity (for a review of threats to internal validity, see Campbell and Stanley 1963; McDermott’s chapter in this volume). Thus, it is difficult to conclude that the hypothesized independent variable *caused* the hypothesized dependent variable. Some nonexperimental designs can also be difficult and expensive to implement and both the internal and external

validity of their conclusions may be undermined by other potential threats such as attrition or history.

Some of these observational designs have the added problem of not directly assessing attitude change. Instead, these designs rely on the inference that an association between a hypothesized ingredient or cause of attitudes and attitudes themselves implies influence or change. Instead of measuring attitudes in these designs, one could instead ask respondents to report the direction and extent of any attitude change or to report their past attitudes about an issue and use this as a measure of attitude change. Research in psychology, however, suggests that respondents cannot do this accurately because they may not be able to accurately report their past attitudes (e.g., Markus 1986).

In studying attitude change, these observational designs have at least one other potential difficulty, which is that it is often difficult to know or measure the information to which people have been exposed that might influence their attitudes. People often cannot remember all the information about an attitude object to which they have been exposed and yet this information may influence their attitudes even if it is not recalled (Lodge, McGraw, and Stroh 1989). A researcher cannot, therefore, rely on self-reports to measure either exposure or the ingredients of respondents' attitudes. In longitudinal designs the researcher often has to infer that some event or stimulus occurring between measures of attitude change is responsible for observed changes. Often, this is done by looking at the nature of attitude change and inferring the information that must have led to that change. For example, Krosnick et al. (2000) found that strong Democrats became more convinced that global warming was happening over time and inferred that this was because strong Democrats were attending to and/or being persuaded by Democratic leaders. Similarly, they found that Republicans became less convinced that global warming was

happening and they inferred that this was happening because strong Republicans were attending to and/or being persuaded by messages from Republican leaders. This is a theoretically reasonable explanation for their findings, but message exposure and persuasion have to be inferred from the observed pattern of attitudes.

3. Experimental Designs

In order to overcome many of the difficulties with observational designs, political scientists have turned to experiments to test hypotheses about attitude change. There are two hallmarks of experimental designs. First, the experimenter manipulates the independent variable (e.g., what information is provided to respondents about a political candidate). Second, respondents or participants are randomly assigned to conditions (e.g., groups or levels of the independent variable). This allows the researcher to be confident that the only difference across groups or levels is the independent variable. If any changes or differences in the dependent variables are observed, the researcher can therefore be confident that these are caused by differences in the independent variable.

Most of the work on attitude change has relied on two basic experimental designs. First are versions of the pretest-posttest control group described by Campbell and Stanley (1963). In the purest version of this type of design, respondents are randomly assigned to two conditions (represented by the R in Figure 10-1 below). All respondents' attitudes are first measured (O_1 and O_3 in Figure 10-1) and then half of respondents who have been randomly assigned to the experimental condition receive a treatment (X in Figure 10-1). This treatment could be, for example, exposure to a persuasive message or a priming manipulation. Multiple experimental groups exposed to different treatments may also be included.

[Figure 10-1 about here]

The key comparison in this design is whether the difference between O_1 and O_2 in the experimental group is different from the difference between O_3 and O_4 in the control condition. This experimental design controls all threats to internal validity and allows the researcher to conclude with confidence that the experimental treatment caused any differences across conditions.

A second version of this design (shown below in Figure 10-2 and labeled here as the Pretest-Posttest Multiple Experimental Condition Design) uses multiple experimental conditions rather than a control group. In this design, respondents are randomly assigned to two (or more) experimental conditions with different treatments (X_a and X_b in Figure 10-2 below). Again, all respondents' attitudes are measured before and after the treatment (O_1 and O_3 in Figure 10-2 below).

[Figure 10-2 about here]

As with the Pretest-Posttest Control Group Design, the key comparison is whether the difference between O_1 and O_2 in Experimental Group A is different from the difference between O_3 and O_4 in Experimental Group B. If differences are observed across experimental groups, the researcher can confidently attribute these differences to the experimental manipulation.

The second type of experimental design used to study attitude change is what Campbell and Stanley (1963) call the Posttest-Only Control Group Design (see Figure 10-3 below). In this design, respondents are randomly assigned to either a control condition or one (or more) experimental conditions. One group of respondents is randomly assigned to receive a treatment (X in Figure 10-3 below). Then the attitudes of both experimental and control groups are measured.

[Figure 10-3 about here]

A second version of this type of design (shown below in Figure 10-4) uses multiple experimental groups, but no control condition. Respondents are randomly assigned to 2 (or more) experimental conditions in which they are exposed to different experimental treatments. Then respondents' attitudes are measured.

[Figure 10-4 about here]

Weaknesses of Experimental Designs

Although experimental designs provide much higher internal validity than observational designs, they are not without weaknesses. One potential weakness is that in many cases, experiments studying attitude change have used samples of undergraduate students. Although many laboratory experiments replicate when conducted with representative samples (e.g., Krosnick, Visser, and Holbrook 2002), there are also many important ways in which college undergraduates are different from a generally representative sample (e.g., they tend to be more homogenous in terms of socio-economic status, education, age, and often race and ethnicity). When studying attitude change, the homogeneity of age among college undergraduates may be of particular concern as susceptibility to persuasion is greatest during early and late adulthood (Visser and Krosnick 1998). Although the general processes involved in attitude change may be similar in college students and in general population samples and therefore it may be appropriate to use samples of students in research focusing on basic effects and mechanisms, researchers studying potential moderators of attitude change processes on which college students are relatively homogenous (e.g., education, age) should use more heterogeneous samples. The extent to which the use of college students as participants in attitude change research continues to be a topic for future research (see Druckman and Kam's chapter in this volume).

A second weakness of experiments studying attitude change is that they often use stimuli that are artificial and processes that do not accurately mirror the world with which actual people come into contact and are influenced by information about political candidates and issues. For example, in order to avoid the influence of pre-existing attitudes, knowledge, and beliefs about political candidates, researchers will sometimes have study participants form attitudes toward hypothetical candidates or policies (e.g., Lodge et al. 1989; although see e.g., Kaid and Boydston 1987). This allows researchers to very accurately isolate the effects of the information being presented, but this process may be different from the actual process by which people acquire information about candidates.

Experiments also typically vary from real-world attitude formation and change contexts in that information is acquired over a much shorter period of time. An experiment might last an hour in its entirety or in some cases be extended over several days or weeks, but the acquisition of information about policies and candidates often occurs over a period of months or even years. As such, attitude change studied in the laboratory may not reflect the processes that actually occur during campaigns or in response to media coverage or advertisements. Perhaps more importantly, attitude change studied during the brief time period of a laboratory experiment may not reflect the kind of change that persists over time or would be likely to influence behavior, although these consequences of attitudes are ones that are often of great interest to researchers. Very few laboratory studies assess whether attitude change persists over any length of time or has any influence on later behavior (although see Boninger et al. 1990; Druckman and Nelson 2003; Mutz and Reeves 2005).

Experiments examining attitude change in political science also often do not fully incorporate two processes that have a great deal of influence on persuasion processes: selective

exposure and selective elaboration. In facing the “buzzing, blooming confusion” (James 1890) of information that people are faced with every day, people make decisions about what information to be exposed to (e.g., they choose whether to watch a particular TV news show, or read information about a topic on a news website). Despite this, many attitude change experiments do not take into account or allow for selective exposure, although researchers have examined this process separately (e.g., Huang and Price 2001).

Furthermore, people choose which information to elaborate about and which to not think about carefully. Persuasion processes can occur through both more and less thoughtful processes and whether persuasion happens via high or low elaboration may have important consequences for the longevity and effects of the persuasion (e.g., Petty and Cacioppo 1986). Although elaboration has been shown to be a key process in persuasion processes, relatively few experiments in political science have measured or assessed elaboration (although see Nelson and Garst 2005).

Thus, the primary weaknesses of experimental designs, particularly as they have been used to study attitude change in political science, relate to external validity, specifically ecological validity (see McDermott’s chapter in this volume). “External validity refers to the question of whether an effect (and its underlying processes) that has been demonstrated in one research setting would be obtained in other settings, with different research participants and different research procedures” (Brewer 2000, 10). Ecological validity is a type or subcategory of external validity that deals with whether “the effect is representative of everyday life” (Brewer 2000, 12). Problems with the artificiality of the experimental context and processes and lack of sample representativeness may reduce both the external and ecological validity of laboratory experiments assessing attitude change.

Laboratory experiments may therefore explore attitude change and formation processes that do not reflect real world processes and these experiments may show researchers what can happen versus what does happen. For psychologists who are interested in the psychological mechanisms underlying processes like attitude change, this may not pose a great concern. Political scientists studying attitude change, however, typically want to apply their findings to processes of attitude formation or change that do occur in the real world, such as the processes by which campaigns influence voter evaluations of candidates or media coverage of an issue affects the public's evaluation of a particular proposed policy. As a result, the possible problems with both ecological and external validity in laboratory experiments may be of concern to political scientists.

4. Findings from Experimental Attitude Change Research

The literature using experiments to study attitude change in political science is extensive and much too large to be reviewed in detail in this chapter. This literature can be organized in two ways. Experimental attitude change research has had a major influence in a number of substantive areas of research in political science. Most of the research examining attitude change has focused on attitudes towards either political candidates or issues. The use of experimental procedures has been widespread throughout studies examining the processes by which people's attitudes toward political candidates are formed and changed (e.g., Nelson and Garst 2005; McGraw's chapter in this volume) including research on the effects of campaign advertising and media effects such as media priming and agenda setting (e.g., Miller and Krosnick 2000; Nelson's chapter in this volume). Experiments have also been widely used to study attitude change about political issues including processes like issue framing (e.g., Chong and Druckman 2007; Transue 2007; Brader, Valentino, and Suhay 2008; Gartner 2008; Nelson et al.'s chapter in

this volume). Experiments have been used less frequently to assess change in attitudes towards groups in society (e.g., Gaffié 1992; Glaser 2003), and attitudes toward government and other public institutions such as the Supreme Court (e.g., Iyengar, Peters, and Kinder 1982; Mondak 1990).

A second approach to organizing and describing the experimental literature on attitude change is to think about how experimentation has contributed to an understanding of attitude change processes. First, experiments have been used to assess and understand the ingredients of attitudes and what types of persuasive attempts do and do not lead to attitude change, including the content of persuasive messages, the source of these messages, and factors that influence resistance to persuasion (e.g., Andreoli and Worchel 1978; Bizer and Petty 2005). For example, Bizer and Petty (2005) found that simply reporting that one is “opposed” to a policy or candidate led to greater resistance to persuasion than saying that one is “supportive” of the policy or candidate. This occurred regardless of which position was framed as “opposition.” In one study, study participants who were asked to report how much they “opposed” their least liked candidate showed greater resistance to persuasion than those who were asked how much they “favored” their most liked candidate, regardless of which candidate they preferred.

A second area in which experimentation has provided key insights is in understanding potential moderators of attitude formation and change processes. Experiments have helped researchers understand when and for whom particular types of persuasive messages or information influence attitudes. This research has focused on the role of respondent characteristics such as race or gender (e.g., Kuklinski and Hurley 1994; Peffley and Hurwitz 2008), pre-existing attitudes, orientations or abilities such as political knowledge, sophistication,

or expertise (e.g., Druckman 2004; Gartner 2008), and personality or individual differences (e.g., need for cognition or need to evaluate; Druckman and Nelson 2003; Kam 2005).

For example, Gartner (2008) found that the effect of information about casualty predictions in the war in Iraq (i.e., half of respondents were randomly assigned to read a messages reporting that experts thought casualties would go down in the future and half were randomly assigned to read a message reporting that experts thought casualties would go up in the future) had a greater effect on attitudes toward the war among respondents who knew little about the war than among respondents who knew more about the war.

Finally, experimentation has provided a great deal of insight into the processes and mechanisms by which attitudes are formed and changed. This includes insights into at least three aspects of these processes. Researchers have studied processes that influence how people weigh individual beliefs or pieces of information in forming attitudes (e.g., Nelson, Clawson, and Oxley 1997; Miller and Krosnick 2000; Transue 2007). For example, Valentino, Hutchings, and White (2002) found that a respondents who read a persuasive message criticizing George W. Bush paired with subtle racial cues (e.g., pictures of Blacks to illustrate some of the arguments) used racial attitudes in assessing George W. Bush more than respondents who received the same message paired with neutral visual images (e.g., the Statue of Liberty). Other areas of research have also focused on the weights assigned to different pieces of information, for example research examining framing effects (see Nelson et al.'s chapter in this volume).

Researchers have also examined the processes by which people integrate information into overall attitudes (e.g., Lodge et al. 1989). For example, Lodge et al. (1989) demonstrated that participants primarily formed evaluations of candidates online (as information was received)

rather than via a memory-based process (whereby evaluations are formed at the time they are reported based on the information available about the candidate in memory).

Finally, researchers have examined mediators of the effects of persuasive messages such as emotions and cognitive processes. For example, Brader, Valentino and Suhay (2008) found that people were less supportive of immigration when an article about immigration was accompanied by a Latino cue and that this effect was mediated by anxiety about immigration. Thus, reading the article about the costs of immigration paired with a Latino salience cue led respondents to be the most anxious about immigration and, as a result, the most supportive of reducing immigration.

Of course, many experiments include elements of more than one of these approaches. For example, Miller and Krosnick (2000) found that people who see many news stories about a political issue weighed that issue more heavily when evaluating a presidential candidate than an issue about which they saw few news stories (commonly known as news media priming). The priming effect was strongest among knowledgeable study participants who trusted the media (two moderators of the effect) and the primary mechanism was the perceived importance of the issue. Thus media coverage influenced evaluations of candidates via issue importance and this effect was moderated by respondents' knowledge and trust in the media.

Similarly, McGraw, Hasecke and Conger (2003) not only examined the processes by which information is integrated into candidate evaluations (contrasting online and memory-based processes), they further examined when and for whom each of these processes is more or less likely to occur. For example, they found that attitudinal ambivalence and uncertainty are both associated with more memory-based (versus online) processing. These are just two examples of

many that simultaneously examine more than one aspect of attitude change, providing a more complex, in many cases more realistic, picture of attitude change processes.

5. Recommendations for Future Research

Although experimentation has contributed a great deal to our understanding of attitude change and persuasion in political science, future research could make even greater contributions by minimizing some of the weaknesses of the experimental research done to date. First, researchers need to work to increase external validity, particularly ecological validity. This means that future studies should be designed to assess what does happen rather than what could happen by using stimulus materials and other experimental procedures that mirror the attitude formation and change processes that occur in the real world (as some researchers have already begun to do; e.g., Chong and Druckman 2007; Gartner 2008). Furthermore, researchers need to assess these processes with representative samples to increase the generalizability of their findings. Finally, researchers can increase the external validity of this research, not necessarily by making experiments more realistic, but by using experiments along with other types of designs (e.g., observational designs with higher external validity, but lower internal validity) to show that the findings from multiple types of designs show similar findings (e.g., Forgette and Morris 2006).

In addition to increasing external validity, researchers designing and conducting future attitude change experiments could also increase their impact by not only assessing attitude change, but also assessing whether the attitude change that is observed lasts over any meaningful period of time or impacts behavior. These consequences of attitudes are one of the primary reasons political scientists are interested in attitudes, and thus demonstrating that observed attitude change also leads to these consequences (e.g., behavioral changes) is key for

demonstrating the importance of the research. Finally, future experiments studying attitude change should take into account (either by measuring or manipulating) cognitive processes, such as selective exposure and elaboration, that likely play a role in persuasion processes in actual political campaigns or media effects.

References

- Abelson, Robert P., Donald R. Kinder, Mark D. Peters, and Susan T. Fiske. 1982. "Affective and Semantic Components in Political Person Perception." *Journal of Personality and Social Psychology* 42: 619-30.
- Ajzen, Icek, and Martin Fishbein. 2005. "The Influence of Attitudes on Behavior." In *The Handbook of Attitudes*, eds. Doris Albarracin, Blair T. Johnson, and Mark P. Zanna. Mahwah, NJ: Erlbaum.
- Allport, Gordon W. 1935. "Attitudes." In *A Handbook of Social Psychology*, ed. Carl Murchison. Worcester, MA: Clark University Press.
- Andreoli, Virginia, and Stephen Worchel. 1978. "Effects of Media, Communicator, and Message Position on Attitude Change." *Public Opinion Quarterly* 42: 59-70.
- Baumgartner, Hans, and Jan-Benedict E.M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38: 143-56.
- Belli, Robert F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Over-Reporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63: 90-108.
- Bem, Daryl J. 1970. *Beliefs, Attitudes, and Human Affairs*. Belmont, CA: Brooks/Cole.
- Berinsky, Adam J. 2002. "Political Context and the Survey Response: The Dynamics of Racial Policy Opinion." *Journal of Politics* 64: 567-84.
- Bizer, George Y., and Richard E. Petty. 2005. "How We Conceptualize our Attitudes Matters: The Effects of Valence Framing on the Resistance of Political Attitudes." *Political Psychology* 26: 553-68.
- Boninger, David S., Timothy C. Brock, Thomas D. Cook, Charles L. Gruder, and Daniel Romer. 1990. "Discovery of Reliable Attitude Change Persistence Resulting from a Transmitter Tuning Set." *Psychological Science* 1: 268-71.
- Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American*

- Journal of Political Science* 52: 959-78.
- Brewer, Marilyn. 2000. "Research Design and Issues of Validity." In *Handbook of Research Methods in Social and Personality Psychology*, eds. Harry T. Reis, and Charles M. Judd. Cambridge: Cambridge University Press.
- Cantril, Hadley. 1934. "Attitudes in the Making." *Understanding the Child* 4: 13-15.
- Campbell, Angus, Phillip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: John Wiley and Sons.
- Campbell, Donald T. and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.
- Chong, Dennis, and James N. Druckman. 2007. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101: 637-55.
- Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. David E. Apter. New York: Free Press.
- Converse, Philip E. 1970. "Attitudes and Non-Attitudes: Continuation of a Dialogue." In *The Quantitative Analysis of Social Problems*, ed. Edward R. Tufte. Reading: MA: Addison-Wesley.
- Davis, James A., Tom W. Smith, and Peter V. Marsden. 2007. *General Social Surveys, 1972-2006 Cumulative Codebook*. Chicago: NORC, University of Chicago.
- Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98: 671-86.
- Druckman, James N., and Arthur Lupia. 2000. "Preference Formation." *American Review of Political Science* 3: 1-24.
- Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47: 729-45.
- Eagly, Alice H. and Shelly Chaiken. 1993. *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Javanovich College Publishers.
- Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. New Haven: Yale University Press.
- Forgette, Richard, and Jonathan S. Morris. 2006. "High-Conflict Television News and Public Opinion." *Political Research Quarterly* 59: 447-56.
- Gaffié, B. 1992. "The Processes of Minority Influence in an Ideological Confrontation." *Political Psychology* 13: 407-27.
- Gartner, Scott Sigmund. 2008. "The Multiple Effects of Casualties on Public Support for War:

- An Experimental Approach." *American Political Science Review* 102: 95-106.
- Glaser, James M. 2003. "Social Context and Inter-Group Political Attitudes: Experiments in Group Conflict Theory." *British Journal of Political Science* 33: 607-20.
- Henry, P. J., and David O. Sears. 2002. "The Symbolic Racism 2000 Scale." *Political Psychology* 23: 253-83.
- Himmelfarb, Samuel, and Carl Lickteig. 1982. "Social Desirability and the Randomized Response Technique." *Journal of Personality and Social Psychology* 43: 710-17.
- Huang, Li-Ning, and Vincent Price. 2001. "Motivations, Goals, Information Search, and Memory about Political Candidates." *Political Psychology* 22: 665-92.
- Huber, Gregory A., and Kevin Arceneaux. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51: 957-77.
- Iyengar, Shanto, Mark D. Peters, and Donald R. Kinder. 1982. "Experimental Demonstrations of the 'Not-So-Minimal' Consequences of Television News Programs." *The American Political Science Review* 76: 848-58.
- James, William. 1890. *The Principles of Psychology*. New York: Holt.
- Kaid, Lynda Lee, and John Boydston. 1987. "An Experimental Study of the Effectiveness of Negative Political Advertisements." *Communication Quarterly* 35:193-201.
- Kam, Cindy D. 2005. "Who Toes the Party Line? Cues, Values, and Individual Differences." *Political Behavior* 27: 163-82.
- Kinder, Donald R. 1998. "Opinion and Action in the Realm of Politics." In *Handbook of Social Psychology*, ed. Daniel T. Gilbert, Susan T. Fiske and Gardner Lindzey. New York: McGraw-Hill.
- Kinder, Donald R., and D. Roderick Kiewiet. 1979. "Economic Discontent and Political Behavior: The Role of Personal Grievances and Collective Economic Judgments in Congressional Voting." *American Journal of Political Science* 23: 495-527.
- Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Rudd, and V. Kerry Smith. 2002. "The Impact of 'No Opinion' Response Options on Data Quality: Prevention of Non-Attitude Reporting or an Invitation to Satisfice?" *Public Opinion Quarterly* 66: 371-403.
- Krosnick, Jon A., Allyson L. Holbrook, and Penny S. Visser. 2000. "The Impact of the Fall 1997 Debate about Global Warming on American Public Opinion." *Public Understanding of Science* 9: 239-60.

- Krosnick, Jon A., and Richard E. Petty. 1995. "Attitude Strength: An Overview." In *Attitude Strength: Antecedents and Consequences*, eds. Richard E. Petty, and Jon A. Krosnick. Hillsdale, NJ: Erlbaum.
- Kuklinski, James H., and Michael D. Cobb. 1998. "When White Southerners Converse About Race." In *Perception and Prejudice*, eds. Jon Hurwitz, and Mark Peffley. New Haven: Yale University Press.
- Kuklinski, James H., and Norman L. Hurley. 1994. "On Hearing and Interpreting Political Messages: A Cautionary Tale of Citizen Cue-Taking." *Journal of Politics* 56: 729-51.
- Lodge, Milton, Kathleen McGraw, and Patrick Stroh. 1989. "An Impression-Driven Model of Candidate Evaluation." *American Political Science Review* 83: 399-419.
- Markus, Gregory B. 1986. "Stability and Change in Political Attitudes: Observed, Recalled, and Explained." *Political Behavior* 8: 21-44.
- McGraw, Kathleen M., Edward Hasecke, and Kimberly Conger. 2003. "Ambivalence, Uncertainty, and Processes of Candidate Evaluation." *Political Psychology* 24: 421-48.
- Miller, Joanne M., and Jon A. Krosnick. 2000. "News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens are Guided by a Trusted Source." *American Journal of Political Science* 44: 301-15.
- Miller, Joanne M., and Jon A. Krosnick. 2004. "Threat as a Motivator of Political Activism: A Field Experiment." *Political Psychology* 25: 507-23.
- Miller, Richard B., and David W. Wright. 1995. "Detecting and Correcting Attrition Bias in Longitudinal Family Research." *Journal of Marriage and Family* 57: 921-9.
- Mondak, Jeffery J. 1990. "Perceived Legitimacy of Supreme Court Decisions: Three Functions of Source Credibility." *Political Behavior* 12: 363-84.
- Mutz, Diana C., and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Political Science Review* 99: 1-15.
- Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91: 567-83.
- Nelson, Thomas E., and Jennifer Garst. 2005. "Values-based Political Messages and Persuasion: Relationships among Speaker, Recipient, and Evoked Values." *Political Psychology* 26: 489-515.
- Peffley, Mark, and Jon Hurwitz. 2007. "Persuasion and Resistance: Race and the Death Penalty in America." *American Journal of Political Science* 51: 996-1012.
- Petty, Richard E., and John T. Cacioppo. 1986. *Communication and Persuasion: Central and*

- Peripheral Routes to Attitude Change*. New York: Springer-Verlag.
- Rahn, Wendy M., Jon A. Krosnick, and Marijke Breuning. 1994. "Rationalization and Derivation Processes in Survey Studies of Political Candidate Evaluation." *American Journal of Political Science* 38: 582-600.
- Rosenstone, Steven J., and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.
- Taris, Toon. 2000. *A Primer in Longitudinal Data Analysis*. Thousand Oaks, CA: Sage.
- Thurstone, Louis L, ed. 1931. *The Measurement of Social Attitudes*. Chicago: University of Chicago Press.
- Transue, John E. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51: 78-91.
- Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White. 2002. "Cues that Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96: 75-90.
- Visser, Penny S., and Jon A. Krosnick. 1998. "Development of Attitude Strength over the Life Cycle: Survey and Decline." *Journal of Personality and Social Psychology* 75: 1389-410.
- Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60: 63-9.
- Wolosin, Robert J., Steven J. Sherman, and Arnie Cann. 1975. "Predictions of Own and Other's Conformity." *Journal of Personality* 43: 357-78.

Figure 10-1. Pretest-Posttest Control Group Design (Campbell and Stanley 1963, 13)

Experimental:	R	O ₁	X	O ₂
Control:	R	O ₃		O ₄

Figure 10-2. Pretest-Posttest Multiple Experimental Condition Design

Experimental Group A:	R	O ₁	X _a	O ₂
Experimental Group B:	R	O ₃	X _b	O ₄

Figure 10-3. Posttest-Only Control Group Design

Experimental:	R	X	O ₁
Control:	R		O ₂

The key comparison in this design is whether O₁ differs from O₂.

Figure 10-4. Posttest-Only Multiple Experimental Group Design

Experimental Group A:	R	X _a	O ₁
Experimental Group B:	R	X _b	O ₂

The key comparison here is whether O₁ is different from O₂.

ⁱ Although attitudes have most often been conceptualized as a single bipolar continuum with positivity toward the object at one end, negativity in the other, and a neutral midpoint, theories of attitudes have also begun to consider the role of attitudinal ambivalence. Attitudinal ambivalence occurs when a person's beliefs or affect toward an attitude object are in conflict with one another (e.g., a person has both positive and negative beliefs about a political candidate; Eagly and Chaiken 1993).

ⁱⁱ Helman, Scott. 2008. "Many Voting for Clinton to Boost GOP Seek to Prolong Bitter Battle." *The Boston Globe*, March 17, p. A1.

ⁱⁱⁱ Analyses of existing data such as those collected by through the American National Election Study (ANES) surveys are exceptions to this, although this type of data analysis using the ANES data are relatively rare in the literature.

11. Conscious and Unconscious Information Processing with Implications for Experimental Political Science

Milton Lodge, Charles Taber, and Brad Verhulst

Affect-driven dual process models dominate contemporary psychological theorizing about how people think, reason, and decide (Chaiken and Trope 1999; Wilson Lindsey, and Schooler 2000; Gawronski and Bodenhausen 2006). Although most dual-process models focus on accuracy-efficiency tradeoffs, hundreds of more recent experiments document the pervasive effects of unconscious thoughts, feelings and behaviors on attitude formation, attitude change, preferences, and decision making. These studies reveal important differences between the influence of conscious and unconscious processing on how people think and reason. The explicit incorporation of unconscious cognition into models of political beliefs challenges the extant understanding of mass beliefs. Much of what we political scientists claim to know about citizens' political beliefs and attitudes is based on verbal self report. The vast majority of the empirical evidence in political behavior research is based directly on verbal responses to explicit questions. This reliance on explicit measures of political attitudes and behaviors is problematic, as these measures assume people have direct access to their 'true' beliefs or attitudes and are willing and able to accurately report them (Wittenbrink 2007).

Most of our daily life is experienced unconsciously, outside awareness. Consequently it is quixotic to focus exclusively on conscious attitudes while ignoring considerations that escape conscious awareness. Recent estimates put the total human capacity for visual sensory processing in the neighborhood of 10 million bits per second, though we can become conscious of only about 40 bits per second (Norretranders 1998). Although the absolute input from other

sensory modalities such as touch, smell, and hearing is considerably less than visual sensory input, the differential processing capacity between conscious and unconscious perception in these domains is similarly lopsided in favor of unconscious processing. Importantly people can only consciously process approximately 7 ± 2 chunks of information, at any given time irrespective of the type of information (Miller 1957). Given these serious limitations on conscious attention, various heuristic devices have evolved to reduce the amount of information that they must process consciously. Where and when conscious information-processing strategies prove to be more or less effective than unconscious information-processing strategies is a critical question. At the very least, the difference between the staggering amount of sensory input and the constraints of conscious processing leaves open the possibility for the introduction of unconscious processing into models of political behavior and decision making.

Both conscious and unconscious processes are continuously at work, not only when people make snap judgments, but contrary to 2000 years of Western thought, even when people are called on to think hard and weigh pros and cons before forming an attitude or making decisions. Research in the cognitive and neurocognitive sciences has used multiple labels to distinguish between these two styles of information processing in the formation and expression of beliefs, attitudes, goals, and behavior, chief among them the concepts explicit and implicit, deliberative and automatic, or System 2 and System 1 processing. This research demonstrates not only that unconscious processing can influence conscious attitudes and behaviors (e.g. Bargh, Chen, and Burrows 1996) but also that consciously activated goals can affect unconscious processing strategies (e.g. Aarts and Dijksterhuis 2000; Chaiken and Maheswaran 1994). Thus, although conscious and unconscious processing strategies can operate independently, it is also common for processing at one level to influence the processing at the other.

Related, though conceptually distinct from these processing strategies, are implicit and explicit attitudes. According to Greenwald, McGee, and Schwartz (1998) “Implicit attitudes are manifest as actions or judgments that are under the control of automatically activated evaluation, without the performer's awareness of that causation” (4). Thus, implicit attitudes are automatic, evaluative tendencies that people hold that influence their thoughts and behaviors outside of their conscious awareness. These implicit attitudes can be measured by a variety of approaches such as the Implicit Association Test (IAT; Greenwald et al. 1998), reaction times in a sequential priming paradigm (Fazio et al. 1986), or emotional transference procedures like the Affect Misattribution Procedure (AMP; Payne et al. 2005). By contrast, explicit attitudes are mediated by controlled, conscious thought and can be measured by survey techniques or other explicit verbal responses.

To simplify our discussion, we favor the terms *conscious* and *unconscious* when referring to information-processing styles, and reserve the terms *implicit* and *explicit* for attitudes or attitude measures. We do not assume a disjuncture between conscious and unconscious processing, as people can process stimuli both consciously and unconsciously and either style of information processing can influence both implicit and explicit attitudes (Monroe and Read 2009).

Conscious processing is simply information processing that we are aware of. The complement to this style of processing – unconscious processing – is all of the information processing we are unaware of. Most habits and heuristics would fall into the unconscious processing category, which does not imply that we are unaware of our habits or that we cannot use heuristics consciously (Lupia 1994; Lau and Redlawsk 2001), nor that we can think carefully and effectively only if we think consciously (Wilson and Schooler 1991; Dijksterhuis 2004).

We can only distinguish empirically between conscious and unconscious information processing, including their joint or independent effects, by using experimental methods, as research on unconscious processing requires control over how information is presented to determine when and how unconscious information is being processed. Thus the control required to identify the cognitive and affective mechanisms involved in these two styles of information processing negates the possibility of observational research.

1. Conscious and Unconscious Processing in the Political Domain

Political scientists routinely acknowledge several unconscious processing mechanisms in accounting for how people think, feel, or behave in the political world. We will focus on three areas of research in political science that incorporate unconscious processing into their explanations of political attitudes and preferences: online processing, implicit attitudes and measurement, and situations where the stimulus may be noticed but its influence on judgments and behaviors is unappreciated.

Online Information Processing

The online model holds that beliefs and attitudes are constructed in real time as people encounter information, and are integrated into existing networks of associations in long-term memory (Anderson and Barrios 1961; Hastie and Park 1986; Lodge, Steenbergen, and Brau 1995). Affect plays the critical role in this online updating process. When people form or revise their impressions of persons, places, events, or issues they spontaneously extract the affective value of the message and within milliseconds update their summary evaluation of the object. This “running tally” integrates new information with one’s prior evaluation of an object and is then restored to memory where it is readily available for subsequent retrieval (Casino and Lodge 2007). A central tenet of the online model is that the process by which people form attitudes is

not routinely mediated by conscious information processing: people do not intentionally form or update tallies, but rather evaluate people, events, and ideas spontaneously.

In an experimental setting the measurement of online processing proceeds in five stages. First, participants evaluate all the information that will subsequently be presented, plus other pieces of information that will not appear in the message, so as to later check for rationalization effects in recall. Next, there is a distracter task, perhaps questions asking for demographics. Then, participants read information about one or more candidates or issues, typically embedded in narrative form as a newspaper article or newscast. In the fourth stage, participants evaluate the candidate or issue. And finally, participants are asked to recall the information in the message, followed by questions probing for gist and details about the candidate's issue positions (Lodge et al. 1995). Note that within this context the measurement of the online evaluation is explicit, but there is now empirical evidence and theoretical rationale suggesting online processing is automatic: people evaluate and integrate their evaluations into a summary judgment effortlessly, outside of conscious awareness.

A well-replicated finding is that people can integrate a great deal of complex information into a summary evaluation in real time but prove unable to recall much of this information after a short time lapse (Hastie and Park 1986; Redlawsk 2001; Steenbergen and Lodge 2003). Specifically, the number of items people accurately recall decays exponentially, so that within days if not minutes the information a person remembers no longer captures the information that was presented in the message and no longer predicts the evaluation. Interestingly, the information people remember when engaged in online processing differs markedly from the information people recall when relying on memory-based processing. Specifically, online tallies show evidence of a primacy effect whereby a person's tally is anchored on the initial information

encountered about the object, whereas memory-based evaluations are heavily dependent on the most recent information encountered. More problematic still in political science practice is that the more people are encouraged to ruminate about a message, the greater the impact of rationalizations on memory (Wilson, Hodges, and La Fleur 1995; Erisen, Taber, and Lodge 2006). When they are called on to stop and think before responding, people tend to overemphasize accessible information and construct an attitude based on what is temporarily accessible (Zaller and Feldman 1992).

Early descriptions of these mechanisms drew too sharp a distinction between memory-based and online processing. An either/or view is theoretically flawed and empirically unfounded. The confusion stems from the failure to discriminate *encoding* from *retrieval* effects. The encoding process is inherently unconscious and occurs automatically regardless of whether subsequent retrieval focuses on the online tally or a broader sample of memory-based considerations. During encoding, affect and cognition become strongly linked in memory and difficult to disentangle. Affective tags are attached to concepts when an object is first evaluated and subsequently strengthened every time a person thinks about the object (Lodge and Taber 2005). Thus, when people rely on online processing, they simply report their general, gut reactions that are embedded in their affective tallies. When asked to explain their attitudes, their justifications rationalize their online tallies, biased by other accessible thoughts.

On retrieval, affect is primary in three senses. First, the affective component of a concept enters the decision stream earlier than the concept's semantic associations (you know you like or dislike Barack Obama before you remember he is a Democrat and an African American). Consequently, affective reactions anchor evaluative judgments (Zajonc 1980). Second, semantic information follows the oft-noted exponential forgetting curve for factual information, while

affective information remains relatively stable over longer periods of time (Lodge et al. 1995).

Third, online evaluations, spontaneously activated on mere exposure to a stimulus, bias the recall of information in affectively consistent ways, resulting in affectively driven rationalizations (Rahn, Krosnick, and Breuning 1994; Erisen et al. 2007). Positive candidate evaluations, for example, heighten the accessibility of other positive considerations in memory. Thus, if people are asked to justify their preferences, they simply search for accessible reasons for their automatic affective reactions rather than the actual information that formed the online evaluation initially.

An exemplary empirical demonstration of online processing was carried out by Betsch et al. (2001) who had participants watch a series of television commercials, telling them they would have to later recall and evaluate the advertised products. Simultaneously, subjects were engaged in a second, cognitively demanding distracter task: they were asked to read aloud the changing stock prices of five hypothetical companies presented on a crawler at the bottom of the TV screen. Although participants were led to believe the task assessed their ability to remember and evaluate the commercials while being distracted, the study actually tested whether they could track the stock ticker information. As predicted by the online model, participants were unable to recall the pertinent stock information, yet their evaluations correlated strongly with the actual stock prices of the five companies. This result points to the automaticity of online evaluations: experimental subjects accurately evaluated the companies' stock performances even when they actively focused on other information and they were unable to recall the stock prices.

From this perspective, an online tally anchors a person's evaluation of an object and spontaneously infuses the encoding, retrieval, and comprehension of subsequent information, its expression as a preference, and readies us to act aversively or appetitively in accord with our

evaluations (Ito and Cacioppo 2005). Importantly, this readiness to act in accord with one's summary evaluation need not be and typically is not mediated by conscious thought.

2. Implicit Attitudes

An appreciation of unconscious information processing of implicit attitudes has recently infiltrated political behavior research. Since all attitudes are latent constructs, they cannot be directly observed but must be inferred from self report or nonverbal responses such as reaction time. For present purposes, let us define an attitude rather generally as an evaluative orientation towards an object, whether a politician, an issue, social group, or abstract concept. Although implicit measures can be used for both explicit and implicit attitudes, the measurement of implicit attitudes demands indirect procedures.

The basic theory underlying tests of associations among concepts and their subsequent effects on behavior is the associative network model of cognition (Anderson 1983, 1993).¹ This theory is predicated on the well-documented observation that activation spreads along semantically and affectively associated pathways (Collins and Quillan 1968; Fazio et al. 1986; Collins and Loftus 1988; Bargh 1999). The more often two concepts have been linked, the more strongly they become associated. This is reflected in faster retrieval of one concept when the other is primed, influence of one concept on interpretations of the other, and greater likelihood of paired retrieval in free memory tasks. Within milliseconds of perceiving a concept (whether word or image), activation spreads automatically to associated concepts, including semantic and affective links. For example, mere milliseconds after seeing a picture of Barack Obama, activation spreads to affective associations (online tallies) and then other related concepts in memory such as "Democrat" or "War in Afghanistan". Because all social concepts are

affectively charged, this process will activate both primary (I like/dislike Obama) and secondary (I love/hate Democrats) evaluative associations (Fazio et al. 1986).

Implicit measures capitalize on this observation by measuring the empirical influence of carefully chosen stimuli on speed of response or content of recall. While specific procedures differ for the various implicit attitude measures, they all seek to assess the uncontrolled, unintended, stimulus driven, autonomous, and unconscious affective responses to stimuli. For example, reaction time approaches rely on the speed of response to a particular primed target as an indicator of the strength of association between the prime and target: the stronger the association (terrorist – bad), the faster the response. Absent conscious awareness, the response bypasses intentionality and more importantly taps beliefs and preferences that the individual may not be willing or able to consciously express.

There are several popular implicit attitude measures, chief among them: the sequential priming paradigm (Fazio et al. 1986), the IAT (Greenwald et al. 1998), and the AMP (Payne et al. 2005). Each procedure relies on slightly different cognitive mechanisms to assess attitudes.

The *Sequential Priming Paradigm* provides the most direct test of the strength of associations (Figure 11-1). Here subjects are presented with a prime followed by a target word or phrase. The subject's task is to respond to the target by pressing one response button if the target is a member of category X or another if the target is a member of the category Y. As the dependent variable is the latency between the initial presentation of the target and the response, participants are instructed to respond "as quickly as possible without making too many errors." For strongly associated concepts, whether semantic (BUSH – REPUBLICAN) or affective (OBAMA – RAINBOW), the activation created by the prime allows people to respond to the target faster than when the concepts are unrelated (TREE – REPUBLICAN) or when non-word

foils are used as primes (BLUM – REPUBLICAN). Note here, this *facilitation effect* measures a relatively strong association between the prime and the target in long-term memory compared to a baseline, while an *inhibition effect* would be signaled by a slower-than-baseline response time. This simple priming paradigm produces robust effects demonstrating the associative nature of both semantic and affective memory.

[Insert Figure 11-1 Here]

A variant of the sequential priming paradigm allows experimenters to assess whether these associations require conscious mediation. Specifically, the experimenter can manipulate the time from the onset of the prime word to the onset of the target word, an interval known as the *stimulus onset asynchrony* (SOA). Conscious expectancies require more than 300 milliseconds to develop (Neely 1977; Posner, Snyder, and Davidson 1980), so any inhibition or facilitation effects observed when the SOA is shorter than 300 milliseconds are necessarily due to automatic activation of non-conscious associations (Bargh et al. 1992). As shown in Figure 11-1, when the SOA is short, the target is presented near the peak of the activation of the stimuli, and as such, it takes less time for a participant to respond to subsequent related stimuli. Alternatively, when the SOA is long, the activation of the prime has returned to near baseline levels of activation when the target stimulus is presented. Thus, no facilitation is expected at a long SOA. Contemporary priming studies routinely present primes at subliminal speeds as short as 14 ms, many times faster than the blink of an eye, to remove all doubt about conscious mediation or control.

The IAT, in contrast to the sequential priming paradigm, capitalizes on response competition rather than spreading activation. The IAT compares the difference between the average response time to two blocks of categorization trials when all of the trials within the block are either affectively congruent (Column 3 in Table 11-1) or affectively incongruent (Column 5

in Table 11-1). The other blocks of trials are used to familiarize the participants with the task. Using the racial IAT as an example, in the congruent condition, European-American stereotypes are paired with pleasant words and African-American stereotypes are paired with unpleasant words. Because the evaluative tendencies for African-American stereotypes and unpleasant words match well-practiced behavioral predispositions for many Americans, participants need not inhibit competing responses before reacting to the stimuli, allowing most people to respond relatively quickly to these trials. Alternatively, in the incongruent exercise, where European-American stereotypes are paired with unpleasant words and African-American stereotypes are paired with pleasant words, the categories are evaluatively incongruent and activate competing automatic behavioral tendencies that respondents must override before they categorize the stimuli, and thus they respond relatively slowly.

In the final procedure we discuss – the AMP (Payne et al. 2005) – the mechanism at work is spontaneous affective transfer. But here, rather than assessing the existing associations between concepts in memory, the AMP creates new associations by repeatedly pairing stimuli toward which people already have an attitude (New York City) with previously neutral stimuli (like Mandarin ideograms), analogous to classical conditioning. During the repeated pairings, the evaluative associations from the attitudinal stimuli transfer to the neutral stimuli. When participants are asked to rate how much they like the previously neutral stimulus, their affect indicates the strength and direction of their attitude toward the previously affective stimulus. Because the affective transfer occurs outside of conscious awareness, respondents are not motivated to alter or misrepresent their attitudes toward the previously neutral stimuli. In fact, participants perform no better than chance at identifying which stimuli were paired. Tests of the AMP demonstrate convergent validity with other attitude measures (Payne et al. 2005).

Each of these procedures has unique strengths and weaknesses. One of the strengths of the IAT in comparison to sequential priming is its strong test-retest reliability (Cunningham, Preacher, and Banaji 2001). Priming experiments – especially those using subliminal primes – are noisy, and participants may miss the prime in the blink of an eye. Sequential priming, however, allows the researcher to assess target-level facilitation and inhibition effects between individual pairs of primes and targets. This is important as it allows one to examine the strength of the association between the specific primes and particular targets. The researcher can modify the list of prime/target pairs to better capture nuanced differences between category subgroups (e.g., liked versus disliked Republican politicians).

The primary drawback of the IAT is that it leads participants to overemphasize the stimulus categories (i.e. race) rather than characteristics of individual stimuli. In one particularly revealing study, participants were presented with pleasant and unpleasant words, and photos of African-American athletes and European-American politicians (Mitchell, Nosek, and Banaji 2003). When participants categorized the photos on the basis of race, they demonstrated the typical implicit racial bias – Blacks were evaluated more negatively than Whites. But, when asked to categorize by occupation (athlete vs. politician), the reverse results were found: participants preferred the African-American faces to the European-American faces.

Forcing participants to make categorical judgments may activate concepts in memory that are not actually part of an individual's personal attitude, but rather tap awareness of cultural stereotypes. To rectify this, Olson and Fazio (2004) asked participants to categorize stimuli in terms of "I Like" or "I Don't Like," rather than good or bad. Importantly, this personalized version of the IAT strengthens the correlation between the IAT and explicit measures of racism although still revealing a depressingly large proportion of Americans who endorse but are

unwilling to express blatantly prejudicial attitudes. Furthermore, the personalized IAT is more strongly linked to the participant's actual behaviors than the traditional IAT (Dovidio, Kawakami, and Gaertner 2002; Olson and Fazio 2004).

Because the AMP is a relatively novel procedure, its strengths and weaknesses are less well-identified. The procedural similarity between the AMP and the sequential priming paradigm suggests that the same limitations may exist for both measures (Deutsch and Gawronski 2009). Early evidence, however, suggests that the AMP may be more resistant to conscious control than either the IAT or sequential priming, as even if participants are motivated to obscure their explicit attitudes, they cannot identify which neutral stimuli were paired with which attitude object (Payne et al. 2005)

One of the key benefits of using implicit rather than explicit attitude measures is that implicit attitudes are typically more predictive of actual behavior (Swanson, Rudman, and Greenwald 2001; Maison, Greenwald, and Bruin 2004; Amodio and Devine 2006). This finding is underscored by the fact that the relationship between implicit and explicit attitude measures is notoriously weak (Nosek 2004). As such, simply using implicit attitude measures can improve our understanding of behavior.

Modified versions of implicit attitude measures can also be used to answer interesting questions that go well beyond the simple measurement of preferences and provide a unique approach to understanding the structure of a person's political attitudes. For example, the central claim of the *hot cognition hypothesis* is that all social concepts are affectively charged (Abelson 1963; Lodge and Taber 2005). This evaluative component is directly associated with the concept in memory, where it is automatically activated on mere exposure to the concept (Fazio et al. 1986; Bargh 1999). This is not to say, however, that automatic semantic associations are less

important. Both types of associations are integral aspects of a person's attitude. In fact, using the sequential priming paradigm, the influence of each type of association can be directly tested, and the unique contributions of the semantic and affective components can be assessed.

Implicit attitude measures also make it possible to compare the associations between concepts that may vary across different groups of people and that may differ from the associations identified by explicit attitude measures. For example, survey evidence suggests that the ideological beliefs of liberals and conservatives are categorically different: liberals and conservatives see the political world in different ways (Conover and Feldman 1984). This suggests that the structure of a liberal's attitudes and thus the associations among concepts in memory depend on his political orientation. Implicit attitude measures provide a method to test these differential associations across groups. Three studies have attempted to do just this and their results merit discussion.

First, we find marked differences between automatic facilitation and inhibition effects for political sophisticates and nonsophisticates (Lodge and Taber 2005). Citizens who have repeatedly thought about and evaluated political leaders, groups, and issues have stronger associations among the ideological concepts, resulting in more extreme experimental facilitation and inhibition effects. Citizens with below average political interest and knowledge have weaker affective and semantic links in memory for many political concepts, and therefore do not display the same pattern of facilitation and inhibition that indicates hot cognition (McGraw and Steenbergen 1995). Thinking about an attitude object, whether intentionally or unintentionally, brings the attitude object into memory, and activates other associated constructs. The more often two constructs are associated in memory, the stronger the association between them. This overall pattern suggests that because of their interest in politics, sophisticates have formed affective and

semantic associations toward a broad range of political objects, and these feelings and thoughts come spontaneously to mind on mere (even subliminal) exposure to the concept.

In a related study, Lavine et al. (2002) demonstrated that primes selectively facilitate different target concepts in different groups of respondents. In this study, high authoritarians responded more quickly to ambiguously threatening prime/target pairs (arms-weapons), while low authoritarians responded more quickly to semantically related neutral prime/target pairs (arms-legs). This pattern of differential activation suggests that authoritarians perceive concepts in a categorically different way when compared with nonauthoritarians. Thus, words have categorically different automatic meanings for different groups of people.

Finally, Taber (2009) assessed the unconscious associations between racially charged political issues, specifically testing whether attitudinal structures corresponded with principled conservatism or symbolic racism. “Affirmative action” was significantly associated with African-American stereotypes for both supporters and opponents of affirmative action. Contrary to the expectations of principled conservatism, however, neither “individualism” nor “big government” was associated with the issue. In a follow-up study, “individualism” and “big government” were associated with affirmative action only when white conservatives were explicitly asked to think about the issue when preparing for a debate with an affirmative action supporter, and not when they expected a like-minded conservative. Such findings suggest that principled conservatism allows affirmative action opponents to rationalize their opposition to affirmative action in ideological terms.

These studies demonstrate that implicit attitude measures make it possible to assess the content of ideological preferences in a way that explicit self-report measures simply cannot do. As is evident from the above discussion, implicit attitude measures can be employed to test a

wide variety of research questions. These measures can be used as both independent and dependent variables, or they can be used to examine the cognitive structure of interconnected (or not) attitudes and associations between concepts.

Unconsciously presented or Unnoticed Stimuli

The final example of how unconscious processing has permeated research in political science is in demonstrations of the impact of unconscious stimuli on how people think and reason about political issues. First, let us operationalize the distinction between conscious and unconscious stimuli. Visual stimuli take approximately 10 milliseconds to reach the *objective* perceptual threshold – as measured by brain activity – and between 60 and 100 milliseconds to reach the *subjective* threshold – after which conscious processing is possible (Bargh and Pietromonaco 1982). If the objective threshold is not passed there is no registration of the stimulus and perception does not occur: a nonevent. If the objective threshold is passed but the subjective is not, we have unconscious perception. In this case, a sensory experience is registered but people are unaware of their perception. Finally, if the subjective threshold is passed, the stimulus event enters conscious awareness and we have the possibility of conscious perception, but only if the individual attends to the stimuli. But just because a stimulus passes the subjective threshold does not guarantee that it will be processed consciously.

Accordingly, a stimulus may be processed unconsciously under two conditions. First, an objectively perceived stimulus may not reach conscious awareness because it occurred too rapidly or peripherally to be noticed, thus necessitating unconscious processing. Alternatively, and this is the most common reason for a stimulus being processed unconsciously in the real world, the individual “sees” the stimulus but fails to recognize its influence on thoughts, feelings,

preferences, and choices. The event influences behavior but its impact remains unappreciated. In both cases, the stimulus influences information processing outside of awareness.

The distinction between conscious and unconscious perception should not be seen as either/or but as interdependent. Our tripartite distinction implies an inherent ordering, since all stimuli that pass the subjective threshold necessarily pass the objective threshold. Thus, it is important to keep in mind that unconscious processes are omnipresent and inevitably influence subsequent thoughts, feelings and behaviors. As such, all conscious thoughts and feelings have an unconscious origin. Consequently, sometimes we just do things without thinking and sometimes thoughts just seem to pop into mind.

Unconscious stimuli are ubiquitous in the real world: the playthings of advertisers who use beautiful women to peddle cigarettes and sports cars, where the American flag and upbeat music provide the backdrop for presidential candidates, or where discordant music sets the tone for political attack ads. Whether consciously unnoticed or simply unappreciated, such “incidental” stimuli commonly and powerfully influence political beliefs and attitudes, in part because they are so easy to manipulate.

A good example of how unconscious stimuli influence conscious thought and behavior is a now-classic set of experiments conducted by Bargh et al. (1996). They primed various concepts outside the participants’ awareness and recorded their subsequent behaviors. Remarkably, after being primed with the concept “elderly,” participants walked more slowly to the elevator after completing the experiment. When participants were primed with words related to rudeness or politeness, they were, correspondingly, more or less likely to interrupt an experimenter engaged in an extended conversation. When primed with African-American (versus European-American)

faces, participants expressed more hostility toward the experimenter. The participants were completely unaware that their behaviors were driven by unconscious motives.

Extending the notion of unnoticed priming into the political arena, Mendelberg (2001) demonstrated that exposure to racially charged advertising propelled racially resentful people to hold more prejudicial beliefs. Specifically, in the 1988 presidential election, George H. W. Bush ran an attack ad against his opponent Michael Dukakis that featured an African-American prisoner named Willie Horton. Horton was released from prison on a weekend furlough, where he escaped from police custody, raped a woman and assaulted her husband. The ad attacked Dukakis's policies on crime; however, it was the racial undertones of the ad that drove racial resentment and led to the increase in negative evaluations of Dukakis (also see Kinder and Sanders 1996). Importantly, when the racial component of the ad was made explicit to viewers, the negative effect on evaluations evaporated. Thus, insofar as people were not consciously aware of the racial component of the ad, the ad was effective.

Thinking carefully, however, does not inevitably negate the impact of unconsciously perceived or unnoticed influences. In a series of experiments, Wilson et al. (1995) and Wilson and Schooler (1991) found that when participants are encouraged to stop and think before making a choice they overanalyze available information, bring to mind less important considerations, and consequently make poorer choices than when making a snap judgment. Current research also suggests that as people increasingly engage in effortful deliberation, the quality of the decision deteriorates in both objective terms (people make the wrong choice) and in subjective terms (higher levels of postdecision regret; Dijksterhuis 2004). In addition to this counterintuitive finding, it appears that people make the best choices if they are distracted from engaging in conscious thought or discouraged from thinking of the reasons for their preferences.

Along these lines, Verhulst, Lodge, and Taber (2007) explored the influence of subliminal primes on evaluations of political candidates when participants are given substantive issue information about the candidate's preferences. Commonly, people believe that the effects of subliminal primes are fleeting and that they would be overwhelmed by actual, substantive information. After all, candidate-voter issue proximity is among the most robust predictors of candidate evaluations and vote choice (Black 1948; Berelson, Lazarsfeld, and McPhee 1954; Downs 1957). Interestingly, we found that subliminal primes influenced candidate evaluations for sophisticated participants: positive primes led sophisticates to like the candidate more, negative primes, less. In a follow-up study, we randomly assigned participants to think carefully about the candidates. Rather than reducing the impact of the subliminal primes, careful, conscious deliberation increased the impact of the primes on candidate evaluations even though participants were unaware of being primed. Again, the effect appeared among the most sophisticated participants, confirming the theoretical hypothesis that only knowledgeable participants have the density of semantic and affective associations necessary for implicit primes to influence the quality of thought. While more research in this area is needed, it now appears that the unconscious information can drive conscious deliberation, culminating in preferences that are strongly influenced by information participants never consciously perceived.

Erisen et al. (2007) addressed this issue from a slightly different angle. Specifically, they encouraged participants to stop and think about political issues while being subliminally primed with smiling, frowning, or neutral cartoon faces. In this set of studies, the subliminal primes biased subsequent cognitive deliberation. Participants primed with positive images listed more positive thoughts, while those primed with negative images list more negative thoughts. Again,

participants were completely unaware of the primes, yet this biased set of prime-induced thoughts led to more extreme attitudes.

In one particularly telling example, Berger, Meredith, and Wheeler (2008) showed that budgetary support for education varied as a function of where people voted – whether in schools, churches, or firehouses – with voters more likely to favor raising state taxes to support education if voting in schools. This effect held even after controlling for their political views. Clearly, the voters knew what building they were in but were, in all likelihood, not consciously aware of its influence on their behavior.

Another major area of research pointing to robust effects of unappreciated influences on judgment is the effect of facial attractiveness on evaluations, attitudes, and behavior. Beautiful-is-good stereotyping is alive in politics, where attractive candidates are seen as possessing more integrity, competence, likeableness and fitness for public office even though people deny being affected by the candidate's appearance (Rosenberg and McCafferty 1987; Verhulst, Lodge, and Lavine 2010). Three large meta-analyses covering over 1000 peer-reviewed psychological studies of physical attractiveness confirm significant experimental and correlational effects on a broad range of social attitudes and behaviors (Eagly and Makhijani 1991; Feingold 1992; Langlois et al. 2000). Typically, physical attractiveness is noticed but people consistently fail to appreciate its impact on evaluations and behavior, yet the magnitude of these effects is roughly the same as other variables in the social sciences (Eagly and Makhijani 1991).

These studies and many more demonstrate the influence of unappreciated information on perception, social judgments, and behavior. The take home point here is that our thoughts, attitudes and behaviors extend deep into our unconsciousness. Similar noticed but unappreciated

effects are no doubt ubiquitous in everyday life outside the laboratory (Bargh 1997), but experimental methods are crucial to tease out the effects.

3. Conclusion: The Unconscious Mind in a Conscious World

People have known for many years that unconsciously perceived thoughts and feelings influence behavior. Recent theoretical advances and experimental procedures allow social scientists to measure and behaviorally validate the effects of unconscious processing and implicit attitudes on complex behaviors. Implicit attitude measures are now at the forefront of contemporary psychology and marketing research, and are now working their way into political science. This research offers the potential to explore more deeply the psychological mechanisms that lead citizens to form all sorts of political attitudes, and examine the way these political attitude influence various political behaviors.

The attitude-behavior connection in most political science research is routinely made by inference and assumption. Political scientists are interested in behavioral variables – voting, petitioning, attending rallies, contributing to campaigns – but our variables are most often verbalized intentions or recollections and not observed behaviors. This accentuates the potential for social desirability to influence responses. Moreover, citizens can only verbalize the thoughts and behavioral intentions that they are aware of. The reconceptualization of political information processing that we propose challenges our discipline’s reigning assumption that political beliefs, attitudes, and behaviors are rooted in conscious considerations, questions our disciplines’ reliance on survey research, and promises better explanations of political behavior. We see strong theoretical and growing empirical reasons to believe that incorporating unconscious processing into our models of political attitudes and behavior will result in a more complete

understanding of the relationships among and between political beliefs, preferences, choices and consequential political behaviors.

References

- Aarts, Henk, and Ap Dijksterhuis. 2000. "Habits as Knowledge Structures: Automaticity in Goal-Directed Behavior." *Journal of Personality and Social Psychology* 78: 53-63.
- Abelson, Robert. 1963. "Computer Simulation of 'Hot' Cognition." In *Computer Simulation of Personality*, eds. Silvan Tomkins, and Samuel Messick. New York: Wiley.
- Amodio, David M., and Patricia G. Devine. 2006. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 91: 652-61.
- Anderson, John. 1983. *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- Anderson, John. 1993. *Rules of the Mind*. Hillsdale, NY: Erlbaum
- Anderson, Norman H., and Alfred A. Barrios. 1961. "Primacy Effects in Person Impression Formation." *Journal of Abnormal Social Psychology* 43: 346-50.
- Bargh, John. 1997. "The Automaticity of Everyday Life." In *Advances in Social Cognition* 10, ed. Robert Wyer. Mahwah, NJ: Erlbaum.
- Bargh, John. 1999. "The Cognitive Monster: The Case Against Controllability of Automatic Stereotype Effects." In *Dual Process Theories in Social Psychology*, eds. Shelley Chaiken, and Yacov Trope. New York: Guilford.
- Bargh, John, Shelley Chaiken, Rajen Govender, and Felicia Pratto. 1992. "The Generality of the Automatic Attitude Activation Effect." *Journal of Personality and Social Psychology* 62: 893-912.
- Bargh, John, Annette Chen, and Lara Burrows. 1996. "Automaticity of Social Behavior: Direct Effects of the Trait Construct Stereotype Activation." *Journal of Personality and Social Psychology* 71: 230-44.
- Bargh, John, and Paul Pietromonaco. 1982. "Automatic Information Processing and Social Perception: The Influence of Trait Information Presented Outside of Conscious Awareness on Impression Formation." *Journal of Personality and Social Psychology* 43: 437-49.
- Berelson, Bernard, Paul Lazarsfeld, and William McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.

- Berger, Jonah, Marc Meredith, and S. Christian Wheeler. 2008. "Contextual Priming: Where People Vote Affects How They Vote." *Proceedings of the National Academy of Sciences* 105: 8846-9.
- Betsch, Tilman, Henning Plessner, Christine Schwieren, and Robert Gutig. 2001. "I Like It But I Don't Know Why: A Value Account Approach to Implied Attitude Formation." *Personality and Social Psychology Bulletin* 27: 242-53.
- Black, Duncan. 1948. "On the Rationale of Group Decision Making." *Journal of Political Economy* 56: 23-34.
- Cassino, Daniel, and Milton Lodge. 2007. "The Primacy of Affect in Political Cognition." In *The Affect Effect: Dynamics of Emotion in Political Thinking and Behavior*, eds. Russell Newman, George Marcus, Ann Crigler, and Michael MacKuen. Chicago: University of Chicago Press.
- Chaiken, Shelly, and Durairaj Maheswaran. 1994. "Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgment." *Journal of Personality and Social Psychology* 66: 460-73.
- Chaiken, Shelley, and Yaacov Trope. 1999. *Dual-Process Theories in Social Psychology*. New York: Guilford.
- Collins, Allan, and Elizabeth Loftus. 1975. "A Spreading-Activation Theory of Semantic Processing." *Psychological Review* 82: 407-28.
- Collins, Allan, and Ross Quillian. 1968. "Retrieval Time in Semantic Memory." *Journal of Verbal Learning and Verbal Behavior* 8: 240-7.
- Conover, Pamela J., and Stanley Feldman. 1984. "How People Organize the Political World: A Schematic Model." *American Journal of Political Science* 28: 95-126.
- Cunningham, William, Kristopher Preacher, and Mahzarin Banaji. 2001. "Implicit Attitude Measures: Consistency, Stability, and Convergent Validity." *Psychological Science* 1: 163-70.
- Deutsch, Roland, and Bertram Gawronski, 2009. "When the Method Makes a Difference: Antagonistic Effects on 'Automatic Evaluations as a Function of Task Characteristics of the Measure.'" *Journal of Experimental Social Psychology* 45: 101-14.
- Dijksterhuis, Ap. 2004. "Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making." *Journal of Personality and Social Psychology* 87: 586-98.
- Dovidio, John, Kaerry Kawakami, and Samuel Gaertner. 2002. "Implicit and Explicit Prejudice and Interracial Interaction." *Journal of Personality and Social Psychology* 82: 62-8.
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Addison-Wesley.

- Eagly, Alice, and Mona Makhijani. 1991. "What is Beautiful is Good, but...: A Meta-Analytic Review of Research on the Physical Attractiveness Stereotype." *Psychological Bulletin* 110: 109-29.
- Erisen, Cengiz, Milton Lodge, and Charles Taber. 2007. "The Role of Affect in Political Deliberation." Presented at the annual meeting of the American Political Science Association, Chicago, IL.
- Fazio, Russell, David Sanbonmatsu, Martha Powell, and Frank Kardes. 1986. "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology* 50: 229-38.
- Feingold, Alan. 1992. "Good-looking People are Not What We Think." *Psychological Bulletin* 111: 304-41.
- Gawronski, Bertram, and Galen V. Bodenhausen. 2006. "Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change." *Psychological Bulletin* 132: 692-731.
- Greenwald, Anthony, Debbie McGhee, and Jordan Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74: 1464-80.
- Hastie, Reid, and Bernadette Park. 1986. "The Relationship Between Memory and Judgment Depends on Whether the Judgment Task is Memory-based or On-line." *Psychological Review* 93: 258-68.
- Ito, Tiffany, and John Cacioppo. 2005. "Attitudes as Mental states of Readiness: Using Physiological Measures to Study Implicit Attitudes." In *Implicit Measures of Attitudes*, eds. Brend Wittenbrink, and Norbert Schwartz. New York: Guilford Press.
- Kim, Sung-youn Kim, Milton Lodge, and Charles Taber. 2009. "A Computational Model of the Citizen as Motivated Reasoner: Modeling the Dynamics of the 2000 Presidential Election." *Political Behavior* 32: 1-28.
- Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. Chicago: University of Chicago Press.
- Langlios, Judith, Lisa Kalakanis, Adam Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. 2000. "Maxims or Myths of Beauty? A Meta-analytic and Theoretical Review." *Psychological Bulletin* 126: 390-423.
- Lau, Richard, and David Redlawsk. 2001. "Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making." *American Journal of Political Science* 45: 951-71.
- Lavine, Howard, Milton Lodge, Jamie Polichak, and Charles Taber. 2002. "Explicating the Black Box Through Experimentation: Studies of Authoritarianism and Threat." *Political Analysis* 10: 342-60.

- Lodge, Milton, and Charles Taber. 2005. "The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis." *Political Psychology* 26: 455-82.
- Lodge, Milton, Marco Steenbergen, and Shawn Brau. 1995. "The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89: 309-26.
- Lupia, Arthur, 1994. "Shortcuts versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections." *American Political Science Review* 88: 63-76.
- Maison, Dominika, Anthony G. Greenwald, and Ralph H. Bruin. 2004. "Predictive Validity of the Implicit Association Test in Studies of Brands, Consumer Attitudes, and Behavior." *Journal of Consumer Psychology* 14: 405-15.
- McGraw, Kathleen, and Marco Steenbergen. 1995. "Pictures in the Head: Memory Representations of Political Candidates." In *Political Judgment: Structure and Process*, eds., Milton Lodge, and Kathleen McGraw. Ann Arbor: University of Michigan Press.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Miller, George. 1957. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63: 81-97
- Mitchell, Jason, Brian Nosek, and Mahzarin Banaji. 2003. "Contextual Variations in Implicit Evaluation." *Journal of Experimental Psychology: General* 132: 455-69.
- Monroe, Brian M., and Stephen J. Read. 2008. "A General Connectionist Model of Attitude Structure and Change: The ACS (Attitudes as Constraint Satisfaction) Model." *Psychological Review* 115: 733-59.
- Neely, James. 1977. "Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited Capacity Attention." *Journal of Experimental Psychology: General* 106: 226-54.
- Norretranders, Tor. 1998. *The User Illusion: Cutting Consciousness Down to Size*. New York, Penguin Press.
- Nosek, Brian. 2005. "Moderators of the Relationship Between Implicit and Explicit Evaluation." *Journal of Experimental Psychology: General* 134: 565-684.
- Olson, Michael, and Russell Fazio. 2004. "Reducing the Influence of Extrapersonal Associations on the Implicit Association Test." *Journal of Personality and Social Psychology* 86: 654-67.

- Payne, Keith, Clara Cheng, Olesya Govorun, and Brandon Stewart. 2005. "An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement." *Journal of Personality and Social Psychology* 89: 277-93.
- Posner, Michael, Charles Snyder, and Brian Davidson. 1980. "Facilitation and Inhibition in the Processing of Signals." *Journal of Experimental Psychology: General* 109: 160-74.
- Rahn, Wendy, Jon Krosnick, and Marijke Breuning. 1994. "Rationalization and Derivation Processes in Survey Studies of Candidate Evaluation." *American Journal of Political Science* 38: 582-600.
- Redlawsk, David. 2000. "You Must Remember This: A Test of the On-Line Model of Voting." *Journal of Politics* 63: 29-58.
- Rosenberg, Shawn, and Patrick McCafferty. 1987. "The Image and the Vote: Manipulating Voter's Preferences." *Public Opinion Quarterly* 51: 31-47.
- Steenbergen, Marco, and Milton Lodge. 2003. "Process Matters: Cognitive Models of Candidate Evaluation." In *Electoral Democracy*, eds. Michael MacKuen, and George Rabinowitz. Ann Arbor: University of Michigan Press.
- Swanson, Jane E., Laurie Rudman, and Anthony G Greenwald. 2001. "Using the Implicit Association Test to Investigate Attitude-Behaviour Consistency for Stigmatised Behavior." *Cognition & Emotion* 15: 207-30.
- Taber, Charles. 2009. "Principles of Color: Implicit Race, Ideology, and Opposition to Race-Conscious Policies." Unpublished paper, Stony Brook University.
- Verhulst, Brad, Milton Lodge, and Howard Lavine. 2010. "The Attractiveness Halo: Why Some Candidates are Perceived More Favorably than Others." *Journal of Nonverbal Behavior* 34: 111-17.
- Verhulst, Brad, Milton Lodge, and Charles Taber. 2007. "Automatic Projection: How Incidental Affect Alters the Perceptions of Political Candidates." Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Wilson, Timothy, Sarah Hodges, and S. LaFleur. 1995. "Effects of Introspecting About Reasons: Inferring Attitudes from Accessible Thoughts." *Journal of Personality and Social Psychology* 69: 1-28.
- Wilson, Timothy, Samuel Lindsey, and Tonya Schooler. 2000. "A Model of Dual Attitudes." *Psychological Review* 107: 101-26.
- Wilson, Timothy, and Jonathan Schooler. 1991. "Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions." *Journal of Personality and Social Psychology* 60: 181-92.

Wittenbrink, Bernd. 2007. "Measuring Attitudes Through Priming." In *Implicit Measures of Attitudes*, eds. Bernd Wittenbrink, and Norbert Schwartz. New York: The Guilford Press.

Zajonc, Robert. 1980. "Feeling and Thinking: Preferences Need no Inferences." *American Psychologist* 35: 117-23.

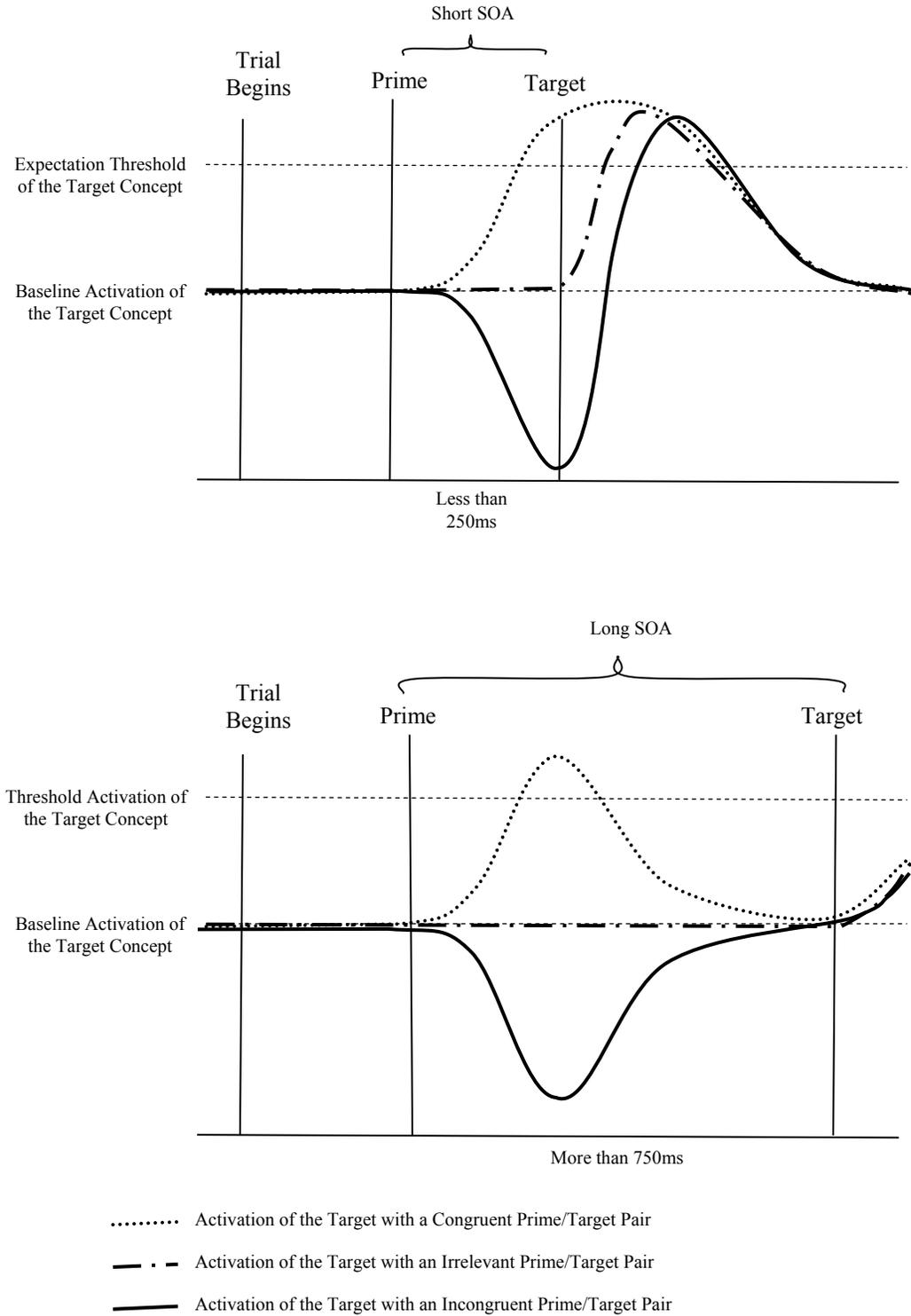
Zaller John, and Stanley Feldman. 1992. "A Simple Theory of Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 35: 579-616.

Table 11-1. A Schematic Figure of the Racial IAT using pleasant and unpleasant words and Euro-American and Afro-American Stereotype words

Experimental Stages	1	2	3	4	5
Task Description	Pleasant-Unpleasant Categorization	Euro-American- Afro-American Stereotypes Categorization	Congruent Categorization Task	Euro-American- Afro-American Stereotypes Categorization (Reversed)	Euro-American- Afro-American Stereotypes Categorization (Reversed)
Task Instructions	<ul style="list-style-type: none"> • Pleasant Unpleasant • 	<ul style="list-style-type: none"> • Euro-American Afro-American • 	<ul style="list-style-type: none"> • Pleasant Unpleasant • • Euro-American Afro-American • 	<ul style="list-style-type: none"> • Euro-American Afro-American • 	<ul style="list-style-type: none"> • Euro-American Afro-American •
Sample Stimuli	<ul style="list-style-type: none"> DEATH • SAD • • LAUGH • KITTEN GRIEF • VOMIT • • LOVE • JOY 	<ul style="list-style-type: none"> • SCIENTIST • COLLEGE LAZY • • ASSERTIVE STUPID • • THOUGHTFUL AGGRESSIVE • ATHLETIC • 	<ul style="list-style-type: none"> DEATH • SAD • • COLLEGE • THOUGHTFUL AGGRESSIVE • • KITTEN ATHLETIC • • SCIENTIST 	<ul style="list-style-type: none"> • ATHLETIC ASSERTIVE • • STUPID COLLEGE • • LAZY THOUGHTFUL • • AGGRESSIVE SCIENTIST • 	<ul style="list-style-type: none"> • ATHLETIC ASSERTIVE • • STUPID COLLEGE • • LAZY THOUGHTFUL • • AGGRESSIVE SCIENTIST •

Note: Dots beside the words indicate whether the left or right button should be pressed when the word is presented.

Figure 11-1. Spreading Activation in a Sequential Priming Paradigm for Short and Long SOA



ⁱ A fully axiomatized computational model, which we call *John Q. Public* (JQP), appears in Kim, Taber, and Lodge (2010).

12. Political Knowledge

Cheryl Boudreau and Arthur Lupiaⁱ

In many political surveys, many citizens fail to answer, or provide incorrect answers to, fact-based questions about political figures and institutions. A common inference drawn from such failures is that citizens' poor performance on surveys reflects their incompetence in democratically meaningful contexts such as voting booths.

The scholarly home for such findings is the academic literature on political knowledge. A common analytic definition of political knowledge is that it is a measure of a citizen's ability to provide correct answers to a specific set of fact-based questions.ⁱⁱ Typical political knowledge questions include "What is the political office held by [name of current vice president, British prime minister, or Chief Justice of the United States]?" and "Which political party has the most seats in the U.S. House of Representatives?" Many people have used responses to survey-based political knowledge questions to criticize the public for its general incompetence.

In recent years, these criticisms have come under increasing scrutiny (e.g., Graber 1984; Popkin 1994). Some scholars raised questions about the practice of basing broad generalizations of citizen competence or knowledge on a relatively small set of idiosyncratic, fact-based survey questions (e.g., Lupia 2006). Others uncovered logical and factual errors in claims about the kinds of political knowledge that are needed to make important political choices competently (e.g., Lupia and McCubbins 1998; Gibson and Caldeira 2009).

A common theme in this new research is that many critiques of the public are based on vague or erroneous assumptions about a key relationship – the relationship between how survey respondents answer political knowledge questions and these same respondents’ abilities to accomplish politically relevant tasks (by which we mean the ability to make the choice one would have made if knowledgeable about a set of relevant facts). Indeed, many scholars simply presumed that survey-based political knowledge measures can be treated as valid representations of citizens’ general knowledge of politics. They also presumed that not offering correct responses to these survey questions was equivalent to incompetence at politically relevant tasks such as voting. Given how often this literature criticized citizens for what they did not know, it is ironic that its authors gave so little thought to the conditions under which these presumptions were true.

Experimental political science has added clarity, precision, and new insight into such matters. Unlike non-experimental studies, experiments allow scholars to 1) randomly assign subjects to treatment and control groups, 2) systematically manipulate relevant aspects of survey-interview and decision-making environments, and 3) directly observe the answers subjects give (and the choices they make) under different conditions. These features of experiments have enabled scholars to clarify how particular aspects of the survey production process contribute to citizens’ poor performance on survey-based political knowledge questions. Experiments have also clarified the relationship between the ability to answer certain questions and the ability to make important decisions competently. To be sure, there are many things about politics that citizens do not know. But experiments show that what citizens actually know about politics, and how such knowledge affects their choices, is very different than the conventional wisdom alleges.

In this chapter, we report on two kinds of experiments that clarify what voters know and why it matters. In section 1, we address the question “What is political knowledge?” by describing experiments that manipulate the survey context from which most political knowledge measures are derived. These experiments reveal that existing knowledge measures are significantly affected by question wording, variations in respondents’ incentives to think before they answer, whether respondents feel threatened by unusual aspects of survey interview contexts, and personality variations that make some respondents unwilling to give correct answers to survey interviewers even when they know the answers.

In section 2, we describe experiments that clarify the relationship between what citizens know and their competence. These experiments compare the choices that people make given different kinds of information. They show when people can (and cannot) make competent decisions despite lacking answers to fact-based political knowledge questions. Collectively, these clarifications provide a different view of citizen competence than is found in most non-experimental work on political knowledge. There are many cases for which not knowing the answers to survey-based political knowledge questions reveals very little about citizens’ competence.

1. Experiments on Properties of Survey-Based Knowledge Measures

Many citizens do not correctly answer political knowledge questions on traditional surveys. This claim is not controversial. Less clear is what these questions tell us about citizens’ knowledge more generally.

Non-experimental studies have attempted to show that responses to such questions constitute valid measures of political knowledge in two ways. One way is to

correlate answers to fact-based questions with factors such as interest in politics and turnout. The underlying assumption is that because people who turn out to vote or are interested in politics also are more likely to answer the survey questions correctly, the questions are valid measures of knowledge. This approach is problematic because these underlying factors are not themselves credible measures of political knowledge. For example, a person can be interested in politics without being knowledgeable and can be knowledgeable without being particularly interested.

Another attempted means of validating survey responses as knowledge measures is to use factor analysis. Here, the claim is that if the same kinds of people answer the same kinds of questions correctly, then the questions must be effectively tapping a more general knowledge domain. The error in this claim can be seen by recalling previous critiques of the use of factor analysis in intelligence estimation. As Gould describes (1996, 48) “the key error of factor analysis lies in reification, or the conversion of abstractions into putative real entities.” In other words, factor analysis yields meaningful results only if the selection of questions themselves is derived from a credible theory of information and choice. Lupia (2006, 224) finds that the selection of specific questions is almost entirely subjective, reflecting the often idiosyncratic tastes of the authors involved. Hence, modern factor analytic claims are of little relevance to the question of whether survey-based political knowledge questions capture citizens’ true knowledge about politics.

Gibson and Caldeira (2009) offer an experiment that questions the validity of existing data and also provides a more effective means of measuring what citizens know.

They argue (2009, 429) that “much of what we know – or think we know – about public knowledge of law and courts is based upon flawed measures and procedures.”

In particular, many of the most famous survey-based political knowledge measures come from open-ended recall questions (i.e., questions where respondents answer in their own words rather than choosing from a small set of answers). An example of such a question, from the American National Election Studies (ANES), is as follows:

“Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like.... What about ... William Rehnquist – What job or political office does he NOW hold?”

From the 1980s through 2004, the ANES hired coders to code transcribed versions of respondents’ verbatim answers as simply “correct” or “incorrect.” The codes, and not the original responses, were included in the public ANES dataset. For decades, scholars treated the codes as valid measures of respondents’ knowledge. While many analysts used this data to proclaim voter ignorance and incompetence, an irony is that almost no one questioned whether the coding procedure was itself valid.

Gibson and Caldeira changed that. They raised important questions about which verbatim responses should be counted as correct. In 2004, for example, William Rehnquist was Chief Justice of the United States. Upon inspecting the transcribed versions of ANES responses, Gibson and Caldeira found many problems with the coding. For example, the ANES counted as correct only responses that included “Chief Justice” and “Supreme Court.” A respondent who said that Rehnquist is on the Supreme Court without saying Chief Justice or a respondent who simply said that he was a federal judge were coded as incorrect.

Gibson and Caldeira's examination of the transcripts often revealed at least partial knowledge of the topic in answers that had been coded as incorrect. Gibson and Caldeira argued (2009, 429) that past practices likely produced “a serious and substantial underestimation of the extent to which ordinary people know about the nation’s highest court.”

To assess the extent of this underestimation, they embedded an experiment in a nationally drawn telephone survey of 259 respondents. Respondents were asked to identify the current or most recent political office held by William Rehnquist, John G. Roberts, and Bill Frist. A control group was asked these questions in the traditional (open-ended) format. A treatment group was asked to identify the same individuals in a multiple choice format.

With respect to Chief Justice Rehnquist, and using the traditional ANES method of scoring open-ended responses as correct or incorrect, 12 percent of respondents correctly identified Rehnquist as Chief Justice. Another 30 percent identified him as a Supreme Court justice, but because these responses did not explicitly refer to him as Chief Justice, the ANES measure would have counted these responses as incorrect. The treatment group, by contrast, was asked to state whether Rehnquist, Lewis F. Powell, or Byron R. White was Chief Justice (with the order of the response options randomized across respondents). When asked the question in this format, 71 percent correctly selected Rehnquist. Gibson and Caldeira observed comparable results for Bill Frist and John Roberts.

The substantive impact of Gibson and Caldeira’s (2009, 430) findings is that “[T]he American people know orders of magnitude more about their Supreme Court than

most other studies have documented.” The broader methodological implication is that the combination of open-ended questions, the ANES’ coding scheme, and a lack of fact-checking by critics of citizen knowledge who used the ANES data contributed to an overly negative image of the public. Gibson and Caldeira’s work subsequently caused the ANES to restructure how it solicits and codes political knowledge (Krosnick et al. 2008). Hence, in this case, an experimental design not only influenced our understanding of political knowledge, but also improved how the concept is now measured.

Other experiments on survey interview attributes suggest further trouble for conventional interpretations of traditional political knowledge measures. Prior and Lupia (2008, 169) ask whether “seemingly arbitrary features of survey interviews” affect the validity of knowledge measures. They contend that the typical survey-based political knowledge assessment occurs in an unusual circumstance. Interviewers have incentives to complete interviews quickly. Respondents often do not want to prolong the interview. Questions are asked, and answers are expected, in quick succession. Moreover, political knowledge questions typically appear in the survey with no advance notice. And the typical survey provides no incentive for respondents to answer the questions correctly.

This “pop quiz” atmosphere is very different than circumstances in which having particular kinds of political knowledge matters most, such as elections. Election dates are typically known in advance. Hence, people who wish to become informed have an opportunity to do so before they cast a vote.

To determine the extent to which odd survey interview attributes contribute to poor performance on political knowledge quizzes, Prior and Lupia assigned over 1200 randomly selected members of Knowledge Networks’ national Internet panel to one of

four experimental groups. The control group was asked fourteen political knowledge questions in a typical survey interview environment, with little time to answer the questions (sixty seconds from the moment that the question first appeared on screen) and no motivation to answer correctly. Treatment groups received greater opportunity and/or incentive to engage the questions. One treatment group was offered one dollar for every question answered correctly. Another group was offered twenty-four hours to respond to the fourteen questions. The third treatment group was offered time and money.

Even though Prior and Lupia's questions were selected to be quite difficult, each of the treatments produced a significant increase in questions answered correctly. Compared to the control group, simply offering a dollar for correct answers increased the average number of correct answers by 11 percent. Offering extra time produced an 18 percent increase over the control group. Time and money together increased the average number of questions answered correctly by 24 percent relative to the control group.

The effect of money alone is noteworthy as the only difference between the control and treatment groups is that the latter is paid a small amount for each correct answer. Treatment group respondents did not have time to look up correct answers. Hence, the treatment group's performance gain indicates that low respondent motivation is a determinant of existing political knowledge measures.

Looking at experimental effects across population groups reinforces the conclusion. The largest effects are on respondents who report that they follow politics "some of the time" (rather than "most of the time" or "not at all"). For them, simply paying a dollar yields a 32 percent increase in correct answers relative to members of the control group who report following politics "some of the time." Hence, for people whose

attention to politics is infrequent, the typical survey interview context provides insufficient motivation for searching the true content of their memories. This finding implies that “conventional knowledge measures confound respondents’ recall of political facts with variation in their motivation to exert effort during survey interviews [and, hence,] provide unreliable assessments of what many citizens know when they make political decisions” (Prior and Lupia 2008, 169).

Other experiments examine how social roles and survey contexts interact to affect respondent performance. McClone, Aronson, and Kobrynowicz (2006) noted that men tend to score better than women on survey-based political knowledge tests. Conventional explanations for this asymmetry included the notion that men are more interested in politics.

McClone et al. saw “stereotype threat” (Steele and Aronson 1995) as an alternative explanation. In a typical stereotype threat experiment, members of stigmatized and non-stigmatized groups are randomly assigned to experimental groups. A treatment group is given a test along with a cue suggesting that members of their stigmatized group have not performed well in the past. The control group receives the test without the cue.

McClone et al.’s phone-based experiment was conducted on 141 undergraduates, 70 men and 71 women. A ten-question index measured political knowledge. Questions were drawn from sources such as the ANES. Stereotype threat was manipulated in two ways. The first manipulation was interviewer gender. Respondents were randomly assigned to be interviewed by men or women. The second manipulation pertained to the “threat cue.” Treatment groups were told that “the survey you are participating in this

evening has been shown to produce gender differences in previous research.” Control groups were not told of any gender differences.

They found that men scored higher than women overall, which is consistent with non-experimental findings. However, the experimental variations affected the inequality. When there was no threat cue, or when women were interviewed by other women, there was no significant difference between men’s and women’s scores. When the threat cue or male interviewers were introduced, the asymmetry emerged -- *but neither factor affected men’s scores*. The effects were confined to women and decreased their scores. This experiment suggests important limits on the extent to which survey-based political knowledge tests can be considered valid measures of what the population as a whole knows about politics.

Other experiments show that whether survey questions allow or encourage “don’t know” responses affects political knowledge measures. These experiments suggest that “don’t know” options may cause scholars to underestimate what the public knows about politics. The reason is that some people are less likely than others to offer answers when they are uncertain.

For example, Jeffrey Mondak and his colleagues designed several split-ballot experiments (i.e., random assignment of survey respondents to experimental conditions) in two surveys, the 1998 ANES Pilot Study and a Tallahassee-based survey. In each survey, the control group received knowledge questions that began with a “don’t know”-inducing prompt, “Do you happen to know...” and interviewers were instructed not to probe further after an initial “don’t know” response. Treatment groups received questions with identical substantive content, but different implementation. In both surveys,

treatment groups heard a guess-inducing phrase such as “even if you're not sure I'd like you to tell me your best guess.” In the ANES version, moreover, the interviewer first recorded any “don't know” responses and then probed further for substantive answers to determine whether respondents who initially responded “don't know” actually knew about the concept in question.

In each experiment, respondents were significantly less likely to choose “don't know” when they were encouraged to guess. Interviewer probing decreased “don't knows” even further (Mondak 2001). Moreover, women were significantly more likely than men to respond “don't know” even when encouraged to guess (Mondak and Anderson 2004). In this analysis, discouraging “don't knows” reduced the extent to which men outperform women (in terms of questions answered correctly) by about half.

Such experiments also show that many respondents chose “don't know” for reasons other than ignorance. Mondak and Davis (2001) analyzed the responses offered by respondents who initially claimed not to know the answer. These responses were significantly more likely to be correct than responses that would have emerged from blind guessing. Taken together, Mondak and his colleagues show that many previous political knowledge measures confound what respondents know with how willing they are to answer questions when they are uncertain of themselves.

Building on these findings, Miller and Orr (2008) designed an experiment where the “don't know” option was not just discouraged, it was eliminated altogether. It was run on 965 undergraduates via Internet surveys. Each respondent received eight multiple-choice political knowledge questions and each question contained three substantive response options. What differed across their experimental groups was question format.

The first group's questions encouraged "don't know." The second group's questions discouraged "don't know." For the third group, the "don't know" option was simply unavailable.

Miller and Orr found that discouraging "don't know" (rather than encouraging it) led to a substantial drop in the use of the "don't know" option. They also found that discouraging "don't know" (rather than encouraging it) corresponded to an increase in the average percentage of correct answers given per respondent.

The most interesting thing about the comparison between the "don't know"-encouraged and "don't know"-omitted groups is that the increase in percent correct (from 61 percent to 70 percent) was higher than the increase in percent incorrect (from 21 percent to 29 percent). This is interesting because each question had three response options. Hence, if the "don't know"-encouraged and "don't know"-omitted groups were equivalent, and if all that omitting "don't know" options does is cause respondents to guess haphazardly, then respondents who would have otherwise chosen "don't know" should have only a one-in-three chance of answering correctly. Hence, if respondents were simply guessing, the increase in average percent incorrect should be roughly double the increase in average percent correct. Instead, the increase in corrects was larger than the increase in incorrects. Miller and Orr's experiment shows that the "don't know" option attracts not just respondents who lack knowledge about the questions' substance but also people who possess relevant information but are reticent to respond for other reasons (such as lack of confidence or risk aversion).

While Miller and Orr's work suggests that "don't know" responses hide partial knowledge, research by Sturgis, Allum, and Smith (2008) suggests a different conclusion.

They integrated a split ballot experiment into a British telephone survey. Each respondent was asked three knowledge-related questions. Each question contained a statement and the respondent was asked to say whether it was true or false. One thousand and six respondents were randomly assigned to one of three conditions. In one condition, the question's preamble included the “don't know”-encouraging phrase, “If you don't know, just say so and we will skip to the next one.” In a second condition, that phrase was substituted with the “don't know”-discouraging phrase, “If you don't know, please give me your best guess.” In the third condition, the original “don't know”-encouraging statement was included but the response options were changed. Instead of simply saying true or false, respondents in this group could say whether the statement was “probably true,” “definitely true,” “probably false,” or “definitely false.” Moreover, in the first and third conditions, respondents who initially said “don't know” were later asked to provide their best guess.

Sturgis et al. find that discouraging “don't know” responses significantly decreased their frequency (from 33 percent to 9 percent). Providing the “definitely” and “probably” response options also reduced “don't know” responses (to 23 percent). Turning their attention to partial knowledge, they then analyze the answers given by respondents who initially responded “don't know” and then chose true or false after interviewer probing. For two of the three questions, probing elicited correct answers at a rate no better than chance. For the other question, two-thirds of the new responses elicited were correct. From these results, they conclude that “when people who initially select a ‘don't know’ alternative are subsequently asked to provide a ‘best guess,’ they fare statistically no better than chance.” But their results suggest a different interpretation –

that there are types of people and question content for which encouraging “don’t knows” represses partial knowledge and that there is still much to learn about the types of questions and people that make such repressions more or less likely. Further, with only three true/false questions, their experiment does not provide a sufficient basis for privileging their conclusions over those of Miller and Orr. As Miller and Orr (2008, 779) note, “The availability of three options from which to choose [the Miller-Orr method] may motivate respondents to draw on their partial knowledge, whereas the true/false format might not.”

In sum, many scholars and analysts use fact-based survey questions to draw broad conclusions about public ignorance. Experiments on survey-based political knowledge measures have shown that many underappreciated attributes of survey interview contexts (including question wording, respondent incentives, and personality variations) are significant determinants of past outcomes. Hence, as a general matter, survey-based political knowledge measures are much less valid indicators of what citizens know about politics than many critics previously claimed.

2. When Do Citizens Need to Know the Facts?

In addition to shedding light on the validity of survey-based political knowledge measures, experiments also clarify the conditions under which knowing particular facts about politics is necessary, sufficient, or relevant to a citizen’s ability to make competent choices. Non-experimental research on this topic has often tried to characterize the relationship between political knowledge and politically relevant choices with brief anecdotes or with regression coefficients that presume a simple linear and unconditional relationship between the factors. These characterizations are never derived from direct

evidence or rigorous theory. For example, many non-experimental critics of voter competence have merely presumed that any fact they deem worth knowing must also be a necessary condition for others to make competent political decisions.

With experiments, scholars can evaluate hypotheses about relationships between knowledge of specific facts and competence at various politically relevant tasks. To illustrate how and why experiments are well suited for this purpose, we discuss several examples. These examples reveal conditions under which people who lack certain kinds of information do, and do not, make competent choices nevertheless.

Lupia and McCubbins (1998) acknowledge that many citizens lack factual knowledge about politics, but they emphasize that uninformed citizens may be able to learn what they need to know from political parties, interest groups, and the like (i.e., speakers). Lupia and McCubbins use a formal model to identify conditions under which citizens can make competent choices as a result of such interactions.

They use experiments to evaluate key theoretical conclusions. In some of these experiments, subjects guess the outcomes of unseen coin tosses. Coin tosses are used because they, like many elections, confront people with binary choices. Coin tosses are also easy to describe to experimental subjects, which makes many interesting variations of a coin-toss guessing game easier to explain.

Before subjects make predictions, another subject (acting as “the speaker”) makes a statement about whether the coin landed on heads or tails. Subjects are told that the speaker is under no obligation to reveal what he or she knows about the coin truthfully.

After receiving the speaker’s statement, subjects predict the coin toss outcome.

To evaluate key theoretical hypotheses, Lupia and McCubbins systematically manipulate multiple aspects of the informational context, including whether speakers get paid when subjects make correct (or incorrect) predictions. They also vary other factors that cause lying to be costly or increase the probability that false statements will be revealed (i.e., verification). In other words, to evaluate conditions under which subjects can learn enough to make correct predictions, Lupia and McCubbins vary attributes of the speaker and the context in which subjects predict coin toss outcomes.

Lupia and McCubbins' experiments show that under the detailed sets of conditions identified by their theory, subjects almost always make correct predictions. Specifically, when subjects perceive a speaker as being knowledgeable and having common interests (i.e., as benefiting when subjects predict correctly), subjects trust the speaker's statements. When they are in conditions under which such perceptions are likely to be correct, these subjects make correct predictions at a very high rate – one that is substantially greater than chance and often indistinguishable from the predictions they would have made if they knew the coin toss outcome in advance. Similarly, when a sufficiently large penalty for lying or probability of verification is imposed upon the speaker, subjects trust the speaker's statements and make correct predictions at very high rates.

These experiments highlight conditions under which uninformed citizens can increase their competence by learning from others. But the experiments evaluate only a few of Lupia and McCubbins' theoretical implications. One question that their experiments leave open is whether a speaker's statements are equally helpful to more and less knowledgeable citizens. On one hand, it is possible that a speaker's statements will be

more helpful to citizens who already know a lot about the choices they face. On the other hand, a speaker's statements may be more helpful to people who know less.

Boudreau (2009) replicates Lupia and McCubbins' experiments but substitutes math problems for coin tosses. An advantage of using math problems is that subjects vary in their levels of preexisting knowledge. Some subjects know a lot about how to solve math problems. Others do not. A second advantage is that there exists a valid, reliable, and agreed-upon measure of how knowledgeable subjects are about this type of decision – SAT math scores. Thus, Boudreau collects subjects' SAT math scores prior to the experiments. She uses the experiments to clarify conditions under which a speaker's statements about the answers to the math problems help low-SAT subjects perform as well as high-SAT subjects.

Boudreau finds that when the speaker is paid for subject success, is subject to a sufficiently large penalty for lying, or faces a sufficiently high probability of verification, both low-SAT and high-SAT subjects achieve large improvements in their decisions (relative to counterparts in the control group, who do the problems without a speaker). Low-SAT subjects improve so much that it reduces the achievement gap between them and high-SAT subjects. This gap closes even when the size of the penalty for lying or probability of verification is reduced (and is, thus, made more realistic because the speaker may have an incentive to lie). This result occurs because high-SAT subjects do not improve their decisions (and apparently ignore the speaker's statements), but low-SAT subjects typically improve their decisions enough to make them comparable to those of high-SAT subjects. By using an experimental task for which subjects vary in their

levels of knowledge, Boudreau further clarifies the conditions under which less informed citizens can make competent choices.

One of the strengths of Boudreau's and Lupia and McCubbins' experiments is that subjects make decisions for which there is an objectively correct or incorrect choice under different conditions. This approach is advantageous because it allows them to measure precisely whether and when a speaker's statements help subjects make a greater number of correct decisions than they would have made on their own. However, what does it mean for citizens to make correct decisions in electoral contexts? Lau and Redlawsk (1997, 2001) address this question by conducting experiments in which subjects learn about and vote for candidates in mock primary and general elections. Subjects in Lau and Redlawsk's experiments are provided with different types of information about fictional candidates. Subjects access this information by clicking on labels (such as "Walker's stand on defense spending") that appear on computer screens. Subjects can also learn about a candidate's partisanship, ideology, and appearance, as well as endorsements and polls. After subjects gather information about the primary election candidates they vote for one of these candidates. Subjects repeat this process for the general election candidates. At the completion of the experiments, subjects receive all of the information that was available for two candidates from the primary election (not just the information they clicked on during the experiment). Subjects are then asked whether they would have voted for the same candidate if they had all of this information when they made their decisions.

Lau and Redlawsk (1997) find that subjects, in the aggregate, are adept at voting correctly. According to one of their measures, approximately 70 percent of subjects voted

correctly. However, Lau and Redlawsk (2001) identified conditions that hinder subjects' ability to vote correctly. For example, they find that although heuristics significantly increase the ability to vote correctly among subjects who score high on their political knowledge and political interest index, they decrease less knowledgeable and less interested subjects' ability to vote correctly.ⁱⁱⁱ Lau and Redlawsk also find that characteristics of the information environment limit subjects' ability to vote correctly. Specifically, subjects are less likely to vote correctly when the number of primary candidates increases from two to four and when the choice between the candidates is more difficult (i.e., the candidates are more similar). Thus, although Lau and Redlawsk observe high levels of correct voting in the aggregate, they also show that there are conditions under which aspects of the information environment have detrimental effects on subjects' ability to make correct decisions.

Continuing Lau and Redlawsk's emphasis on the political environment, Kuklinski et al. (2001) use experiments to assess the effects that other aspects of the environment have on citizens' decisions. In contrast to Lau and Redlawsk's focus on correct voting, Kuklinski et al. contend that the ability to make tradeoffs is fundamental to being a competent citizen. They conduct survey experiments in which they measure subjects' ability to make tradeoffs among competing goals for health care reform.

Subjects in Kuklinski et al.'s experiments view seven different health care goals (e.g., universal coverage, no increase in taxes, uniform quality of care), and they rate on a scale of one to ten how much of each goal a health care plan must achieve for them to consider the plan acceptable. The key to this experiment is that the health care goals conflict with one another; that is, no health care plan can realistically achieve all of the

goals. In various treatment groups, Kuklinski et al. manipulate the conditions under which subjects rate the seven health care goals. In one group, subjects are given general information about the need for tradeoffs when designing any program. In a second group, subjects are given motivational instructions that encourage them to take their decisions seriously. In a third group, subjects are given both general information and motivational instructions. In a fourth group, subjects are given diagnostic information about the exact tradeoffs involved in health care reform (e.g., that we cannot provide health coverage for everyone and simultaneously keep taxes low). In a fifth group, this diagnostic information is provided along with motivational instructions. Kuklinski et al. then observe whether and under what conditions information and/or motivation improves subjects' ability to make tradeoffs (measured as the extent to which subjects reduce their demands for conflicting goals), relative to subjects in the control group who do not receive any information or motivational instructions.

Kuklinski et al.'s experiment reveals conditions under which new information improves subjects' ability to make tradeoffs and eliminates differences between more and less knowledgeable subjects. Specifically, where control group subjects tended not to make tradeoffs, treatment group subjects were much more likely to do so when both general information and motivational instructions were provided. Kuklinski et al. also show that diagnostic information about health care tradeoffs induces subjects to make these tradeoffs, regardless of whether they are motivated to do so and regardless of their knowledge level. Indeed, when diagnostic information is provided, less knowledgeable subjects are just as capable of making tradeoffs as more knowledgeable subjects. In this way, Kuklinski et al. demonstrate that when information in the environment is

sufficiently diagnostic, it can substitute for preexisting knowledge about politics and eliminate differences between more and less knowledgeable subjects.

Other experiments clarify the effect of policy-specific knowledge on citizens' abilities to express their opinions. Using survey experiments, Gilens (2001) suggests that policy-specific knowledge may be more relevant than more general conceptions of political knowledge. Gilens randomly determines whether subjects receive specific information about two policy issues (crime and foreign aid). Subjects in the treatment group receive information about two news stories, one showing that the crime rate in America has decreased for the seventh year in a row and one showing that the amount of money spent on foreign aid has decreased and is now less than one cent of every dollar that the American government spends. Control group subjects simply learn that two news stories have been released, one pertaining to a government report about the crime rate and one pertaining to a report about American foreign aid. Gilens then asks all subjects about their level of support for government spending on prison construction and federal spending on foreign aid.

Gilens's results demonstrate that the provision of policy-specific information significantly influences subjects' opinions. Specifically, treatment group subjects (who learn that the crime rate and foreign aid spending have decreased) are much less likely than the control group to support increasing government spending on prison construction and decreasing American spending on foreign aid. Gilens shows that policy-specific information has a stronger influence on subjects who possess high levels of general political knowledge. Gilens suggests that citizens' ignorance of policy-specific facts (and

not a lack of general political knowledge) is what hinders them from expressing their opinions effectively on certain policy issues.

Kuklinski et al. (2000) also assess whether policy-specific facts are relevant to citizen opinions. In contrast to Gilens, Kuklinski et al. distinguish between citizens who are uninformed (i.e., who lack information about particular policies) and citizens who are misinformed (i.e., who hold incorrect beliefs about particular policies). Indeed, Kuklinski et al. suggest that the problem facing our democracy is not that citizens are uninformed, but rather that citizens confidently hold incorrect beliefs and base their opinions upon them.

Kuklinski et al. assess experimentally whether and when the provision of correct policy-specific information induces citizens to abandon incorrect beliefs and express different opinions. In one treatment group, subjects receive six facts about welfare (e.g., the percentage of families on welfare, the percentage of the federal budget devoted to welfare) before they express their opinions. In another treatment group, subjects first take a multiple-choice quiz on these six facts about welfare. After answering each quiz question, subjects in this treatment group are asked how confident they are of their answer. The purpose of the quiz and the follow-up confidence questions is to gauge subjects' beliefs about welfare and how confidently they hold them. In the control group, subjects simply express their opinions about welfare policy.

Kuklinski et al. also conduct follow-up experiments in which the information about welfare is made more salient and meaningful than the six facts provided in the initial experiments. To this end, Kuklinski et al. ask subjects what percentage of the federal budget they think is spent on welfare and what percentage of the budget they

think should be spent on welfare. Immediately after answering these two questions, subjects in the treatment group are told the correct fact, which for most subjects is that actual welfare spending is lower than either their estimate or their stated preference. Control group subjects answer these two questions but do not receive the correct fact about actual welfare spending. At the completion of these experiments, subjects in both groups express their level of support for welfare spending.

In both experiments, Kuklinski et al. find that subjects are grossly misinformed about welfare policy. Indeed, the percentage of subjects who answer particular multiple-choice quiz questions incorrectly ranges from 67 percent to 90 percent. Even more troubling is their finding that subjects who have the least accurate beliefs are the most confident in them. Kuklinski et al. also show that the opinions of subjects who receive the six facts about welfare are no different than the opinions of subjects in the control group, which indicates that treatment group subjects either did not absorb these facts or failed to change their opinions in light of them. However, when a fact about welfare is presented in a way that explicitly exposes and corrects subjects' incorrect beliefs (as in the follow-up experiments), subjects adjust their opinions about welfare spending accordingly. In this way, Kuklinski et al. demonstrate that policy-specific information can induce misinformed citizens to abandon their incorrect beliefs and express informed opinions, but only when it is presented in a way that is meaningful and relevant to them.

Taken together, experiments have shown conditions under which knowledge of the kinds of facts that have been the basis of previous political knowledge tests are neither necessary nor sufficient for competence at subsequent tasks. Some of these conditions pertain to the kind of information available. Other conditions pertain to the

context in which the information is delivered. Collectively, these clarifications provide a different view of citizen competence than is found in most non-experimental work on political knowledge. While there are many cases in which lacking information reduces citizens' competence, experiments clarify important conditions under which things are different. In particular, if there are relatively few options from which to choose (as is true in many elections), if people can be motivated to pay attention to new information, if the information is highly relevant to making a competent decision, and if people's prior knowledge of the topic, the speaker, or even the context leads them to make effective decisions about whom and what to believe, then even people who appear to lack political knowledge as conventionally defined can vote with the same level of competence as they would if better informed.

3. Conclusions

Our argument in this chapter has been as follows. First, there are many ways in which the survey questions that scholars use to measure political knowledge are not valid indicators of what citizens know about politics. Second, for circumstances in which such measures are valid, it is not clear that this knowledge is necessary, sufficient, or relevant to a citizen's ability to perform important democratic tasks, such as voting competently. Third, experiments have shed important new light on both of these issues. Collectively, experiments are not only helping scholars more effectively interpret existing data, they are also helping scholars develop better knowledge measures. In a very short period of time, experiments have transformed the study of political knowledge.

Yet, experiments have just begun to scratch the surface of the multifaceted ways in which thought and choice interact. They have examined only a few of the many

attributes of survey interviews that can affect responses. They have also examined only a few of the many ways that particular kinds of political knowledge can affect politically relevant decisions and opinions. Their different results have also raised questions about whether and when particular types of information eliminate differences between more and less knowledgeable citizens (compare Lau and Redlawsk [2001] to Kuklinski et al. [2001] and Boudreau [2009]). As research in fields of study such as political cognition, the psychology of the survey response, and political communication evolve, more questions will be raised about the validity of extant knowledge measures and whether particular kinds of knowledge are relevant to democratic outcomes. While such questions can be studied in many ways, experiments should take center stage in future research. Hence, the experiments we have described represent the beginning, rather than the end, of a new attempt to better understand what people know about politics and why it matters.

References

- Boudreau, Cheryl. 2009. "Closing the Gap: When Do Cues Eliminate Differences between Sophisticated and Unsophisticated Citizens?" *Journal of Politics* 71: 964-76.
- Gibson, James L., and Gregory A. Caldeira. 2009. "Knowing the Supreme Court? A Reconsideration of Public Ignorance of the High Court." *Journal of Politics* 71: 429-41.
- Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95: 379-96.
- Gould, Stephen J. 1996. *The Mismeasure of Man. Revised Edition*. New York: W.W. Norton and Co.
- Graber, Doris. 1984. *Processing the News: How People Tame the Information Tide*. New York: Longman.
- Krosnick, Jon A., Arthur Lupia, Matthew DeBell, and Darrell Donakowski. 2008. "Problems with ANES Questions Measuring Political Knowledge." Retrieved

from

www.electionstudies.org/announce/newsltr/20080324PoliticalKnowledgeMemo.pdf

- Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, and Robert F. Rich. 2001. "The Political Environment and Citizen Competence." *American Journal of Political Science* 45: 410-24.
- Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *Journal of Politics* 62: 790-816.
- Lau, Richard R. and David P. Redlawsk. 1997. "Voting Correctly." *American Political Science Review* 91: 585-98.
- Lau, Richard R. and David P. Redlawsk. 2001. "Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making." *American Journal of Political Science* 45: 951-71.
- Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* New York: Cambridge University Press.
- Lupia, Arthur. 2006. "How Elitism Undermines the Study of Voter Competence." *Critical Review* 18: 217-32.
- McClone, Matthew S., Joshua Aronson, and Diane Kobrynowicz. 2006. "Stereotype Threat and the Gender Gap in Political Knowledge." *Psychology of Women Quarterly* 30: 392-8.
- Miller, Melissa, and Shannon K. Orr. 2008. "Experimenting with a "Third Way" in Political Knowledge Estimation." *Public Opinion Quarterly* 72: 768-80.
- Mondak, Jeffery J. 2001. "Developing Valid Knowledge Scales." *American Journal of Political Science* 45: 224-38.
- Mondak, Jeffery J., and Belinda Creel Davis. 2001. "Asked and Answered: Knowledge Levels When We Will Not Take 'Don't know' for an Answer." *Political Behavior* 23: 199-224
- Mondak, Jeffery J., and Mary R. Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-based Differences in Political Knowledge." *Journal of Politics* 66: 492-513.
- Popkin, Samuel L. 1994. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

- Prior, Markus, and Arthur Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall from Political Learning Skills." *American Journal of Political Science* 52: 168-82.
- Steele, Claude M., and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African-Americans." *Journal of Personality and Social Psychology* 69: 787-811.
- Sturgis, Patrick, Nick Allum, and Patten Smith. 2008. "An Experiment on the Measurement of Political Knowledge in Surveys." *Public Opinion Quarterly* 72: 90-102.

ⁱ We thank Adam Seth Levine, Yanna Krupnikov, James Kuklinski, Nicholas Valentino, and John Bullock for helpful comments.

ⁱⁱ The Oxford English Dictionary defines "knowledge" as "(i) expertise, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject; (ii) what is known in a particular field or in total; facts and information; or (iii) awareness or familiarity gained by experience of a fact or situation." In this definition, information is required for knowledge, but it is not the same as knowledge. We (the authors) regard what is measured by responses to the survey questions referenced in this chapter as "political information." Whether or not such responses constitute a "political knowledge" measure first requires that we ask the question "knowledge of *what*?" Scholars working in this area rarely ask this question and simply confound information and knowledge. While answers to fact-based survey questions can provide evidence of a citizen's knowledge of the *specific facts* mentioned in *specific questions*, the extent to which such responses provide reliable evidence of broader kinds of political knowledge is typically limited (Lupia 2006). These limitations are often due to the narrow focus and small number of such questions that appear in any given survey. That said, given that the literature is widely known as referring to all such data as evidence of "political knowledge" and given that we wish to focus our space in this handbook on how experimental work has advanced that very literature, we follow the flawed naming convention to minimize confusion.

ⁱⁱⁱ The index combines subjects' levels of political knowledge, political behavior, political interest, political discussion, and media use.

IV. Vote Choice, Candidate Evaluations, and Turnout

13. Candidate Impressions and Evaluations

Kathleen M. McGrawⁱ

In order for citizens to responsibly exercise one of their primary democratic duties, namely voting, two preliminary psychological processes must occur. First, citizens must learn something about the candidates; that is, come to some understanding, even if amorphous, about the candidates' characteristics and priorities. Second, citizens must reach a summary judgment about the candidates. There is a long and distinguished history of scholarly studies of the linked processes of voting, perceptions of candidates, and evaluations of them. In his recent review of the voting behavior literature, Bartels (2010) notes, "The apparent failure of causal modeling [of observational data] to answer fundamental questions about voting behavior produced a variety of disparate reactions," including scholars turning to experimentation to better understand these basic processes. My goal in this chapter is to outline experimental workⁱ that has contributed to our understanding of citizens' impressions and evaluations of political candidates, and to identify questions that future experiments might answer.

1. Clarification of basic concepts

Two sets of conceptual distinctions should be made clear at the outset. First, the chapter title distinguishes between candidate *impressions* and *evaluations*. By "impressions" I mean an individual's mental representation – the cognitive structure stored in memory – consisting of knowledge and beliefs about another person.ⁱⁱ "Evaluation", on the other hand, refers to a summary global judgment ranging from very negative to very positive. Depending on the underlying cognitive processes that

contribute to the formation of the evaluation (i.e., online or memory-based), it may or may not be part of the cognitive representation.

Second, I started this chapter by linking candidate impressions and evaluations to voting choices, and it is undoubtedly the case that those processes are connected. There is an understandable tendency among researchers and readers to assume that processes of political impression formation and evaluation are tantamount to voting choices, but it is an error to do so. Impression formation and evaluation are concerned with single targets, whereas a vote choice, like any choice, requires selecting from a set of two or more alternatives. Whereas most observational and experimental studies in this research tradition are situated in a campaign context, it is worth noting that impressions and evaluations of individuals who eventually become candidates for office very often begin prior to the electoral context. For example, many Americans had developed impressions and evaluations of Arnold Schwarzenegger long before he decided to run for office (and most of those Americans never had, and never will have, the opportunity to decide whether or not to vote for him). The same holds true for Hillary Rodham Clinton. I do not believe these are unusual examples. Many local, state, and national political leaders are known from other areas of life before entering the electoral arena. Moreover, citizens continue to update their impressions and evaluations once candidates become elected officials, and are no longer seeking office. It is unfortunate that research hasn't captured these pre- and post- candidate phenomena.

Moreover, it is a mistake to assume, even when a vote choice is the clear endpoint of candidate evaluation, that the underlying processes are equivalent. Behavioral decision theorists have long recognized that “*choosing* one alternative from a set can invoke different

psychological processes than *judging* alternatives” (Johnson and Russo 1984, p. 549). Lau and Redlawsk (2006) have provided the clearest discussion of this disjunction in the political science literature. Global evaluations of individual candidates need to be translated into a decision about *how* to vote, and maybe even a decision about *whether* to vote. We still know very little about how the processes of impression formation, evaluation, and vote choice are linked. Consequently, my focus in the remainder of this chapter is on impressions and evaluations, rather than voting.

2. Shortcomings of Observational Studies

While the many observational studies on candidate impressions and evaluation have provided valuable insights, they also suffer from some limitations. Two bear mentioning. The first is the ability to draw strong inferences about causal determinants. It is widely understood that variables such as the American National Election Studies (ANES) candidate like-dislike questions are contaminated by post-hoc rationalizations, and so researchers must be cautious in treating them as causes of the vote (Rahn, Krosnick, and Breuning 1994; Lodge and Steenbergen 1995). The same problem exists for virtually all of the factors presumed to be causal determinants in the observational literature (e.g., trait inferences, emotional responses, perceived issue positions, etc.). Experiments provide scholars with the opportunity to evaluate the causal impact of theoretically meaningful predictors.ⁱⁱⁱ

Second, candidates and other politicians in the real world inevitably suffer from what experimentalists refer to as confounds. In a research design, confounding occurs when a second (or more) extraneous variable perfectly covaries with the theoretical variable of interest. So, for example, the two candidates in the 2008 Presidential election (Obama and McCain) differed on

several important dimensions that may have been consequential for citizens' evaluations of them – partisan affiliation, race, age, experience, image, vice presidential picks, campaign strategies, favorability of media coverage and so on, making it difficult for analysts to specify with any degree of precision which characteristics were critical. Experimentation allows researchers to disentangle – that is, manipulate independently – candidate and situational factors to determine which have a meaningful causal impact.

3. Principal Characteristics of Experiments on Candidate Impressions and Evaluations

The experimental designs used to study candidate impressions and evaluations can be characterized along three key dimensions. The first is the static versus dynamic dimension. The vast majority of experiments in this tradition rely on a static, one-shot presentation of information about a political candidate, utilizing (more or less realistic) paper “campaign fact sheets,” videotapes, or electronic formats. Not only is the presentation of information static, but participants are typically provided with all of the information that the researcher deems relevant, rather than being responsible for seeking out information. Exposure, in other words, is held constant.

In the late 1980s, Rick Lau and David Redlawsk began to develop a dynamic process tracing methodology, where the information that is available about candidates comes and goes, and research participants are responsible for choosing the information they wish to learn (Lau and Redlawsk 2006). This represents a significant methodological development. While very different in many important respects, the static and dynamic paradigms do share two features. First, in the research conducted to date, both involve a simulated campaign that takes place in a single, short period of time (Mitchell 2008 is an exception). Second, both paradigms have for the

most part relied on hypothetical candidates. I do not know of any experimental studies involving real candidates over an extended period of real time.

A second dimension involves the research participants. While the majority of experimental studies in this tradition rely on samples of convenience rather than representative samples, some studies rely on college students while others draw from nonstudent volunteer populations. Recent analyses suggest a trend away from a heavy reliance on college student samples in experimental work more generally in political science.^{iv} The potential problems associated with college student samples, problems which often times are exaggerated, are well-known, and beyond the scope of this paper. Suffice it to say that I concur with Druckman and Kam (chapter in this volume) who claim that, “student subjects do *not* intrinsically pose a problem for a study’s external validity.” In particular, there is little theoretical reason to believe that the basic cognitive processes underlying the formation of impressions and evaluations of candidates vary across different samples. And, to the extent that theoretically meaningful differences might exist (say, for example, in terms of cognitive ability and political sophistication), experimental researchers have been open to exploring the impact of those moderators.

A third important distinction is how the partisan affiliation of the candidate is handled in the experimental design.^v Three possibilities exist. The first is to vary the partisan affiliation of the stimulus candidate so that partisanship becomes an independent variable that is fully crossed with other manipulated variables (e.g., Riggle et al. 1992). The second approach is to make explicit the partisanship of the target candidate, and hold it constant (e.g., Lodge, McGraw, and Stroh 1989). Finally, in some studies, the partisan affiliation of the candidate is left out altogether

(McGraw, Hasecke, and Conger 2003, is an example, but there are many others). Partisan attachments exert an enormous impact on citizens' impressions and evaluations of political candidates, which would suggest that partisan affiliation should be central to these experiments. However, experimentalists face tradeoffs, as all researchers do, and the first option – manipulation of candidate partisanship -- comes with a sometimes hefty cost, namely doubling the number of treatments, and so the number of participants. But the failure to manipulate partisanship (i.e., by holding it constant or ignoring it) carries risks, beyond abstract worries about external validity. First, if partisanship is not manipulated, it is impossible to ascertain whether the impact of a key manipulation holds for candidates of both parties. Second, in some instances, the presence of information about a candidate's partisan affiliation can serve to dampen, and even eliminate, the impact of other manipulated variables. For example, the Riggle et al. 1992 study found that candidate attractiveness had no impact when partisanship was available. Similarly, Stroud, Glaser, and Salovey (2006), found that emotional expressions on the part of candidates had no impact when partisan information was provided. In short, when researchers do not manipulate partisan affiliation, they risk failing to detect contingency effects or overstating the impact of other factors.

4. The Content of Candidate Impressions

As noted in the previous section, impressions are cognitive structures, consisting of what we know and believe about another person, the information we have learned, and the inferences we have drawn. Citizens' impressions of political officials are rich and multifaceted, consisting of trait inferences, knowledge about political attributes such as partisanship, beliefs about issue positions and competencies, personal characteristics and history, family, group associations, and

hobbies and personal proclivities^{vi} (Miller, Wattenberg, and Malanchuk 1986; McGraw, Fischle, and Stenner 2000). While the content of political impressions is undoubtedly wide ranging, the experimental literature has tended to focus on traits and policy positions.

Traits are a central component of ordinary and political impressions, and so they have received a tremendous amount of theoretical and empirical attention. Because traits are unobservable, they must be inferred from the observable qualities of the politician. Because behavior is often ambiguous, there is rarely an inevitable correspondence between a particular behavioral episode and the resulting trait inference. Although there are a seemingly infinite number of traits available in ordinary language, the most common traits used to characterize politicians tend to fall into a limited number of categories: competence (“intelligent,” “hard-working”), leadership (“inspiring,” “(not) weak”), integrity (“honest,” “moral”), and empathy (“compassionate,” “cares about people”). It is clear that trait inferences are consequential for evaluations of political candidates and vote choices (Funk 1996; Kinder 1998). Of the four dimensions, competence appears to be most influential, at least in terms of evaluations of presidential candidates (Markus 1982; Kinder 1986). Much of the available data are cross-sectional, raising the very real possibility that trait inferences are rationalizations, rather than causes of evaluations. However, experimental work has verified that traits play a causal role in shaping candidate evaluations (Huddy and Terkildsen 1993; Funk 1996).

Policy positions also play a prominent role in impressions and evaluations of candidates. Although the authors of *The American Voter* (Campbell et al. 1960) suggested issues are a relatively peripheral component of evaluations and vote choices, more recent work suggests a more prominent role. For example, issues that people consider important have a substantial

impact on presidential candidate evaluations (Krosnick 1988). Relatedly, experimental and observational research on media priming effects indicates that issues that are highlighted in the media have a sizable impact on evaluations of political leaders (Iyengar and Kinder 1987; Krosnick and Kinder 1990).

Having established that traits and policy positions are important components of candidate impressions and that they play a causal role in evaluations, it is important to determine their sources. It is customary to categorize citizens as flexible information processors, capable of engaging in both data-driven (individuating) and theory-driven (stereotypic) processing, in line with the theoretical predictions drawn from dual processing models (Fiske and Neuberg 1990). By individuation, I mean judgments based on the specific information that is available, without reference to stereotypic categories. It is clear from the literature that citizens' judgments about traits and policy positions are rooted to some extent in the actual behaviors manifested by candidates (see, for example, Kinder (1986) for sensible candidate-specific differences in trait perceptions, and Ansolabehere and Iyengar (1995) for evidence that citizens learn issue positions from advertising). While the content of citizens' impressions can be grounded in the information they learn about specific candidates, that information need not be accurate or objective. After all, candidates (and their opponents) structure communication strategies to manipulate these perceptions (I return to this theme in the following section).

Trait and policy inferences also result from stereotyping, a consequence of categorizing individuals into different groups or types. As a type of cognitive structure, stereotypes contain elements that can be applied by default to individual members of the group, particularly in low information settings. While many stereotypes might be activated when thinking about

candidates, political scientists have tended to focus on four: physical appearance, gender, race, and partisanship. The first three promote stereotyping because they are physical characteristics that are activated by visual cues. There is good reason to believe that stereotypes based in these categories are especially powerful, as inferences drawn from physical cues tend to be even more automatic than those drawn from verbal sources (Gilbert 1989). Gender, race and partisanship also promote stereotyping because they are politically meaningful categories that play a prominent role in American politics.

Of the four stereotype categories, gender has received the most scholarly attention, and the results from observational and experimental studies converge.^{vii} All else equal, female candidates are ascribed stereotypic feminine traits whereas male candidates are described in stereotypic masculine terms. In addition, gender-based stereotypes extend to perceived competency in various policy domains, as well as to inferences about partisanship and ideology. Huddy and Terkildsen's (1993) experiment provides the most sophisticated analysis of the complex links among candidate gender, traits, and issue competency.

Racial considerations are also consequential for a wide range of public opinion phenomena.^{viii} However, on the specific question of whether racial stereotypes (beliefs about the characteristics of members of racial groups; as opposed to the affective phenomenon of racial prejudice) have an impact on candidate perceptions and evaluations, there is surprisingly little evidence and that which exists is decidedly mixed. Different experiments have reached different conclusions about the extent to which African American candidates are inferred to possess positive and negative trait characteristics (Williams 1990; Colleau et al. 1990; Moskowitz and Stroh 1994; Sigelman et al. 1995). White Americans infer that African American candidates

support more liberal policy positions (Williams 1990; Sigelman et al. 1995; McDermott 1998), consistent with the broad racial gap in policy preferences observed in the population at large (Kinder and Sanders 1996). The extent to which these policy inferences are produced by racial or partisan stereotypes is unclear, however, because the majority of African American citizens and candidates are affiliated with the Democratic Party, and the aforementioned studies did not independently manipulate candidate partisanship.

More generally, as Hutchings and Valentino conclude, “we still are unsure why whites do not support black political candidates” (2004, 400). Several recent studies suggest promising avenues for future research. Schneider and Bos (2009) make the insightful argument that previous research on racial stereotypes and candidate inference has been misguided in the assumption that stereotypes of African American *politicians* are equivalent to stereotypes of African American *people in general* and their data on this point are compelling, with little overlap in the content of the two. Hajnal’s (2006) research suggests that the candidate’s status is important, with African American incumbents viewed more favorably than African American challengers. Berinsky and Mendelberg’s (2005) analysis of Jewish stereotypes -- in particular, the links between, and consequences of, acceptable and unacceptable stereotypes -- provides a psychologically astute framework for understanding political stereotypes and judgment more generally.

The third stereotypic category to be considered is partisanship, and here observational and experimental studies nicely converge in the domain of policy inferences: citizens hold clear and surprisingly consensual beliefs about the policy positions and issue competencies that “go with” partisan affiliation (e.g., Feldman and Conover 1983; Rahn 1993). Hayes (2005), in an

extension of Petrocik's (1996) theory of issue ownership, demonstrates that citizens also associate specific traits with candidates from the two parties (i.e., the public views Republicans as stronger leaders and more moral, while Democrats are seen as more compassionate and empathetic). Hayes' study is observational, and experimental confirmation of this finding would be useful. Hayes also assumes that partisan trait inferences are derived from the policy positions taken by candidates of different parties. This assumption is consistent with experimental work; while people frequently make inferences between candidate traits and issue information, they are more likely to infer candidate traits from issue positions rather than the reverse (Rapoport, Metcalf, and Hartman 1989). The causal model suggested – but not yet tested – by these disparate findings is that the partisan affiliation of the candidate produces inferences about policy positions (grounded in both actual communication strategies and stereotypes), which in turn generate trait inferences.

The final stereotypic category is physical appearance, which has an impact on trait inferences and evaluation.^{ix} Absent any other information about a candidate's qualities, more attractive facial appearances produce more positive trait inferences and evaluations (Rosenberg et al. 1986; Sigelman, Sigelman, and Fowler 1987). However, as noted earlier, when information about partisanship is available, physical attractiveness appears to have no impact, suggesting limits to its impact (Riggle et al. 1992). Facial maturity has also been implicated. People attribute more warmth, honesty and submissiveness to baby-faced adults (possessing large eyes, round chins and thick lips) whereas more mature facial features (characterized by small eyes, square jaws and thinner lips) elicit attributions of dominance and strength (Zebrowitz 1994). In a creative experimental demonstration of the political consequences of facial maturity, Keating,

Randall and Kendrick (1999) manipulate the facial images of recent presidents through digital techniques, finding that subtle changes in facial features affected the trait ratings of well-known leaders in a theoretically meaningful fashion.

Two recent experimental research programs demonstrating the potentially powerful effects of facial appearance are noteworthy. Iyengar and his colleagues (e.g., Bailenson et al. 2008), utilizing a morphing technology to digitally alter a candidate's appearance to make it more similar to research participants, show that facial similarity produces more positive trait inferences and higher levels of candidate support, above and beyond the impact of partisan and policy similarity. These facial similarity effects are particularly evident among weak partisans and independents, and for unfamiliar candidates (see Iyengar's chapter in this volume). Todorov shows that rapid judgments of competence, based on facial appearances, predict the outcomes of real gubernatorial, Senate, and House campaigns (and, consistent with the trait literature previously described, inferences of competence from facial appearance, rather than other trait inferences, are key; Hall et al. (2009) provide a good review of Todorov's research program). These facial appearance effects appear to be automatic and outside of conscious awareness, consistent with contemporary psychological understandings of automaticity and social thought (Andersen et al. 2007).

Taken as a whole, the experimental literature provides considerable empirical evidence that stereotyping processes have an impact on candidate impressions. Yet, there are also many important questions that remain unaddressed. For example, too few studies in this tradition manipulate, or take into account, the partisanship of the target candidate, making it impossible to determine the magnitude (if any) of other characteristics on trait and issue inferences when this

politically important characteristic is made salient. Similarly, there has been little consideration of the characteristics of individuals and situations that moderate the impact of factors such as race, gender, and physical appearance. Finally, it is clear that traits and issue positions are important components of impressions and causal determinants of candidate evaluations. However, as implied in the preceding discussion, we know very little about how trait and issue inferences are linked, and the underlying causal dynamics that connect a given candidate's characteristics, the intervening inferences, and the resulting summary evaluation.

5. Cognitive Process Models of Candidate Evaluation

Experiments have been critical in the development and testing of the two models put forth to describe the processes underlying the formation of evaluations of politicians.^x The first posits that evaluations are formed online, with continuous updating of the summary evaluation as new information is encountered (Hastie and Park 1986; Lodge et al. 1989; Lodge and Steenbergen 1995). Under the alternative, memory-based processing, opinions are constructed at the time an opinion is expressed, by retrieving specific pieces of information from long-term memory and integrating that information to create a summary judgment (Zaller 1992; Zaller and Feldman 1992).

The two most important experimental studies of the cognitive process models of candidate evaluation are Lodge and Steenbergen (1995) and Lau and Redlawsk (2006). Neither, interestingly, is notable for the manipulation of specific independent variables that shed light on the causal mechanisms that produce each type of processing. Rather, their importance is the result of ambitious research designs and careful measurement.

Lodge and Steenbergen (1995) represent an advance on prior experimental studies of

candidate evaluation in at least three ways. For one, it is the first study to consider these processes in a comparative context, i.e., participants received information about two candidates running for office. Importantly, despite the comparative context, the experimental instructions and dependent variables focus on evaluations of the candidates, not a choice between them. Second, Lodge and Steenbergen extend the time frame beyond the typical brief, single laboratory sitting, by collecting recall and evaluation data over a month-long period. Their empirical analyses provide support for the dominance of online processing, consistent with most of the previous experimentally-based literature.^{xi} The third important advance of the Lodge and Steenbergen study lies in the compelling implications they draw from their results. They argue forcefully that experimental, as opposed to observational, methods are necessary to understand candidate evaluation processes, because it is only with experiments that researchers can know with certainty the information that individuals have received. Second, they argue that if political scientists focus on information holding and retention, they are likely to underestimate the extent to which campaigns and media have an impact on citizens. Finally, Lodge and Steenbergen reach the normative conclusion that information holding (i.e., recall) is not a proper standard of “good citizenship.” Rather, they contend that what really matters is the types of information that citizens receive, and whether they are responsive to that information.

Lau and Redlawsk’s (2006) inventive research has also substantially expanded our understanding of a wide range of phenomena relevant to candidate evaluation and vote choice. As I previously noted, Lau and Redlawsk developed an innovative dynamic process tracing methodology that stands in stark contrast to the static campaign paradigm used in previous experimental studies. Contrary to Lodge and Steenbergen (1995), the election context and the

necessity of an eventual vote is salient and central to their research design. In regards to the two processing models, Lau and Redlawsk's results (also see Redlawsk 2001) suggest that evaluations of political candidates have a significant memory-based component when a vote choice is required. Their data are compelling on this point, and so provide a significant challenge to the conclusions reached in the Stony Brook studies, which implied voting is the result of online processes (despite the fact that none of those studies required a vote choice from the participants). While Lau and Redlawsk may very well be correct in their conclusion that voting – making a choice -- promotes memory-based processing, an alternative explanation exists. The learning environment in the dynamic process tracing methodology is complex. Research participants learn about four or six candidates at once, with information streaming by at a fast pace -- a pace that may undermine their ability to encode information and update multiple online tallies. Consequently, it is possible that task complexity, rather than the vote choice per se, is responsible for the memory-based results, because complex tasks disrupt normal (in this context, online) processing routines (Kruglanski and Sleeth-Keppler 2007). I recognize that the complexity of Lau and Redlawsk's design was deliberate, because they believe the method matches the complexity of real-world learning about candidates. As a result, the “methodological artifact” explanation put forth here is consistent with their preferred conclusion, namely, that voting promotes memory-based processing. Additional research will be needed to evaluate these competing explanations.

Surprisingly, aside from the vote choice and other contextual factors considered by Lau and Redlawsk (2006) and by Rahn, Aldrich, and Borgida's (1994) study of information format, there has been no other research examining how structural or contextual factors influence the

propensity to engage in online versus memory-based processing, and certainly this is an avenue for future research. For example, one might imagine that different visual images (i.e., personalizing or depersonalizing the candidate) might have an impact (McGraw and Dolan (2007) present evidence consistent with this logic). In contrast to the scarcity of research focusing on the contextual moderators of the two processing models, there has been a good deal of research identifying individual difference moderators. Politically sophisticated individuals are more likely to engage in online processing, whereas those who are less sophisticated tend to engage in memory-based processing (McGraw, Lodge, and Stroh 1990; McGraw and Pinney 1990; McGraw et al. 2003; McGraw and Dolan 2007). In addition, people with a high need to evaluate and people who are “entity theorists” (i.e., who believe other people’s personalities are stable and fixed) are more likely to manifest online processing (McGraw and Dolan 2007).

There is a surprising disconnect between the two empirical traditions that I have reviewed to this juncture. That is, research on the content of candidate impressions has not considered process, or “how citizens assemble their views, how they put the various ingredients together” (Kinder 1998, 812). Similarly, work on online and memory-based processing, incorporating global net tallies, has not been sensitive to the possibility that different kinds of information and inferences may have their impact through different psychological process mechanisms.

One final thought on the two processing models of candidate evaluation: it is increasingly common for scholars, both experimentalists and scholars working with observational data, to reject the either-or approach, despite the fact that those kinds of conclusions characterize most of the empirical work. Rather, scholars often conclude that a hybrid model, incorporating both online and memory-based components, may be more psychologically realistic (Hastie and

Pennington 1989; Zaller 1992; Just et al. 1996; Lavine 2002; McGraw 2003; Lau and Redlawsk 2006; Taber and Lodge's chapter in this volume). In theory, a hybrid model is almost certainly correct. But in practice, it is not clear what this really means, how we would go about empirically testing for it, or the conditions under which we would be more likely to expect hybrid processing to occur.

6. Considering the Impact of Strategic Candidate Behaviors

Much of the experimental literature on candidate impressions and evaluations fails to take into account the self-presentational tactics that politicians engage in to influence their constituents.^{xii} I have noted elsewhere that “these processes are two sides of the same coin, and a complete understanding of what ordinary citizens think about politicians will be out of reach until political psychologists take into account the strategic interplay between elites and the mass public” (McGraw 2003, 395). There, I took a positive approach by reviewing research that is suggestive as to this strategic interplay; here, I focus on some of the many unanswered questions that experiments are well suited to answer.

Fenno's (1978) influential presentation of “home style” -- referring to three sets of activities, in which elected representatives engage -- provides a good starting point. The first activity set is self-presentation along three dimensions: conveying qualifications (competence and honesty), a sense of identification with constituents, and empathy. These dimensions dovetail nicely with the trait ascriptions examined in the research I have described in this chapter. However, the ultimate objective in Fenno's description of political self-presentation is not high approval ratings (evaluations) or trait inferences, but rather the more nebulous and fragile concept of *trust*, including the willingness on the part of constituents to provide leeway to

the representative on legislative decisions. There has been no experimental work linking politicians' self-presentational tactics, trait inferences, and constituent trust.

Fenno's second category of "home style" activities involves allocation of resources to the district (travel back home, staff, casework, communication efforts, etc.) Virtually all analyses of the impact of constituency service are observational, and the research results from those analyses are unpromising: "nearly every scholar who has analyzed resource data, beginning with Fenno, has turned up negative results" (Rivers and Fiorina 1991, 17). In part, as Rivers and Fiorina (1991) demonstrate, the null results are attributable to a problem of endogeneity (i.e., that allocation of resources is dependent upon the representative's expectation of electoral success). There has been very little experimental work aimed at understanding representatives' responsiveness to constituents or how various types of constituency service influence citizens' impressions and evaluations of elected officials (Cover and Brumberg (1982) and Butler and Brockman (2009) provide notable exceptions), but such designs have the potential to provide a great deal of insight into these important phenomena.

Unlike the first two, Fenno's third "home style" activity, explanation of Washington activity to constituents, has been the focus of a fair amount of experimental work (see McGraw (2003) for a review). It is clear from that work that different types of explanations for policy decisions and corrupt or scandalous acts have systematic effects on a host of judgments: attributions of credit and blame, inferences about specific trait characteristics, and global evaluations of politicians. This is a research area, too, however, where a number of unanswered questions remain. For example, we know very little about the consequences of political explanations outside of the laboratory (so this is a call for more observational studies). In

addition, the experimental work has focused on particular types of explanations – excuses and justifications^{xiii} – and, as a result, we know much less about the impact of other accounts -- in particular, apologies. This should be of interest, given the apparent proliferation of apologies in contemporary politics and the high level of attention to apology discourse in the media.

It is also worth considering how experimental work might contribute to our understanding of the impact of “position-taking” (Mayhew 1974) on evaluations and impressions. Three areas seem promising. The first would be to consider how citizens respond to elites’ attempts to modify their issue positions in order to be consistent with public opinion and, in particular, when and why citizens view such movements positively as “democratic responsiveness” (Stimson, MacKuen, and Erikson 1995) or pejoratively as “pandering” (Jacobs and Shapiro 2000). McGraw, Lodge, and Jones (2002) provide some preliminary experimental evidence on these questions, finding that the arousal of suspicion of pandering is the result of a complex interplay between situational cues and agreement with the message. If aroused, suspicion of pandering does have negative consequences for evaluations of public officials.

Second, consider Petrocik’s (1996) theory of “issue ownership,” which argues that parties develop reputations for being more skilled at handling certain policy domains. Specific candidates, in turn, are perceived as more credible over issues “owned” by their party, and so they strive to make the issues associated with their party “the programmatic meaning of the election and the criteria by which voters make their choice” (Petrocik 1996, 828). In a campaign advertising study, Ansolabehere and Iyengar (1994) provide convincing arguments for why experimental investigations of the issue ownership hypothesis are warranted (e.g., in campaigns where other candidate characteristics, such as sex, race, or prior experience is confounded with

party, it is impossible to determine which characteristic has “ownership” of the issue).

Ansola-behere and Iyengar find effects on voting choice but not on other dimensions of candidate evaluation (consistent with the argument that the two are distinct processes), suggesting that further experimental work on when and why issue ownership is consequential for candidate impressions and evaluations would be useful.

The third aspect of candidate position taking that would benefit from experimental investigation is ambiguity. Political scientists have long emphasized that politicians often have an incentive to adopt ambiguous issue positions (Downs 1957; Key 1958), because “by shunning clear stands, they avoid offending constituents who hold contrary positions; ambiguity maximizes support” (Page 1976, 742). Scholars have also formalized the incentives for taking ambiguous positions, as well as the circumstances under which they ought to be avoided. While there is some observational work examining the causes and consequences of ambiguous position taking (Page and Brody 1972; Campbell 1983), there is little experimental work which has examined the factors that lead citizens to view a position as ambiguous, or the conditions under which ambiguous positions produce positive or negative consequences for a candidate who espouses them (Tomz and van Houweling (2009) provide a provocative recent exception).

7. Conclusions

Experiments have provided a platform for important advances in our understanding of candidate impressions and evaluations. Yet, as I have taken care to highlight, there are many significant questions that remain to be answered. In particular, we need to understand the reciprocal linkages among citizen inferences, evaluations and choices, and candidate strategies, by developing more comprehensive theories that integrate psychological and political principles.

It is my hope that the next generation of political science experimentalists will meet these challenges.

References

- Andersen, Susan M., Gordon B. Moskowitz, Irene V. Blair, and Brian A. Nosek. 2007. "Automatic Thought." In *Social Psychology: Handbook of Basic Principles*, eds. Arie W. Kruglanski and E. Tory Higgins. New York: The Guilford Press.
- Ansolabehere, Stephen S., and Shanto Iyengar. 1994. "Riding the Wave and Issue Ownership: The Importance of Issues in Political Advertising and News." *Public Opinion Quarterly* 58: 335-57.
- Ansolabehere, Stephen S., and Shanto Iyengar. 1995. *Going Negative: How Political Advertising Divides and Shrinks the American Electorate*. New York: The Free Press.
- Bailenson, Jeremy N., Shanto Iyengar, Nick Yee, and Nathan A. Collins. 2008. "Facial Similarity between Voters and Candidates Causes Influence." *Public Opinion Quarterly* 72: 935-61.
- Bartels, Larry M. 2010. "The Study of Electoral Behavior." In *The Oxford Handbook of American Elections and Political Behavior*, ed. Jan E. Leighley. New York: Oxford University Press.
- Berinsky, Adam J., and Tali Mendelberg. 2005. "The Indirect Effects of Discredited Stereotypes." *American Journal of Political Science* 49: 846-65.
- Butler, Daniel M., and David E. Broockman. 2009. "Who Helps DeShawn Register to Vote? A Field Experiment on State Legislators." Unpublished manuscript, Yale University.
- Campbell, James E. 1983. "The Electoral Consequences of Issue Ambiguity: An Examination of the Presidential Candidates' Issue Positions from 1968 to 1980." *Political Behavior* 5: 277-91.
- Campbell, Angus, Philip Converse, Warren Miller, and Donald Stokes. 1960. *The American Voter*. New York: Wiley.
- Colleau, Sophie M., Kevin Glynn, Steven Lybrand, Richard M. Merelman, Paula Mohan, and James E. Wall. 1990. "Symbolic Racism in Candidate Evaluation: An Experiment." *Political Behavior* 12: 385-402.
- Cover, Albert D., and Bruce S. Brumberg. 1982. "Baby Books and Ballots: The Impact of Congressional Mail on Constituent Opinion." *American Political Science Review* 76: 347-59.

- Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know about Politics and Why it Matters*. New Haven, CT: Yale University Press.
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.
- Feldman, Stanley, and Pamela Johnston Conover. 1983. "Candidates, Issues, and Voters: The Role of Inference in Political Perception." *Journal of Politics* 45: 810-39.
- Fenno, Richard E. 1978. *Home Style: House Members in their Districts*. Boston: Little, Brown.
- Fiske, Susan T., and Steven L. Neuberg. 1990. "A Continuum Model of Impression Formation: From Category-Based to Individuating Processes as a Function of Information, Motivation, and Attention." In *Advances in Experimental Psychology* Vol. 23, ed. Mark P. Zanna. San Diego, CA: Academic Press.
- Funk, Carolyn L. 1996. "Understanding Trait Inferences in Candidate Images." In *Research in Micropolitics* Vol. 5, eds. Michael X. Delli Carpini, Leonie Huddy, and Robert Y. Shapiro. Greenwich, CT: JAI Press.
- Gilbert, Daniel T. 1989. "Thinking Lightly about Others: Automatic Components of the Social Inference Process." In *Unintended Thought*, eds. John S. Uleman, and John A. Bargh. New York: The Guilford Press.
- Graber, Doris A. 1984. *Processing the News: How People Tame the Information Tide*. New York: Longman.
- Hacker, Kenneth L. 2004. *Presidential Candidate Images*. Lanham, MD: Rowman and Littlefield.
- Hajnal, Zoltan L. 2006. *Changing White Attitudes toward Black Political Leadership*. New York: Cambridge University Press.
- Hall, Crystal C., Amir Goren, Shelly Chaiken, and Alexander Todorov. 2009. "Shallow Cues with Deep Effects: Trait Judgments from Faces and Voting Decisions." In *The Political Psychology of Democratic Leadership*, eds. Eugene Borgida, Christopher M. Federico, and John L. Sullivan. New York: Oxford University Press.
- Hayes, Danny. 2005. "Candidate Qualities through a Partisan Lens: A Theory of Trait Ownership." *American Journal of Political Science* 49: 908-23.
- Hastie, Reid, and Bernadette B. Park, B. 1986. "The Relationship between Memory and Judgment Depends on whether the Task is Memory-based or On-line." *Psychological Review* 93: 258-68.

- Hastie, Reid, and Nancy Pennington. 1989. "Notes on the Distinction between Memory-based Versus On-line Judgments." In *On-line Cognition in Person Perception*, ed. John M. Bassili. Hillsdale, NJ: Erlbaum.
- Huddy, Leonie, and Nayda Terkildsen. 1993. "Gender Stereotypes and the Perception of Male and Female Candidates." *American Journal of Political Science* 37: 119-47.
- Hutchings, Vincent L., and Nicholas Valentino. 2004. "The Centrality of Race in American Politics." *Annual Review of Political Science* 7: 383-408.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.
- Jacobs, Lawrence R., and Robert Y. Shapiro. 2000. *Politicians don't Pander: Political Manipulation and the Loss of Democratic Responsiveness*. Chicago: University of Chicago Press.
- Johnson, Eric J., and J. Edward Russo. 1984. "Product Familiarity and Learning New Information." *Journal of Consumer Research* 11: 542-50.
- Just, Marion R., Ann N. Crigler, Dean E. Alger, Timothy E. Cook, and Montague Kern. 1996. *Crosstalk: Citizens, Candidates, and the Media in a Presidential Campaign*. Chicago: University of Chicago Press.
- Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415-40.
- Keating, Caroline F., David Randall, and Timothy Kendrick. 1999. "Presidential Physiognomies: Altered Images, Altered Perceptions." *Political Psychology* 20: 593-610.
- Kelley, Stanley, and Thad Mirer. 1974. "The Simple Act of Voting." *American Political Science Review* 61: 572-79.
- Key, Vladimir O. 1958. *Politics, Parties, and Pressure Groups* 4th ed.. New York: Crowell.
- Kinder, Donald R. 1986. "Presidential Character Revisited." In *Political Cognition: The 19th Annual Carnegie Symposium on Cognition*, eds. Richard R. Lau and David O. Sears. Hillsdale, NJ: Lawrence Erlbaum.
- Kinder, Donald R. 1998. "Opinion and Action in the Realm of Politics." In *The Handbook of Social Psychology* Vol. 2, 4th Ed, eds. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. New York: McGraw-Hill.
- Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided By Color: Racial Politics and Democratic Ideals*. Chicago: University of Chicago Press.

- Krosnick, Jon A. 1988. "The Role of Attitude Importance in Social Evaluation: A Study of Policy Preferences, Presidential Candidate Evaluations, and Voting Behavior." *Journal of Personality and Social Psychology* 55: 196-210.
- Krosnick, Jon A., and Donald R. Kinder. 1990. "Altering the Foundations of Support for the President Through Priming." *American Political Science Review* 84: 497-512.
- Kruglanski, Arie W., and David Sleeth-Keeper. 2007. "The Principles of Social Judgment." In *Social Psychology: Handbook of Basic Principles*, eds. Arie W. Kruglanski, and E. Tory Higgins. New York: The Guilford Press.
- Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing During Election Campaigns*. New York: Cambridge University Press.
- Lavine, Howard. 2002. "On-line Versus Memory-based Process Models of Candidate Evaluation." In *Political Psychology*, ed. Kristen R. Monroe. Mahwah, NJ: Erlbaum.
- Lodge, Milton, Kathleen M. McGraw, and Patrick Stroh. 1989. "An Impression-driven Model of Candidate Evaluation." *American Political Science Review* 83: 399-420.
- Lodge, Milton, and Marco Steenbergen. 1995. "The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89: 309-26.
- Markus, Gregory B. 1982. "Political Attitudes during an Election Year: A Report on the 1980 NES Panel Study." *American Political Science Review* 76: 538-60.
- Mayhew, David. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale University Press.
- McDermott, Monika L. 1998. "Race and Gender Cues in Low-Information Elections." *Political Research Quarterly* 51: 895-918.
- McGraw, Kathleen M. 2003. "Political Impressions: Formation and Management." In *Handbook of Political Psychology*, eds. David Sears, Leonie Huddy, and Robert Jervis. New York: Oxford University Press.
- McGraw, Kathleen M., and Thomas Dolan. 2007. "Personifying the State: Consequences for Attitude Formation." *Political Psychology* 28: 299-328.
- McGraw, Kathleen M., Mark Fischle, and Karen Stenner. 2000. "What People 'Know' Depends on How They are Asked." Unpublished manuscript, Ohio State University.
- McGraw, Kathleen M., Edward Hasecke, and Kimberly Conger. 2003. "Ambivalence, Uncertainty, and Processes of Candidate Evaluation." *Political Psychology* 24: 421-48.

- McGraw, Kathleen M., and Valerie Hoekstra. 1994. "Experimentation in Political Science: Historical Trends and Future Directions." In *Research in Micropolitics*, eds. Michael X. Delli Carpini, Leonie Huddy, and Robert Y. Shapiro. Greenwich, CT: JAI Press
- McGraw, Kathleen M., Milton Lodge, and Jeffrey Jones. 2002. "The Pandering Politicians of Suspicious Minds." *Journal of Politics* 64: 362-83.
- McGraw, Kathleen M., Milton Lodge, and Patrick Stroh. 1990. "On-line Processing in Candidate Evaluation: The Effects of Issue Order, Issue Salience and Sophistication." *Political Behavior* 12: 41-58.
- McGraw, Kathleen M., and Neil Pinney. 1990. "The Effects of General and Domain-Specific Expertise on Political Memory and Judgment Processes." *Social Cognition* 8: 9-30.
- Miller, Arthur H., Martin P. Wattenberg, and Oksana Malanchuk. 1986. "Schematic Assessments of Presidential Candidates." *American Political Science Review* 79: 359-72.
- Mitchell, Dona-Gene. 2008. "It's about Time: The Dynamics of Information Processing in Political Campaigns." Unpublished dissertation, University of Illinois at Urbana-

Champaign.

Moskowitz, David, and Patrick Stroh. 1994. "Psychological Sources of Electoral Racism." *Political Psychology* 15: 307-29.

Nimmo, Dan, and Robert L. Savage. 1976. *Candidates and their Images: Concepts, Methods, and Findings*. Santa Monica, CA: Goodyear Publishing.

Page, Benjamin I. 1976. "The Theory of Political Ambiguity." *American Political Science Review* 70: 742-52.

Page, Benjamin I., and Richard Brody. 1972. "Policy Voting and the Electoral Process: The Vietnam War Issue." *American Political Science Review* 66: 979-95.

Petrocik, John. 1996. "Issue Ownership in Presidential Elections, with a 1980 Case Study." *American Journal of Political Science* 40: 825-50.

Rahn, Wendy M. 1993. "The Role of Partisan Stereotypes in Information Processing about Political Candidates." *American Journal of Political Science* 37: 472-96.

Rahn, Wendy M., John H. Aldrich, and Eugene Borgida. 1994. "Individual and Contextual Variations in Political Candidate Appraisal." *American Political Science Review* 88: 193-99.

Rahn, Wendy M., Jon A. Krosnick, and Marike Breuning. 1994. "Rationalization and Derivation Processes in Survey Studies of Political Candidate Evaluation." *American Journal of Political Science* 38: 582-600.

Rapoport, Ronald B., Kelly L. Metcalf, and Jon A. Hartman. 1989. "Candidate Traits and Voter Inferences: An Experimental Study." *Journal of Politics* 51: 917-32.

Redlawsk, David. 2001. "You Must Remember This: A Test of the On-line Model of Voting." *Journal of Politics* 63: 29-58.

Riggle, Ellen D., Victor C. Ottati, Robert S. Wyer, James H. Kuklinski, and Norbert Schwarz. 1992. "Bases of Political Judgments: The Role of Stereotypic and Nonstereotypic Judgment." *Political Behavior* 14: 67-87.

Rivers, Douglas, and Morris P. Fiorina. 1991. "Constituency Service, Reputation, and the Incumbency Advantage." In *Home Style and Washington Work: Studies of Congressional Politics*, eds. Morris P. Fiorina and David W. Rohde. Ann Arbor: University of Michigan Press.

Rosenberg, Shawn, Lisa Bohan, Patrick McCafferty, and Kevin Harris. 1986. "The Image and the Vote: The Effect of Candidate Presentation on Voter Preference." *American Journal*

of Political Science 30: 108-27.

Schneider, Monica, and Angela Bos. 2009. "It Don't Matter if You're Black or White? Exploring the Content of Stereotypes of Black Politicians." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.

Sigelman, Lee, Carol K. Sigelman, and C. Fowler. 1987. "A Bird of a Different Feather? An Experimental Investigation of Physical Attractiveness and the Electability of Female Candidates." *Social Psychological Quarterly* 50: 32-43.

Sigelman, Carol K., Lee Sigelman, Barbara Walkosz, and Michael Nitz, 1995. "Black Candidates, White Voters: Understanding Racial Bias in Political Perceptions." *American Journal of Political Science* 39: 243-65.

Stimson, James A., Michael B. MacKuen, and Robert S. Erikson. 1995. "Dynamic Representation." *American Political Science Review* 89: 543-65.

Stroud, Laura, Jack Glaser, and Peter Salovey. 2006. "The Effects of Partisanship and Candidate Emotionality on Voter Preference." *Imagination, Cognition, and Personality* 25: 25-44.

Tomz, Michael, and Robert P. van Houweling. 2009. "The Electoral Implications of Candidate Ambiguity." *American Political Science Review* 103: 83-98.

Williams, Linda F. 1990. "White/Black Perceptions of the Electability of Black Political Candidates." *National Political Science Review* 2: 145-64.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zaller, John R., and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36: 579-616.

Zebrowitz, Leslie A. 1994. "Facial Maturity and Political Prospects: Persuasive, Culpable, and Powerful Faces." In *Beliefs, Reasoning, and Decision Making: Psych-Logic in Honor of Bob Abelson*, eds. Roger C. Schank, and Ellen Langer. Hillsdale, NJ: Erlbaum.

14. Media and Politics

Thomas E. Nelson, Sarah M. Bryner and Dustin M. Carnahan

Nobody needs another summary of mass media research right now; there are plenty of fine, current surveys of the field (Kinder 2003). Our chapter discusses specifically how experimental research provides insight into the relationship between the media and the political world. We are especially interested in important questions that experimentation is well suited to address. Experimentation has been vital to the development of scholarship in this area, but we should also recognize when it is best to step away and choose another method.

Causation and experimentation go together hand in glove, and questions of causation are paramount in both lay and scholarly thought about the media (Iyengar 1990). Questions about the social, economic and organizational factors that determine mass media content are certainly fascinating and relevant in their own right. One can argue, however, that such questions eventually beget questions about the ultimate impact of that content on individuals and political processes and institutions.

As an example of the promise and limitation of experimentation in mass media research, consider the media's constant bugbear: public perceptions of ideological bias in the news. The usual form of this complaint is that media organizations subtly stump for liberal causes (Goldberg 2001). While this complaint often amounts to little more than strategic bluster, it is conceivable that the increasing differentiation of the media marketplace will encourage news organizations to become more forthright in displaying overt liberal or conservative commentary (Dalton, Beck, and Huckfeldt 1998). Experimentation can certainly advance our understanding

of the consequences of media bias, ideological or otherwise. Scholars can rigorously examine, using experimental methods, whether rival coverage of current affairs contributes to different perceptions and opinions among their regular consumers (Entman 2004). Experiments are not well suited, however, to investigating claims about the *prevalence* of such biases. Experiments assume that meaningful variation in mass media content exists; establishing the nature and extent of such variation falls to scholars using other methods, such as content analysis (Patterson 1993; Hamilton 2004).

Convincing experimental work in social science frequently demands a leap of faith to take the stimulus and setting of the experiment as valid instantiations of actual political phenomena (Aronson 1977; Boettcher 2004). With research on the mass media, that leap is more like a hop; the method and the phenomenon seem made for each other. The institutions of the mass media serve up, on a daily basis, morsels of content that the experimenter can excise and transport to the laboratory almost effortlessly. Further, while many topics in political scholarship are plagued by problems of causal order, media scholarship has fewer conundrums, at least with respect to conceptions of short-term media effects. Few of us doubt that day-to-day variations in the topic, framing, style, and other features of news media content properly belong in the cause column, and all manner of individual-level political parameters belong in the effect column. Although made for each other, media and experimental research took their time to get acquainted. The earliest scholarly explorations of the political media were observational. This material has been well reviewed elsewhere; for our purposes, a few important lessons from this era stand out. First is the claim of minimal effects, which argues that the political impact of mass media messages pales in comparison with more proximate influences such as friends and family.

Further, since many media messages are open to multiple interpretations, viewers tend to see what they want to see, thereby leading, at most, to a reinforcement of prior views. Experimental work since this time has cast doubt on the blanket claim of minimal effects (or any blanket claim, really). Still, in methodological parlance, the claim of minimal effects resonates as a cautionary tale about limitations to the generalizability of laboratory findings. A statistically significant effect observed in the tightly controlled conditions of the experimental laboratory might be overwhelmed by multiple competing influences in the real world (Kinder 2007).

To say that the problems of reciprocal causation are relatively mild for media scholarship is not to say that causality is unambiguous. The real knots proved to be selection effects and omitted variables, and these problems ultimately proved intolerable for experimentally-minded researchers. For instance, what about selective exposure? The claim is that people avoid material that challenges their political preconceptions and instead feast on a diet of ideologically congenial media fare (Iyengar and Hahn 2009). Selective exposure is a potential objection to any claim of long-term media exposure effects, political or otherwise. Many critics of violent entertainment media claim that they foster aggressive behavior in their consumers, especially children (Eron 2001). A predictable rebuttal to this criticism, and the cross-sectional data that support it, is selective exposure: violent TV does not create aggressive behavior, but simply attracts aggressive viewers. It takes experimentation to answer the selectivity rebuttal and show conclusively that, yes indeed, a steady diet of violent TV can make an erstwhile pacific child more aggressive.

As for omitted variables, this is a common shortcoming of observational research examining the effects of media attention, namely agenda setting and priming. Nonexperimental

and quasi-experimental work indeed shows that media attention to an issue can heighten the public's concern about it, while also causing attitudes about the president to align more closely with attitudes about that issue (McCombs and Shaw 1972). The potential omitted variable is real-world change in the urgency of that issue. Frequently, all of these phenomena covary and it takes Herculean analyses with fine-grained time-series data to sort out cause from effect (Behr and Iyengar 1985) – or experiments, wherein the researcher has perfect knowledge of what precedes what, because she has designed it that way.

Bartels has called mass media research “an embarrassment” because of its repeated failure to demonstrate convincing effects, and his point remains relevant (Bartels 1993). The mass media are, collectively, a huge institution that consume and metabolize tremendous resources, and yet clear, unambiguous media effects are difficult to spot. The minimal effects era yielded valuable insights about the cognitive, social, and institutional forces that check the independent effects of the mass media. The media are, after all, but one player in the game of contemporary politics. Still, the paucity of positive findings no doubt also reflects weak research design.

Philosophers of science talk about the importance of establishing a causal frame as a precondition for a properly posed scientific question (White 1990). In other words, it is not enough to simply ask if a certain variable is consequential; one must locate the potential consequence within a specific set of background conditions. This leads us to question the claim about minimal effects in circumstances where equal and opposite forces intersect (Zaller 1992). Showing powerful and unambiguous media effects is just as challenging as providing unambiguous evidence that campaigns matter (Vavreck 2009). This is in part because the causal

frame is vague. If campaigns do not matter, then why don't the talented people who run them simply unilaterally disarm? They do not do so because, under a counterfactual causal frame of no opposition activity, one would surely find that campaigns matter an awful lot. Much of the same is said about the criticism that research has exaggerated the political impact of communication frames because, in the real world, the opposition offers its own frame, thereby canceling out the effect of the original frame. It is doubtless the case that in many high-stakes political contests, two equally powerful and widespread frames will compete for public attention, resulting in little net movement in public opinion. Even in such circumstances, it would be inaccurate to conclude that frames do not matter. Two equally powerful locomotives, placed nose to nose and running at full speed, are not going to move very far. It would be a misinterpretation to claim, however, that neither one of them is having an effect, as if the chemical and mechanical forces that supply their power cease to exist. Experimentation allows us to create that counterfactual causal frame, enabling us to investigate the *how* and *why* of social phenomena, not merely their net impact.

1. Media species

The very term *media effects* could be seen as an unconscionable overgeneralization: not all media are alike. Still, most experiments concentrate on the effects of variation within a source category (e.g., television news), rather than across. With the proliferation of media sources, however, it becomes even more urgent to address the question of whether there is an identifiable, unique impact of a particular modality for transmitting information. Observational studies relating consumption habits to political outcomes can certainly be helpful, but eventually raise the same thorny questions about selectivity.

The news

Back in the pre-Internet age, research focused on differences between newspaper and television media effects. Much of this work was inspired by concerns about differences in political engagement and sophistication between those who take in a steady diet of rich printed news compared to those who subsist on sweet but substanceless television. Neuman, Just, and Crigler's (1992) book *Common Knowledge* demonstrated, by using the intersection of survey and experimental data, that while there is more information to be had from print sources, many people actually learn more from television. Television is easy to understand and to decipher, while print is harder and takes more effort. And it is not just selection bias; Neuman et al. have found that medium matters in the lab, a finding which is in line with work both by earlier scholars (Andreoli and Worchel 1978) and more recent research (Druckman 2003).

Sometimes content differences between media are trivial, thus affording a stricter test of modality effects. Stories in the printed version of the *New York Times* are word-for-word identical to those in the online version, but each modality contains something the other lacks. The printed version, for instance, provides cues to a story's importance, via placement and headline. Readers and viewers follow the suggestions of editors and producers, and spend a lot more time on material that is deemed important by the people who put together the news product. So-called indexed news sources instead provide a menu of story choices, and the consumer selection of those choices is presumably guided by more idiosyncratic interests. From the standpoint of normative democratic theory, a case could be made that news producers should encourage citizens to eat their media vegetables, because important stories are not always entertaining. Absent cues to importance, consumers may well adopt a more personalized news

consumption pattern, picking and choosing items that suit their own individual whims, which is exactly what Althaus and Tewksbury (2000) found.

Entertainment

TV news and newspapers obviously differ in many ways, but they share the goal of informing the consumer. What about the political impact of other kinds of programs, who primarily set out to entertain? News magazines and soft news broadcasts, political talk radio, late night comics and even fictional primetime dramas often reference political affairs and figures. And what of the rise of feature-length political documentaries such as *Fahrenheit 9/11* and *An Inconvenient Truth*? They make no pretense of neutrality, but seek to deliver a partisan message with enough panache to hold the viewer's attention for ninety minutes or more. Because audiences welcome entertainment programming as a diversion from serious fare, they might process such content passively, abandoning the normal activities of deliberation and counterargument characteristic of the news viewer (Zaller 1992). Despite these differences, or perhaps because of them, exposure to entertainment media can exert some of the same influences long attributed to exposure to news media, specifically priming and agenda setting. Holbrook and Hill (2005) show that exposing audiences to entertainment media in the form of a crime drama leads participants to emphasize the importance of addressing the issues of crime and violence in their post-test responses.

Additionally, entertainment media -- unlike hard news -- often seek to elicit sharp emotional reactions in viewers, ranging from sympathy to rage. Experiments are particularly adept at shedding light on how these emotional responses affect public opinion and political evaluations. Holbert and Hansen (2008) offer one example of how emotional reactions -- in this

case, anger – can mediate the relationship between debate exposure and perceived candidate performance. When exposed to the movie *Fahrenheit 9/11*, participants reported increased levels of anger toward both major party candidates in the 2004 presidential election, which then affected how participants evaluated each candidate's debate performance. Kim and Vishak's (2008) experimental study presented evidence that the fact-based nature of news programming promoted the use of memory-based processing -- leading to increased recall of factual information and the use of that information in forming and expressing political attitudes. Entertainment media, on the other hand, promoted online information processing; individuals could recall little factual detail from their program (an episode of *The Daily Show*) but still responded to the information in their reported impressions. Concerns about the political enfeeblement of audiences have become even more acute with the rise of soft news. Most of the extant work has been observational (see Baum 2003 and Prior 2003 for an overview of the debate). Experimental work clearly has as much to offer to this debate as it did to *Common Knowledge*.

New media

A truism holds that research lags behind technological and social change, but a new generation of media scholars has embraced experimentation as the most appropriate method to examine the far-reaching implications of information-technology developments. A notable example is the study of computer-mediated (or online) discussion. Political discussion and deliberation has long been thought of as an asset for democracy, fostering understanding on divisive issues through allowing citizens to defend their positions and explicate their reasoning (Goodin 2008). The Internet provides an unprecedented virtual public forum for diverse voices to

assemble and discuss pressing issues. However, some scholars have argued that online discussion is substantively different from face-to-face (FTF) deliberation in terms of both process and consequence. Does online political discussion differ from FTF discussion in ways that make it less valuable for the promotion and maintenance of democracy?

For all their merits, observational studies might not supply the best answers to these questions, as they depend upon unreliable self-report measures of how people interact in online political conversations or how these behaviors differ from traditional FTF interactions. Experimentation, on the other hand, provides investigators with the opportunity to observe precisely how these types of discussions vary -- and therefore differ in terms of their deliberative value -- through allowing a direct comparison between these two forms.

Based on much of this experimental evidence, deliberative theorists apparently have little to worry about with regards to the future of political discussion in a new media environment; individuals that participate in online political discussions of various sorts are more willing to express their opinions as a result of anonymity (Ho and McLeod 2008) and receive comparable levels of political information through their conversations as FTF discussants. Still, democracy's ills will not all be solved by online discussion. For example, anonymity in discussion appears to undermine the credibility of the source, thereby making participants less likely to trust and learn from their fellow discussants (Postmes, Spears, and Sakhel 2001).

Other areas of research concerning media effects are also affected by the rise of new media, including, perhaps what we even choose to call media. An abundance of observational studies show that younger generations increasingly access information from new media sources, including social networking sites like facebook.com, youtube.com, and a myriad of blogs (see

Prensky 2001; Kohut 2008; Winograd and Hais 2009; Madden 2009). As the environment for sources like these is extremely dynamic, very little experimental work has yet been conducted, although we can only suspect how promising work in this area could be. Is information from these alternative media sources treated with a higher level of skepticism than media from traditional sources (see Cassese et al. 2010 for an early exploration of this phenomenon)? Do people learn as much from these sources? Does online content's malleable form force us to question what it means to be a journalist? All of these questions could be explored experimentally, and to great end.

2. Media Effects

Priming and agenda setting

We now turn to phenomena that do not depend on a specific communication modality; that is, they can, in theory, occur whether the medium is newspaper, Internet, or smoke signal. We define *media effect* in this context as a signature consequence of a distinctive practice of mass media organizations that can -- perhaps not easily -- be separated from the mere informational content they report. For example, the mass media did not kill Michael Jackson, but decisions they have made about how to report his death -- from the sheer volume of coverage to the balance taken between discussions of his public and private lives -- will shape his legacy far beyond the mere fact of his passing.

Experiments can take us far toward understanding the consequences of such media decisions. If there is a single book that has done more than any other to accentuate the advantages of experimentation, that book would be *News that Matters* (Iyengar and Kinder 1987). Interestingly, for all the ground broken by this volume, its theoretical claims are really

extensions and qualifications of the minimal effects school. The two important phenomena the book explores – agenda setting and priming -- are functions of media *attention*, not media *content*. Furthermore, there are no direct effects on political opinions of any sort. In keeping with the "cognitive miser" model of human social thought, the book argues that the media do not really change attitudes, they simply stir the mixture of ingredients that constitute our opinions. Research on priming and presidential evaluation is largely disinterested in the content of the primed stories. Just about any issue that the news highlights is likely to exert extraordinary influence on judgments of presidential competence. The racial priming literature has a special concern with content. The landmark study is Mendelberg's (2001) *The Race Card*, which offers an implicit-explicit (IE) model of racial priming: (1) messages can prime racial attitudes, making them assume a more prominent role in subsequent political evaluations; and (2) implicit cues are usually more powerful primes than are explicit cues. We hereby offer an extended digression into controversy over the racial priming hypothesis, as it provides a textbook example of experimentation's vitality in political science, especially for investigating the *conditions* under which a phenomenon occurs, and the *mechanisms* responsible for its occurrence.

The critiques of the IE model are largely methodological, not theoretical. In a large-scale, survey-based experiment, Huber and Lapinski (2008) failed to replicate Mendelberg's findings that implicit racial cues are more effective in activating racial sentiments than are explicit cues. Huber and Lapinski's experiments were a conceptual, not direct replication of Mendelberg in terms of stimuli, procedures, measures and sample. Conceptual replications test the robustness of a finding, determining whether or not theoretically irrelevant alterations in design, materials, or procedure moderate the effect. A conceptual replication failure is an ambiguous signal, however.

It could mean that the theory is not an authentic representation of actual political processes, or that the phenomenon revealed by the original research is real, but narrow and trivial. Huber and Lapinski argue the latter, attributing Mendelberg's positive findings to her sample – which they argued to be especially likely to exhibit the implicit/explicit difference. Huber and Lapinski's sample – which did not display any equivalent moderating effect for the explicitness of the racial signal – was more representative of the general population.

Huber and Lapinski's null finding does not therefore undermine the *qualified* claim that, *for some people*, implicit racial cues have meaningfully different consequences than explicit racial cues. By concentrating her experiment among a sample of such individuals, however, Mendelberg has effectively exaggerated the generalizability of her findings. This criticism about sample unrepresentativeness has dogged laboratory experiments for decades (see Druckman and Kam's chapter in this volume). In her defense, we should note that Mendelberg anticipated the critique against using college sophomores by literally taking her experiment on the road, toting her equipment to the homes of her nonstudent participants.

A third possible explanation for the failure of conceptual replications – invoked by Mendelberg (2008) in her rebuttal to Huber and Lapinski – is the lack of correspondence among separate operationalizations of the key constructs: manipulations, measures, moderators, and mediators. The crucial variable in this research is the *explicitness* of the racial cue. In theory, a number of messages might activate racial sentiments; Mendelberg, in fact, lists seventeen experiments that rely on a wide range of stimuli. In theory, again, some of these cues will be recognized as making obvious appeals to racial antipathy, in violation of the norm of racial

equality. Once messages rise to the level of explicitness, the egalitarian norm allegedly kicks in and the recipient suppresses any incipient prejudices.

Mendelberg, however, argues that Huber and Lapinski's experiments fail to properly operationalize explicitness by utilizing two stimuli that prime racial feelings in equal measure but above and below the threshold of explicitness. Huber and Lapinski wisely include a manipulation check, designed to measure whether the stimuli provides a legitimate test of the IE hypothesis, by asking participants whether they thought the messages were "good for democracy." While participants exposed to explicit ads were found to evaluate explicit ads as "somewhat bad" to a relatively greater degree than those exposed to the implicit ads (between fifteen and twenty percent in the implicit condition compared to twenty-five percent in the explicit condition), Mendelberg complains that the absolute proportion is small and changes little, suggesting that a majority of respondents in the explicit condition do not feel the irresistible tug of racial egalitarianism.

Framing

Framing research puts media content front and center by claiming it is not simply *whether* an issue is covered, but *how* it is covered that matters. As the debate over implicit racial cues illustrates, so much of an experiment's value depends on the translation of an abstract concept into a concrete treatment -- that is, the operationalization of the independent variable. At a theoretical level, researchers studying framing must tame an unruly concept that, however intuitive it might seem, is defined in different ways by different scholars, both across and within fields (Schaffner and Sellers 2009). Even if a consensus could be obtained on the dictionary definition of framing, there are many variations in the experimental operationalization of this

concept. We argue that messages that convey entirely different objective information should not be considered alternative frames. By analogy, think about two different ways of changing visual perspective. One could look out in a certain direction, say, Southeast, describe what one sees, and then pivot on one's foot to a different compass point, and then describe that view. The two descriptions will likely be quite different because they will refer to different sets of objects. Next, consider slowly circling around a stationary object and describing it from different vantage points. Each description will refer to the same object, but will still vary, since different features of the object will come into prominence. Alternative frames do not change the object of description, merely the way it is characterized.

We can pose many fascinating questions about why one news organization might cover an issue in one particular way, while another organization might cover the same issue quite differently (Price and Tewksbury 1997). Experiments would be a very poor choice of method to gain traction on such questions, but they can help us understand whether such variations make any difference. The question of *whether* framing affects opinion and behavior is often followed, in the next breath, with the question of *when* it carries such effects (Druckman 2001). This is a question concerning the boundary conditions governing framing effects. Most of us believe framing happens, but surely it doesn't *always* happen.

3. Message processing

Media scholars are rarely content simply to investigate *whether* a particular effect happens, or even *when*; they also want to know *why*. If experimentation's principal value to the social sciences is investigating causation, then a close second is surely investigating psychological mechanism. Observational research has made strides in incorporating measures

that reveal, if crudely, psychological processes (e.g., reaction-time measures in telephone interviews). Still, the experimenter's kit overflows with specialized tools for revealing what takes place between stimulus and response (but see Bullock and Ha's chapter in this volume).

Scholars with a rationalist inclination make the unsurprising but still important claim that the chief psychological effect of media consumption is *learning*. In other words, the media teach us what we need to know to make sensible political decisions (Lau and Redlawsk 2006). Psychological theory points to other processes that are less obvious, while helping to unpack the generic learning effect. Various dual mode theories suggest two broad categories of psychological response to communication, especially to persuasive messages: 1) a more thoughtful, effortful "central," "systematic," or "piecemeal" route; and 2) a quicker, superficial "peripheral," "heuristic," or "stereotypic" route (Cacioppo and Petty 1982; White and Harkins 1994; Mutz and Reeves 2005;). Gone are the days when media scholars assumed uniform effects across the general public. Even studies employing the proverbial college sophomore have shown important moderator effects. Individual traits and qualities, such as political knowledge and sophistication (Nelson, Oxley, and Clawson 1997), trust in media (Gunther 1992), value orientation (Johnson 2007), need for cognition and evaluation (Neuman et al. 1992), and race (White 2007) significantly moderate media effects. Braverman (2008) examines both involvement with the material as well as the way the message is transmitted, and finds an interactive effect between the two. Experiments thus provide an excellent way to look at individual differences in processing across media sources, something that survey research cannot tap in such a controlled manner.

Psychological theorizing about framing initially posited that it functions much like priming; that is, frames subtly draw our limited attentional resources toward some considerations, and away from others. Subsequent research has expanded the set of psychological processes implicated by framing. Price and Tewksbury (1997) argue that frames alter the applicability of stored information to the framed issue. Under certain frames, a cognition may no longer be perceived to fit the issue. Nelson and colleagues have argued that frames also operate in a more mindful way by affecting judgments of the importance or relevance of cognitions (Nelson, Clawson, and Oxley 1997).

Research on message processing surfaces in scholarship on new media. The distinctive qualities of new communication forms – dynamism, decentralization, nonlinearity, and the fading boundary between producer and consumer – suggest a sea change in the acquisition and use of information. Hypermedia structure is said to mimic cognitive structure at the individual level, with semantic nodes linked by association to form a conceptual web. This distinctive structure, in theory, facilitates the absorption and retention of hypermedia content (Eveland and Dunwoody 2002). Wise and coauthors manipulate the amount of content available to a website visitor and, by taking readings of the heart rates of the participants, find that in a richer media environment (one that displays more stories), participants use higher levels of cognitive resources (Wise, Bolls, and Schaefer 2008). Other scholars, rather than manipulating the amount of information, instead manipulate the interactivity of the website, finding that more interactive websites lead to higher levels of processing (Sicilia, Ruiz, and Munuera 2005). They suggest that the relationship between number of hyperlinks and depth of processing may be nonlinear, and that if individuals are presented with too much information, they shut down.

4. External validity

A method that maximizes our confidence in causal hypotheses is not much good if all we learn about is what happens in our laboratory. Generalizability has several dimensions, including *mundane realism* and *psychological realism*. We can define mundane realism as verisimilitude: the correspondence between features and procedures of the experiment and those prevalent in the real world. Psychological realism refers to the engagement and arousal of similar psychological processes to those that prevail in analogous situations in the real world. Solomon Asch's (1955) classic experiments on conformity resemble nothing we know in the real world, and yet it cannot be denied that conformity pressure was positively suffocating for the subjects in those famous studies.

The importance of this distinction is apparent in many of the research traditions on mass media effects, including scholarship on modality effects on political learning. To provide the proper causal frame, we should ask not whether newspaper consumers are more knowledgeable than, say, those who get their news almost exclusively from television; instead, we need to ask whether or not *the same individual* would learn more or less if they got their news from a different source.

Framed in this manner, experiments become a natural research choice, but that is just the first of our decisions. We could, for example, take representative samples of coverage from two distinct media, randomly assign research participants to receive one set or the other, and compare their knowledge and understanding of politics following these treatments. But is this the best way to go about it? That partly depends on what we mean by a *modality effect*, and this is where questions about generalizability emerge in high relief. From a strict mundane realism

perspective, the aforementioned approach is best. But we can take a sharp scalpel and separate questions about the effects of *typical* coverage across diverse media from questions about the *inherent* differences between media. Media differ not just in the obvious perceptual qualities, but also in characteristic ways that journalism is practiced. These differences are not inherent to the perceptual qualities of the media, they amount to institutional folkways. As we all know, newspapers present far more information than a typical news broadcast. It doesn't have to be this way, but it is.

So should we remain faithful to inherent differences or typical differences? Neither answer is obviously better than the other, since they represent equally legitimate framings of the question about media effects. The former formulation is likely to appeal more to media researchers with a pragmatic political orientation: they want to know what happens in the world of actual politics, and how characteristic practices leave their mark. Media researchers with deeper psychological interests will want to know what it is about television that leads to differing levels of information absorption and refinement relative to print or other media, irrespective of the tendency for print media to include a greater overall volume of information.

The external validity gauntlet was thrown down with flourish by Iyengar and Kinder's seminal experiments. These researchers invested great effort in putting together stimuli, procedures, and a laboratory environment that closely mimicked the typical real-world news consumption experience. This standard for mundane realism has been matched, but never exceeded (Brader 2005). Was all this effort worth it? One of the most important reasons, in our judgment, that the work has had such lasting impact is because of the exacting measures taken to anticipate and preempt criticism on external validity grounds. Such perceptions matter, but are

they based on a myth about the weakness of experiments conducted under less externally valid conditions? There are, of course, strong claims about the superiority of evidence collected under conditions of greater mundane realism but, for the most part, these are speculations or common sense bromides. Critics who fret about the verisimilitude of social science experiments should take note that cutting-edge experimental physics uses laboratory conditions that have not occurred anywhere in the universe since approximately one half-second after the commencement of the Big Bang (Kaku 2008). What little systematic evidence we have does not strongly support the case that more realism equals more validity (Anderson, Lindsay, and Bushman 1999). Investments in external validity could be costly in ways beyond simply the expenditure of resources. Experimenters did not repair to the laboratory just because it was close to their offices; the laboratory setting is simply a natural extension of the logic of experimental control. Isolating the effect of one variable requires controlling for the effects of systematic and unsystematic error. The former contributes to Type I error (false positives); the latter to Type II (false negatives). Laboratory settings, all things being equal, help to minimize the impact of variables that would water down the impact of the experimental stimuli and lead to a false rejection of the null hypothesis.

This is the great advantage of Iyengar and Kinder's studies, to some. That their experiments yielded positive evidence of agenda setting and priming, despite all the potential distractions of their soft laboratory setting, suggests that such phenomena are likely to have real impact in the real world. Yet perhaps even a soft laboratory is not enough. Sniderman and Kinder are rightly lauded as pioneers in applying experimental methods to the study of political communication, and yet both have publicly criticized framing research for its excessive reliance

on (laboratory) experimentation. The critiques boil down, once again, to mundane realism. Sniderman and Theriault (2004) complain that the typical experimental *treatment* makes crucial departures from real-world frames; Kinder (2007) says that the laboratory *setting* is too unreal because it rivets the participant's attention to stimuli that, outside of the laboratory, might never register with us.

Such a view brings us full circle to questions about the ultimate scientific purpose of experimental work. To paraphrase McGuire (1983), experiments are better attuned to investigating what *can* happen than what *will* happen. They are singularly excellent tools for theory building, which is a realm of abstraction and ideal. Core concepts and vital mechanisms are thrown into high relief, while nuisance factors, complications, and contingencies are set aside for another day. As a class, experimenters are unconcerned with achieving an exact calibration of the magnitude of the effects of their variables outside the laboratory. For all the attention given to mundane realism in their studies, Iyengar and Kinder never issued any precise claims about how much the media can affect the public's agenda. For example, how much media emphasis would be needed, in terms of amount and prominence of news stories, to move global warming to the top of the national agenda? We doubt anyone is prepared to make precise estimates of such quantities on the basis of laboratory findings, however realistic the conditions under which they were obtained.

Furthermore, increasing the external validity of an experiment can be associated with an increase in ethical concerns. Many media experiments involve deceiving, however temporarily, the study participants. In order for the experiment to capture how a person might really react to a piece of media, it seems logical that the person should think that the treatment is real. However,

this brings with it a wealth of other problems -- how long lasting are media effects, even those done in the lab? If they last long enough for the person to think they are real, we might expect that the participant's attitude is somehow affected. Given the possibility of potentially long-lasting effects, we believe that researchers should carefully consider the ethical implications of experiments high in external validity.

We hasten to add that confidence in causal inference does *not* require a laboratory setting. The resourceful, inventive researcher can conduct field experiments that have all the rigor of a classic experimental design but add a naturalistic setting and/or manipulation. Not only does a well-designed field experiment negate external validity concerns, but it provides an avenue toward greater precision in estimating the magnitude of relationships between variables of interest. In other words, it not only helps to answer the "So what?" question, but also the "How much?" question.

Nevertheless, the number of field experiments is dwarfed by that of laboratory investigations, for the obstacles facing the field experimenter are formidable (for an extended discussion, see Gerber's chapter in this volume). The literature on media effects is as guilty as any, suggesting that this might be a growth area for enterprising researchers. One exemplary effort was inspired by the civic journalism movement, which seeks to move journalistic practice from the mere recounting of facts to the promotion of public discussion and participation. Researchers collaborated with media outlets, who systematically varied their production to represent "old" and "new" journalistic styles; consumers of these different journalistic products were later surveyed along various dimensions such as intent to vote, engagement in political discussion, and involvement with civic groups (Denton and Thorson 1998).

Once in a great while, opportunistic investigators can even take advantage of naturally occurring manipulations to conduct a *quasi-experiment*. Such investigations typically lack random assignment and/or carefully controlled manipulations, but they more than make up for these shortcomings by supplying systematic observations of the effects of tangible variation in the phenomena we care about. An example is Mondak's (1995) study of the consequences of the 1992 Pittsburgh newspaper strike. Mondak shows that the strike did not cause the good people of Pittsburgh to suffer decrements in national or international political awareness relative to demographically comparable residents of Cleveland.

5. Conclusion

The media never sleep, nor does innovation in communications technology. Today's consumer of the latest in trendy communication toys is tomorrow's befuddled techno has-been, beseeching his or her teenager to "make this damned thing work". The media technology universe has changed so drastically that we will likely witness the demise of the printed daily newspaper within years, not decades. We must wonder if media scholarship's conventional wisdom will similarly obsolesce.

Fortunately, clear thinking about concepts and processes never goes out of style. The accelerating pace of change in information technology *is* unnerving, but we have to be careful about confusing superficial with fundamental change. What is *really* changing? The sheer amount of available information and the ease with which it can be accessed by ordinary people? The size and breadth of the communication networks commanded by citizens? The practice of journalism by people who are not professional journalists? And what are the consequences of such changes? Just as the Internet has opened up opportunities for new forms of criminal

behavior, so too is it likely that the political consequences of innovations in information technology won't all be beneficial.

Sound science doesn't always require a good theory, but it is not a bad place to start and there are plenty of fascinating and pertinent theories of mass communication to choose from. Perhaps, with the diversification and personalization of mass media sources, strong priming and agenda setting effects will become rare. Perhaps the rise of user generated content will usher in a new era of political trust and involvement. Perhaps ideological polarization will accelerate with the proliferation of partisan information sources.

It seems pointless to speculate about the future direction of information technology change, and the havoc it will create for politics. If the giants of 20th century media research anticipated YouTube, mobile broadband, blogs, RSS feeds, podcasts, social networking websites, and the like, they didn't tell us. It does seem safe, if a little cowardly, to predict that experimentation will be a vital part of the scholarly analysis of such developments.

References

- Althaus, Scott L., and David Tewksbury. 2000. "Patterns of Internet and Traditional News Media use in a Networked Community." *Political Communication* 17: 21-45.
- Anderson, Craig A., James J. Lindsay, and Brad J. Bushman. 1999. "Research in the Psychological Laboratory: Truth Or Triviality?" *Current Directions in Psychological Science* 8: 3-9.
- Andreoli, Virginia, and Stephen Worchel. 1978. "Effects of Media, Communicator, and Message Position on Attitude Change." *Public Opinion Quarterly* 42: 59-70.
- Aronson, Elliot. 1977. "Research in Social Psychology as a Leap of Faith." *Personality and Social Psychology Bulletin* 3: 190-5.
- Asch, Solomon E. 1955. "Opinions and Social Pressure." *Scientific American* November: 31-5.

- Bartels, Larry M. 1993. "Messages Received: The Political Impact of Media Exposure." *American Political Science Review* 87: 267-85.
- Baum, Matthew A. 2003. "Soft News and Political Knowledge: Evidence of Absence or Absence of Evidence?" *Political Communication* 20: 173-190.
- Behr, Roy L., and Shanto Iyengar. 1985. "Television News, Real-World Cues, and Changes in the Public Agenda." *Public Opinion Quarterly* 49: 38-57.
- Boettcher III, William A. 2004. "The Prospects for Prospect Theory: An Empirical Evaluation of International Relations Applications of Framing and Loss Aversion." *Political Psychology* 25: 331-62.
- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49: 388-405.
- Braverman, Julia. 2008. "Testimonials Versus Informational Persuasive Messages: The Moderating Effect of Delivery Mode and Personal Involvement." *Communication Research* 35: 666-94.
- Cacioppo, John T., and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42: 116-31.
- Cassese, Erin, Weber, Christopher, Hauser, David, and Steven Nutt. 2010. "Media, Framing, and Public Opinion: How does YouTube Fit in?" Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Dalton, Russell J., Paul A. Beck, and Robert Huckfeldt. 1998. "Partisan Cues and the Media: Information Flows in the 1992 Presidential Election." *American Political Science Review* 92: 111-26.
- Denton, Frank, and Esther Thorson. 1998. "Effects of a Multimedia Public Journalism Project on Political Knowledge and Attitudes." In *Assessing Public Journalism*, eds. Edmund B. Lambeth, Philip Meyer, and Esther Thorson. Columbia, MO: University of Missouri Press.
- Druckman, James N. 2001. "The Implications of Framing Effects for Citizen Competence." *Political Behavior* 23: 225-56.
- Druckman, James N. 2003. "The Power of Television Images: The First Kennedy-Nixon Debate Revisited." *The Journal of Politics* 65: 559-71.
- Entman, Robert M. 2004. *Projections of Power : Framing News, Public Opinion, and U.S. Foreign Policy*. Chicago: University of Chicago Press.

- Eron, Leonard D. 2001. "Seeing is Believing: How Viewing Violence Alters Attitudes and Aggressive Behavior." In *Constructive and Destructive Behavior: Implications for Family, School, and Society*, eds. Arthur C. Bohart and Deborah J. Stipek. Washington, DC: American Psychological Association.
- Eveland, William P., Jr., and Sharon Dunwoody. 2002. "An Investigation of Elaboration and Selective Scanning as Mediators of Learning from the Web Versus Print." *Journal of Broadcasting & Electronic Media* 46: 34-53.
- Goldberg, Bernard. 2001. *Bias: A CBS Insider Exposes how the Media Distort the News*. Washington, DC: Regnery Publishing.
- Goodin, Robert E. 2008. *Innovating Democracy : Democratic Theory and Practice After the Deliberative Turn*. New York: Oxford University Press.
- Gunther, Albert C. 1992. "Biased Press or Biased Public? Attitudes Towards Media Coverage of Social Groups." *Public Opinion Quarterly* 56: 147-67.
- Hamilton, James. 2004. *All the News that's Fit to Sell: How the Market Transforms Information into News*. Princeton, NJ: Princeton University Press.
- Ho, Shirley S., and Douglas M. McLeod. 2008. "Social-Psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication." *Communication Research* 35: 190-207.
- Holbert, R. Lance, and Glenn J. Hansen. 2008. "Stepping Beyond Message Specificity in the Study of Emotion as Mediator and Inter-Emotion Associations Across Attitude Objects: Fahrenheit 9/11, Anger, and Debate Superiority." *Media Psychology* 11: 98-118.
- Holbrook, Robert A., and Timothy Hill. 2005. "Agenda-Setting and Priming in Prime Time Television: Crime Dramas as Political Cues." *Political Communication* 22: 277-95.
- Huber, Gregory A., and John S. Lapinski. 2008. "Testing the Implicit-Explicit Model of Racialized Political Communication." *Perspectives on Politics* 6: 125-34.
- Iyengar, Shanto. 1990. "The Accessibility Bias in Politics: Television News and Public Opinion." *International Journal of Public Opinion Research* 2: 1-15.
- Iyengar, Shanto, and Kyu S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communications* 59: 19-39.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News that Matters : Television and American Opinion*. Chicago: University of Chicago Press.
- Johnson, Marcia K. 2007. "Reality Monitoring and the Media." *Applied Cognitive Psychology* 21: 981-93.

- Kaku, Michio. 2008. *Physics of the Impossible: A Scientific Exploration into the World of Phasers, Force Fields, Teleportation, and Time Travel* 1st Ed. New York: Doubleday.
- Kim, Young Mie, and John Vishak. 2008. "Just Laugh! You Don't Need to Remember: The Effects of Entertainment Media on Political Information Acquisition and Information Processing in Political Judgment." *Journal of Communication* 58: 338-60.
- Kinder, Donald R. 2003. "Communication and Politics in the Age of Information." In *Oxford Handbook of Political Psychology*, eds. David O. Sears, Leonie Huddy, and Robert Jervis. New York: Oxford University Press.
- Kinder, Donald R. 2007. "Curmudgeonly Advice." *Journal of Communication* 57: 155-62.
- Kohut, Andrew. 2008. "The Internet Gains in Politics." *Pew Internet & American Life Project* January 11. Retrieved from <http://www.pewinternet.org/Reports/2008/The-Internet-Gains-in-Politics.aspx>.
- Lau, Richard R., and David P. Redlawsk. 2006. *How Voters Decide: Information Processing during Election Campaigns*. Cambridge: Cambridge University Press.
- Madden, Mary. 2009. "The Audience for Online Video Sharing Sites Shoots Up." *Pew Internet & American Life Project*. Washington, DC: Pew Research Center.
- McCombs, Maxwell E., and Donald L. Shaw. 1972. "The Agenda-Setting Function of the Mass Media." *Public Opinion Quarterly* 36: 176-87.
- McGuire, William J. 1983. "A Contextualist Theory of Knowledge: Its Implications for Innovation and Reform in Psychological Research." In *Advances in Experimental Social Psychology* Vol. 16, ed. Leonard Berkowitz. San Diego, CA: Academic Press.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Mendelberg, Tali. 2008. "Racial Priming: Issues in Research Design and Interpretation." *Perspectives on Politics* 6: 135-40.
- Mondak, Jeffery J. 1995. "Newspaper and Political Awareness." *American Journal of Political Science* 39: 513-27.
- Mutz, Diana C., and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Political Science Review* 99: 1-15.
- Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and its Effect on Tolerance." *American Political Science Review* 91: 567-83.

- Nelson, Thomas E., Zoe M. Oxley, and Rosalee A. Clawson. 1997. "Toward a Psychology of Framing Effects." *Political Behavior* 19: 221-46.
- Neuman, W. Russell, Marion R. Just, and Ann N. Crigler. 1992. *Common Knowledge: News and the Construction of Political Meaning*. Chicago: University of Chicago Press.
- Patterson, Thomas E. 1993. *Out of Order* 1st Ed. New York, NY: A. Knopf.
- Postmes, Tom, Russell Spears, and Khaled Sakhel. 2001. "Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior." *Personality and Social Psychology Bulletin* 27: 1243-54.
- Prensky, Marc. 2001. "Digital Natives, Digital Immigrants." *On The Horizon*, NCB University Press 9: October.
- Price, Vincent, and David Tewksbury. 1997. "News Values and Public Opinion: A Theoretical Account of Media Priming and Framing." In *Progress in the Communication Sciences*, eds. G. Barnett and F. J. Boster. Greenwich, CT: Ablex.
- Prior, Markus. 2003. "Any Good News in Soft News? The Impact of Soft News Preference on Political Knowledge." *Political Communication* 20: 149-71.
- Schaffner, Brian F., and Patrick J. Sellers, eds. 2009. *Winning with Words: The Origins and Impact of Framing*. New York: Routledge.
- Sicilia, Maria, Salvador Ruiz, and Jose L. Munuera. 2005. "Effects of Interactivity in a Web Site: the Moderating Effect of Need for Cognition." *Journal of Advertising* 34: 31-44.
- Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*, eds. Willem E. Saris, and Paul M. Sniderman. Princeton, NJ: Princeton University Press.
- Vavreck, Lynn. 2009. *The Message Matters : The Economy and Presidential Campaigns*. Princeton, NJ: Princeton University Press.
- White, Ismail. 2007. "When Race Matters and When it Doesn't: Racial Group Differences in Response to Racial Cues." *American Political Science Review* 101: 339-54.
- White, Paul H., and Stephen G. Harkins. 1994. "Race of Source Effects in the Elaboration Likelihood Model." *Journal of Personality and Social Psychology* 67: 790-807.
- White, Peter A. 1990. "Ideas About Causation in Philosophy and Psychology." *Psychological Bulletin* 108: 3-18.

Winograd, Morely, and Michael D. Hais. 2009. *Millennial Makeover: MySpace, YouTube, and the Future of American Politics*. New Brunswick, NJ: Rutgers University Press.

Wise, Kevin, Paul D. Bolls, and Samantha R. Schaefer. 2008. "Choosing and Reading Online News: How Available Choice Affects Cognitive Processing." *Journal of Broadcasting & Electronic Media* 52: 69-85.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York, NY: Cambridge University Press.

15. Candidate Advertisements

Shana Kushner Gadarian and Richard R. Lau ⁱ

No where can democracy be better seen “in action” than during political campaigns. As Lau and Pomper (2002, 47) argue, “Democracy is a dialogue between putative leaders and citizens. Campaigns provide the most obvious and the loudest forums for this dialogue. Candidates try to persuade voters to cast a ballot and to support their cause. Voters respond by coming to the polls and selecting their preferred candidates.” Candidates make their arguments in speeches at campaign rallies and on their web sites, but those venues are primarily experienced by the most committed of partisans. It is only through their campaign advertisements that candidates have any chance of reaching uncommitted voters. And in times of even very approximate party balance, it is uncommitted voters who usually determine election outcomes. Hence political ads are arguably the vehicle through which democracy operates.

Consider the choices facing a political candidate at the outset of a democratic election campaign. To simplify, let us assume a candidate’s goal is to win the election, but she is facing one or more opponents who have the same goal and therefore want to defeat her. All candidates must decide what strategy to follow in order to maximize their chances of winning. But all candidates face resource limits that put very real constraints on what it is possible to do. Therein lies the rub. Candidates must decide how they can get the biggest bang for their buck. Of the myriad of different strategies they could follow, which one is most likely to result in electoral victory? Over the past fifty years, the medium of choice has been television for those candidates with sufficient resources to afford televised ads. But even limiting attention to television only

slightly reduces the options available.

Now reverse the perspective to that of a citizen in a democracy living through an election campaign. The choices the candidates collectively make about what campaign strategies they want to follow determine the campaign environment available to the voter. History shapes the electoral context as well: a longtime incumbent will be better known (for better or worse) at the outset of a campaign, while a candidate mounting his first campaign starts with almost a blank slate. But the citizen still has a great deal of control over how much of that campaign they wish to experience. Some people actively seek out as much information as possible about all candidates running, others do everything they can to avoid anything vaguely political, while still others are so busy taking care of their families that they just don't have much time for anything else. Some people (roughly forty percent in the U.S.) know how they are going to vote before the campaign even begins; others (another forty percent, more or less) know that they will not bother voting. The remainder – approximately twenty percent – will often vote given the added stimulation during presidential elections, but probably will not bother to vote during any less intense political campaign. How they will vote is much more uncertain and potentially open to influence from the campaign.

These two sets of factors – the choices made by candidates at the outset of a campaign (which can be modified during the course of an extended campaign), and the choices made by voters during the course of the campaign – are what make it so challenging to study the effects of actual political campaigns. Candidate X spends all available resources to convince as many citizens as possible to vote for him while Candidate Y is simultaneously doing everything she can think of (and afford) to counter what Candidate X does and to convince those same citizens

to vote for her, while many citizens are either totally oblivious of anything political going on around them, or are aware of an ongoing campaign but are doing their best to avoid it. There is little wonder that political scientists might try to eliminate many of these complications by turning to experiments to study the effects of candidate advertisements.

This chapter will review the experimental literature on the effects of candidate advertisements, primarily on vote choice and candidate evaluation.ⁱⁱ By far the largest number of studies has focused on one question: the effectiveness of negative, as opposed to positive, political ads. This same question has motivated the great bulk of the nonexperimental studies as well. We can use this question to illustrate the difficulty of trying to determine the effectiveness of different campaign strategies by studying real political campaigns. We hope that researchers studying political campaigns will begin systematically studying many aspects of political campaigns in addition to their tone, but for now, by necessity, we are mostly limited to explorations of this one specific question.

Experiments overcome the most vexing difficulties, but they do so by creating an artificial situation that could, in several important ways, crucially misrepresent (or limit) the very phenomenon the scientist is trying to study. At the very least, all researchers studying campaign advertisements experimentally face a number of “practical” challenges that determine the nature of the experiment they run, affect the causal inferences that can be drawn from the research, and influence the breadth of situations to which the findings can be generalized. We conclude this chapter by discussing a number of important questions about candidate advertisements that could be explored experimentally but so far have not been sufficiently addressed.

1. Methodological Difficulties in Studying Real Campaigns through Observational Methods

Fifty years ago when social science research on media effects was in its infancy, Carl Hovland (1959) noted a “marked difference” in the picture of communication effects obtained from experimental and survey methods, with experiments indicating the possibility of “considerable modifiability of attitudes” through exposure to televised communications, while correlational studies usually find that “few individuals ... are affected by communications” (8). Hovland focused on several important methodological reasons for the different results, including 1) the fact that the audience for many real-world communication efforts such as political campaigns are highly selected, so that the communicator often ends up preaching to a choir that already agrees with the message; 2) actual exposure to the communication is almost guaranteed in the lab, while in the real world even people who are exposed to a candidate advertisement may well ignore it; 3) subjects in experiments typically view communications in social isolation, while any real-world communication efforts are experienced in a social context that usually reinforces prior attitudes and thus resists change; 4) in the laboratory, the dependent variable is typically gathered immediately after the communication, while with observational methods the dependent variable (such as voting) may not occur until days or even months after a communication is delivered; and 5) laboratory research typically studies less involving issues, while observational studies typically focus on more important topics. Over the succeeding fifty years, improved theory and methods have provided numerous examples of observational studies documenting rather substantial effects of political communication (see Kinder 2003) so that Hovland’s starting point is no longer true, but these methodological issues continue to describe important differences between experimental and observational studies of candidate

advertisements. Indeed, the advent of television remote controls, the explosion of the number of cable channels, and the Internet have made Hovland's first issue of discretionary exposure to candidate messages more problematic today than it was fifty years ago.

We would add a sixth point to Hovland's list, the fact that today's actual candidates can target their messages to particular audiences and usually have a pretty good idea of how receptive that audience will be to the message, while in laboratory experiments we typically randomly expose subjects to different messages. It is difficult to imagine candidates making decisions about campaign strategy without some idea of what their chances are of winning the election, whom the opponent is likely to be, and what resources they will have in order to accomplish their electoral goals. The conventional wisdom about negative advertising is that it is effective but risky. It can quickly and relatively inexpensively lower evaluations of the target of the attacks, but may result in a backlash that lowers evaluations of the sponsor of the attacks as well, leaving the net benefit somewhat up in the air (Lau, Sigelman, and Rovner 2007). But candidates who expect to lose or who find themselves behind in a race, and candidates with fewer resources than their opponent, have few viable options except to attack. This means that negative advertising is a strategy often chosen by likely losers, which makes it difficult to determine the effectiveness of negative campaigning – or any other campaign strategy that candidates choose with some knowledge of the likely campaign outcome. Did a candidate lose an election because they chose to attack their opponent, or did they choose to attack their opponent because they were going to lose anyway and no other strategy gave them a better chance of reversing their fortunes? Maybe they would have lost by more votes had they chosen some other campaign strategy. Figure 15-1 lists these six problems and their methodological

consequences.

[Figure 15-1 about here]

In statistical parlance, the problem is that choice of campaign strategy is endogenous to the likely outcome of the election. Some unmeasured (unexplained) portion of the dependent variable is related to unmeasured portions of the independent variable, and this correlation violates one of the basic assumptions of regression analysis. The statistical solution is to find one or more “instrumental variables” that are related to the problematic independent variable (campaign strategy) but are not related to the dependent variable. This is a tall order, and the results are only as good as the quality of the instruments that can be found (Bartels 1991). But if reasonable instruments are available, they are used, along with any other independent variables in the model, to predict the problematic independent variable in a first stage regression. The predicted scores from this first stage regression, which have been “purged” of any inappropriate correlation with unmeasured aspects of the dependent variable, are then used to represent or stand in for the problematic independent variable in a second stage regression.

In their studies of the effect of negative advertising, Ansolabehere, Iyengar, and Simon (1999) and Lau and Pomper (2004; see also 2002) provide good examples of this procedure, and because they are based on an analysis of virtually every contested Senate election across multiple election years, the external validity of the findings are difficult to challenge. At the same time, one could question exactly what has been learned from these studies. Campaigns are dynamic events and are usually observed over a period of time. Both Ansolabehere et al. and Lau and Pomper relied on American National Election Studies (ANES) survey data in their individual level analysis, and thus implicitly observed those campaigns during the period the ANES was

interviewing (typically the two months before the November elections). They have one estimate of the tone of each candidate's campaign over this entire eight week period (from each respondent's memory, in Ansolabehere et al.; from a coding of newspaper accounts of the campaigns, in Lau and Pomper). But this by necessity treats as identically positive (or negative) a huge variety of different campaign ads and themes and statements that might have very different effects.ⁱⁱⁱ Similarly, because they did not have enough data to reliably measure campaign tone on a daily or even weekly basis, these authors again implicitly assume that tone has the same effect across that entire final two months of a campaign – a dubious assumption at best. Most of these limitations are not inherent in observational methods, but in practice they put very real constraints on what can be learned from any such study.

Candidate advertisement experiments can resolve some of the inferential difficulties in estimating the effects of ads. Experiments can avoid the endogeneity of campaign strategy that makes determining the causal direction of effects difficult in observational studies. Researchers can also more precisely estimate the effect of the ad itself on turnout or candidate evaluation without needing to adjust for voter characteristics that may drive both actual ad exposure and the political outcomes researchers care about because experimental exposure to candidate ads is randomly assigned rather than determined by the citizens' personal characteristics, such as political interest, past voting behavior, and state of residence. Lastly, experiments allow researchers to test hypotheses about how ads affect behavior in ways that prove difficult to isolate using purely observational methods.

Experiments solve a number of issues raised by observational studies but also come with their own limitations. Political campaigns involve at least two candidates vying to persuade

voters to support their candidacy, use multiple means of persuasion based on the closeness of the race and candidate resources. But experiments by their very nature lend themselves more naturally to studying discrete events – single ads rather than comprehensive ad campaigns. Even the most comprehensive experiments that create a campaign environment (Lau and Redlawsk 2006) cannot completely recreate the “blooming, buzzing” chaos that is a political campaign, and thus the experimental environment is a simplification of the real-world campaign environment. While experiments can isolate how campaign ads affect the public, they may over or underestimate the effect of these ads in a full information campaign environment. The effect of any one particular ad may be washed away in a real campaign by the cumulative impact of candidate visits or debates, and thus experiments that only include single ads may overestimate how ads may affect the public. Yet if the impact of ads depends on relevant policy information or the personal characteristics of the candidates within a campaign that are absent from a more controlled but sterile experimental environment, then using experiments to isolate how ad tone or content affects evaluations may underestimate their effects.

A second potential limitation of studying candidate ads experimentally is that most experiments occur at a single point in time and do not follow up with respondents afterward. Researchers can measure the short-term effect of ads within the span of the experiment, but it is not entirely clear whether the effects captured during an experiment are long lasting or a short-term reaction to the experimental stimuli. Additionally, it is difficult to measure the duration of ad effects within the limited time constraints of most experiments.^{iv} Nor do we know what the cumulative effects of a political ad on attitudes might be (Gerber et al. 2007; Chong and Druckman 2009).

However serious these concerns are, though, the benefit of using experiments is that they clearly establish the causal effect of an ad on dependent variables of interest. Establishing this causal effect may be the first step to observing the effect in a broader political context. In other words, if researchers cannot show that a campaign ad affects candidate evaluations or turnout within an experiment, it seems quite unlikely that they will be able to observe an effect of that same ad using observational methods. We use the experimental literature on negative advertising to review how experiments determine what we know about the effects of campaign ads.

2. Negative Ads and the Likelihood of Voting

A politician's decision to produce and run negative ads – ads that portray an opponent's positions as wrong or that cast doubt on an opponent's character – may influence voters to support the sponsor of the attack and thus increase the probability of turning out to vote. Yet negative ads may also backfire and lower voters' probability of voting for the attack sponsor and serve to demobilize the public. In this section, we explore experimental findings on how candidate advertising affects the probability of voting.

In the paradigmatic study of demobilization, Ansolabehere et al. (1994) demonstrate that exposure to a single negative ad embedded in a fifteen minute newscast decreased experimental subjects' intention to turn out in the next election by five percentage points compared to respondents who saw a positive ad with the same audiovisual script. The many strengths of this experiment include 1) the treatment conditions varied on only two dimensions (tone and sponsoring candidate) while being identical on all other dimensions (visuals, voiceover, issue focus); 2) the studies occurred during ongoing political campaigns; and 3) the experimental setting approximated the home environment where ads might actually be viewed. By varying

only the tone between the treatment and control conditions, the authors had much greater control over the nature of the experimental treatment and were able to conclude that it is the negativism of the ads per se that decreases respondents' intention to turn out. Additionally, by setting the experiments within an actual campaign, the researchers could choose salient ad content and use real candidates, increasing the realism of the lab experience.

In the aforementioned example, the experiments tested the effect of a single candidate ad on candidate evaluations, but very rarely do voters receive only one-sided information in a campaign. Whether negative ads mobilize or demobilize may depend on how many negative ads voters see or hear. Using an experiment with 10,200 eligible voters in the Knowledge Networks panel during the 2000 election, Clinton and Lapinski (2004) varied three factors to test whether negative ads mobilized or demobilized the public: how many ads respondents saw (one versus two ads), whether the ads came from Bush or Gore, and whether the ads were positive or negative. Overall, the authors found no evidence that negative ads systematically increased or decreased voters' probability of turning out, suggesting that exposure to one or two ads in the middle of an ongoing campaign may not systematically affect the decision to vote. Clinton and Lapinski's use of multiple ad treatments that varied tone and the source of the negative ad more closely mimic real-world campaigns than do experiments with single ads, although by using actual candidate ads their tone manipulation inevitably varied more than simply tone.

Krupnikov (2009) argues that exposure to negative advertising may be demobilizing for some voters depending on the timing of exposure. She argues that campaign negativism only affects turnout when exposure comes after a voter selects a candidate but before he can implement the voting decision – a hypothesis that, in practice, would be very difficult to test with

observational methods. In an online experiment using a representative sample of Americans, respondents who received a negative ad after choosing a favored candidate were six percent more likely to say that they would put no effort into turning out than those who received the negative message before choosing their favorite candidate, suggesting a relatively strong demobilization effect but only for some respondents. It is worth noting, however, that the experimental design utilizes candidates devoid of names, written rather than audiovisual treatments, and ads of similar length but different policy areas – and perhaps some of these other differences rather than the timing of ads per se may affect the decision to turn out. Overall, the Ansolabehere et al. experiments provide the most controlled test of the hypothesis that negativity demobilizes the electorate, but other studies' use of multiple ads more convincingly proxy the dynamics and complexity of real campaigns.

So do negative ads demobilize or mobilize the electorate? The results across observational and experimental studies are mixed. A meta-analysis of fifty-seven experimental and observational studies of the so-called demobilization hypothesis demonstrated no consistent effect of negative advertising on turnout (Lau et al. 2007), which suggests that negative ads may matter for a variety of reasons that vary over different campaigns and/or for certain individuals at different times. There may be at least three reasons why the turnout findings differ across experimental and observational studies: 1) cumulative effects, 2) timing, and 3) selective versus broad exposure. Most experimental studies of advertising consider the effects of only one or two negative ads on turnout, although well-funded campaigns can run multiple ads per day for weeks up to an election. Ansolabehere et al.'s experiments looked at the effect of exposure to a single ad for one candidate within campaigns that actually featured a high volume of competing ads. To

the extent that the effects identified in the lab are cumulative, then exposure to multiple negative ads should strengthen the demobilizing effects identified. But if exposure to competing messages with varying tones could cancel each other out, or if there is a relatively low ceiling effect for exposure to repeated attack ads, one-shot experimental designs cannot identify these longer-term effects. Observational studies are better suited to pick up the effect of advertising volume on turnout, but tone measures for individual ads are more difficult to obtain. The timing of advertising exposure may influence turnout in ways that experimental designs and cross-sectional observational studies may account for differently. Experiments are not typically run during campaigns, while observational studies tend to occur right before elections. To the extent that voters are influenced by different factors over the course of a campaign, for example, after they have made up their own minds (Krupnikov 2009) or when it looks like their favored candidate is going to lose, then how far in advance of an election a study is done and the competitiveness of the race may determine whether voters decide to turn out at all. Lastly, negativism of any form is uncomfortable for some voters so, in the real world, those voters may ignore negative ads, meaning that studies may only pick up the effect of attack ads on more interested or less sensitive voters. Yet, when forced to confront negative ads in an experiment, these same voters may be turned off by the negativism and want to disengage from politics. If this is particularly the case for uncommitted voters, then an experiment may find an overall demobilization effect even when these voters would never encounter these ads in a real campaign. While any of these possibilities may explain the differences between observational and experimental research, if there is one crucial factor that determines whether ad exposure will mobilize or demobilize, researchers have yet to identify it.

3. Negative Ads and Candidate Evaluation/Vote Choice

Presumably, candidates decide to use negative ads because they believe that the ads will either decrease the likelihood of voting for the opposing candidate or at least lower the public's evaluations of the opposing candidate. Yet, experimental research on the effects of ads demonstrates that while negative ads may decrease evaluations of the target of the attack, negative ads may have a variety of other consequences including: 1) a backlash against the attacking candidate who loses popularity as a result of sponsoring the attacks (Matthews and Dietz-Uhler 1998), 2) a "victim syndrome" where the target of the attack actually becomes more favorably viewed after a negative ad (Haddock and Zanna 1997), or 3) a "double impairment" where both the source of the negative ad and the target are viewed more negatively.

How negative ads affect candidate evaluation may depend on the relationship between the target of the ad and the voter. Negative ads may reinforce partisan loyalties and affect those who share the partisanship of the ad sponsor (Ansolabehere and Iyengar 1995) or alternatively, negative ad exposure may lead to what Matthews and Dietz-Uhler (1998) call a "black sheep effect" whereby negative ads that come from a liked group cause respondents to downgrade a liked candidate (one who shares the voter's partisanship), while negative ads that come from a less liked candidate do not have this effect. In Matthews and Dietz-Uhler's (1998) experiment, 123 undergraduates read either a positive or negative mock advertisement about family values that was said to be sponsored by either an in-group (same party) or out-group candidate. An in-group sponsor of a positive ad was evaluated more positively than any out-group member, regardless of the type of ad. However, an in-group sponsor of a negative ad was evaluated more negatively than either an in-group sponsor with a positive message or an out-group sponsor of

either type of ad.

Does this black sheep effect apply to all types of in-groups? Schultz and Pancer (1997) found that when the in-group/out-group characteristic is gender rather than party, the black sheep effect is absent. In their experiment, 134 students read positive or negative statements about the integrity and personality of female or male candidates' opponents. When judging a candidate of their own gender, subjects rated the candidate as having greater integrity when the candidate attacked his/her opponent than when she did not. Yet, when judging a candidate of the opposite gender, participants tended to rate the candidate who attacked his/her opponent as having less integrity than those who did not attack. It may be the case that gender is not a strong or salient enough political characteristic to make subjects feel bad when a group member attacks another candidate. Alternately, undergraduate samples may be particularly aware of social norms about gender equality and wish to reflect these norms by not judging women candidates differently than male candidates. While either of these explanations is a possibility, whether a "black sheep" effect occurs broadly or only with partisanship is an open question. The relative dearth of actual female candidates and their tendency to be disproportionately Democratic makes these competing explanations almost impossible to tease apart with observational methods, but it would be a relatively simple matter to simultaneously vary both candidate gender and partisanship in an experiment to reconcile these contradictory findings.

4. Competing Messages

Experiments that consider single ads in isolation may miss the dynamics that occur in real campaigns, when competing candidates provide contending considerations that complicate how voters evaluate candidates. Clinton and Owen (2009) use a large-scale Knowledge Networks

experiment during the 2000 election to test the effect of competing Bush and Gore ads on how decided and undecided subjects make a candidate choice. Respondents with an initial predisposition toward a candidate solidified their vote intention after viewing the competing ads, but undecided subjects were less likely to designate a candidate preference after ad exposure, suggesting that competing messages may inhibit some voters from making a vote choice.

In a study of 274 undergraduates, Roddy and Garramone (1988) test the impact of the type of negative ad respondents see (issue versus image) and the tone of a response to the attack (positive versus negative). The authors created positive and negative television ads for two fictional candidates that focused either on the candidate's issue positions or character. Each respondent saw one of the negative ads paired with either a negative or positive response ad from the candidate targeted by the initial attack. Roddy and Garramone find that when candidates strike first, negative issue ads significantly increase evaluations of the sponsor's character and decrease the probability of voting for the target of the attack more than negative image ads do. Yet when the target of an attack responds with a positive ad rather than a negative ad, it significantly lowers the probability of voting for the sponsor of the original negative ad.

During the 1996 presidential campaign, Kaid (1997) conducted an experiment with a total of 1,128 undergraduates that exposed subjects to both positive and negative Dole and Clinton ads at two points during the campaign – late September and early October. Respondents received a mix of four negative and positive ads from the two candidates in the September session and a different four ads in October. In the September round of experiments, partisans increased evaluations for in-party candidates and lowered their evaluation of out-party candidates. These findings suggest that candidate ads affect individuals differentially based on

their partisan identification and that experiments that utilize nonpartisan candidates may overestimate the effectiveness of ads on the entire electorate.

A major benefit of Kaid's experimental design is that it tests the impact of competing candidate ads over time. Kaid shows that the effect of Dole messages changed significantly during the 1996 campaign. In the first wave of the experiment, exposure to Dole ads increased evaluations of Dole among both Republicans and independents but by October, all respondents took a more negative view of Dole after seeing the Dole spots. However, because this experiment was conducted in the context of an ongoing presidential campaign, it is not clear if it is time per se or some other factor confounded with time in the 1996 campaign (i.e., it became increasingly clear that Dole was going to lose) that explains these results.

5. Other Campaign Ad Effects

In a series of experiments, Brader (2005) demonstrates that campaign ads affect voters through appealing to emotions – particularly enthusiasm and fear. In the experiments, respondents saw either positive ads or negative ads as they watched a newscast. Half of the positive ads included enthusiasm cues while half of the negative ads included fear cues; these emotional appeals came from the addition of music and evocative imagery. Brader shows that ads affect voters by appealing to their emotions but do not simply move voters toward one candidate or away from another. Rather, emotions affect voters by changing the way that voters make decisions. Fear appeals lead voters to make decisions based on contemporary information, while enthusiasm appeals encourage decisions based on prior beliefs. This study represents a growing body of experimental work that examines not only how ads affect the public but the underlying psychological mechanisms of those effects.

6. Practical Challenges in Designing Experiments on Candidate Advertisements

Any researcher adopting experimental methods to study the effects of candidate advertisements faces a number of practical challenges that have to be addressed in designing the research. In the following section, we explicate some practical challenges that scholars face.

Actual Ads or Ads Created Just for the Experiment?

One of the first questions any researcher planning an experiment must decide is whether to study ads actually produced and utilized by a candidate in a real campaign, or to employ ads created explicitly for the experiment. “Borrowing” ads from a real campaign is cheap and easy and has a great deal of external validity, but researchers give up a good deal of control by employing real political ads. Many combinations of strategies that we would like to study experimentally simply do not occur in practice. The obvious solution to this problem is to run an experiment and create your own political ads, randomly assigning different campaign strategies to different candidates. The researcher gets a specifically desired type of ad, but with the added expense of having to create the ad and the commensurate loss of external validity resulting from employing ads that were not created by an actual candidate. A compromise solution is to employ the video from an actual ad (thus claiming some degree of external validity) but manipulate the “voice-over” that accompanies the video, so that all other aspects of the manipulated ad are identical except for the verbal message.

Real Candidates or Mock Candidates?

Another problem with making inferences from real elections is that many candidates (for example, all incumbents) are familiar to many voters from earlier elections. Thus any new information that might be learned about a familiar candidate during the course of a campaign is

interpreted against a background of whatever prior information a citizen has stored about that candidate in memory. Again, the experimental solution is obvious: use candidates that subjects are not familiar with, either by “creating” your own mock candidates, or by using real candidates from a far-away state with whom few subjects are familiar. But then any knowledge we gain as a result of this experiment should be limited to situations where the candidates are new and unfamiliar – either open seat races without any incumbent, or lower level offices where the candidates are just starting their political careers. A lot of elections are like this and there is nothing wrong with studying them. We suspect, however, that most experimentalists utilizing such a design do not imagine that they are primarily studying open seat or lower level elections.

Include Party Affiliation or Ignore It?

Admittedly, a large store of information about an incumbent politician is surely limited to the most politically interested subset of the population. But for many people, a candidate’s party affiliation substitutes for a great deal of more specific information. However, if we include party affiliation in the description of a stimulus candidate, a subject’s own party identification could largely determine candidate evaluations independent of whatever advertisements from that candidate are shown, thus greatly reducing the power of any experimental manipulation to detect significant effects. McGraw’s chapter in this volume raises similar concerns. Avoid party affiliation in the description of a stimulus candidate and it will be much easier to push around evaluations of that person. But would whatever we learn from such an experiment be generalizable only to nonpartisan elections? Surely that would not be the intent of most researchers. Indeed, how campaign ads are perceived and interpreted depends crucially on whether the ad comes from my guy or their guy (Lipsitz et al. 2005). An additional complication

that arises when experiments do not utilize party labels is that respondents may use other cues, such as gender (King and Matland 2003) or personality traits (Rapoport, Metcalf, and Hartman 1989), to infer candidates' partisanship and policy positions. Thus, experimental effects that researchers associate with these other characteristics may actually be a function of candidates' partisanship even when partisanship is absent.

College Sophomores versus Real People as Subjects?

Another question that any researcher employing experimental methods must face is where to get subjects. The easiest and cheapest solution for many (academic) researchers is undergraduates – a young, bright, captive, and generally compliant population. For many topics social scientists study (such as basic cognitive processes), undergraduates are perfectly good subjects. But as Sears (1986) warned, undergraduates are a very homogenous population in terms of their age, life experience, education/intelligence, heightened political awareness, and “less crystallized social and political attitudes” (Sears 1986, 22); this homogeneity may restrict our ability to study factors that could be of great interest to political campaigns. Druckman and Kam's chapter in this volume discusses methods that may greatly increase the generalizability of results produced by student samples, and we do not make strong recommendations about whether campaign ad experiments should choose adult samples over student samples. Whether researchers wish to utilize a broad or narrow database (Sears 1986) should be determined by their theoretical expectations, in addition to practical considerations about the expense and the difficulty of subject recruitment.

It is not impossible to bring “real people” into a laboratory, but it does add considerably to the financial costs and time required to run an experiment. To recruit nonstudent experimental

samples, researchers can include college employees in the sample pools of on-campus labs (Kam, Wilking, and Zechmeister 2007), recruit community members through newspaper advertising, flyers, or Internet message boards (Brader 2005; Lau and Redlawsk 2006), bring the experiment to subjects face-to-face (Ansolabehere and Iyengar 1995; Mendelberg 2001) or use online panels. High quality online samples like those in the YouGov/Polimetrix or Knowledge Networks panels may provide a less expensive way to include adults in experimental samples than sending research assistants to the homes of respondents or paying “real people” to come to the lab. However, when experimentalists rely on online experiments, this inevitably shortens the duration of an experiment and may limit the types of candidate ad experiments that are possible. Shorter online experiments are more common than longer, more in-depth experiments that may allow researchers to embed candidate commercials within newscasts or otherwise make the experimental treatment less obvious or include multiple ads as treatments. Online subjects may feel less pressure to comply with directions and may be more likely to opt out of the experiment by switching to something in their home or on their computer that is more interesting. While online panels can provide more representative samples of the voting public than can on-campus labs, online experiments may be only a partial solution to concerns over external validity if studies themselves are more limited.

Experiment During Campaigns versus Outside of Campaigns?

Another decision facing researchers is whether to run experiments during the course of an ongoing campaign or at another time. Running studies during actual campaigns may increase external validity by utilizing real candidates and salient issues, but there are practical and ethical challenges to waiting for a campaign to begin. On the practical side, limiting studies to campaign

season significantly limits how often researchers can field studies and requires tight coordination with the scheduling of pre-testing and human subjects approval. On the ethical side, when researchers use real candidates during an actual campaign, there is a possibility of affecting candidate evaluations and possibly vote choice with treatments that may or may not reflect candidates' actual policy or personal positions. Respondents do often forget treatments relatively soon after exposure, and debriefing should lower concerns that respondents will think that the treatments are real. However, we know of few studies that follow up with respondents to verify that treatments do not have long-term consequences. Additionally, to the extent that experimental subjects utilize online processing rather than memory-based processing, experiments may affect political attitudes and behaviors even if respondents forget the details of treatments, which may be particularly troubling during the course of a campaign.

7. Conclusion

As experimental research on campaign ads continues to increase in popularity in the discipline, we should consider whether Hovland's critiques of experimental studies of media effects still hold or whether the current literature has overcome these concerns. It is worth noting that many of Hovland's complaints against the external validity of experimental studies of media effects (unrealistic exposure, the social isolation of the experimental subject, the probable but unknown ephemerality of the demonstrated effects, studying trivial and less ego involving issues) are not inherent in the experimental method, but the choices that researchers make when designing experiments may either alleviate or exacerbate these issues. While some current studies directly address issues such as the ephemerality of campaign effects (Clinton and Lapinski 2004; Chong and Druckman 2009) and utilize real candidates to make studies more ego

involving (Kaid 1997; Clinton and Owen 2008), several of the other critiques are overlooked in the current literature and thus provide opportunities for new -- and very important -- scholarship.

In a world of increasing media choice, the problem of audience selectivity is probably more serious now than in Hovland's era and may make demonstrating effects of campaign ads more difficult in observational studies than in experimental ones. That is, in the real world, less interested or less partisan citizens can tune out, turn off, or altogether avoid candidate ads whereas in the lab some percent of subjects receive a political message that they would never otherwise receive. The advent of new technologies and collection of new data that more precisely track when and how many times campaign ads air (Franz et al. 2007) can help both observational and experimental researchers estimate when campaign exposure is probable and what types of individuals are likely to be most affected by exposure. We know of few experimental studies that contend with the issue of selectivity or allow respondents to opt out of experimental treatments (Arceneaux and Johnson 2007; Gaines and Kuklinski's chapter in this volume), but with a more complex media environment, questions about the types of citizens who may be affected by campaign communications may suggest more complex experimental designs in the future.

Political campaigns take place in a social context. Voters do not experience campaigns in social isolation, but rather they talk about events and ads with friends, family, and coworkers, many of whom share their political identities. The makeup of one's social network may affect how ads affect candidate evaluation or vote choice. When voters are surrounded by like-minded compatriots, they may be less likely to accept counter-attitudinal arguments than when faced with a more heterogeneous social network. Yet experiments of campaign ads rarely take these

types of social dynamics into account, confirming Hovland's concerns over social isolation. Unlike social psychology experiments that often introduce elements of group discussion and decision making, most campaign experiments take place at the individual level and include only interactions between experimenter and subject. While researchers lose an element of control, experimenters could invite subjects into the lab in groups rather than individually or could snowball sample respondents from existing social networks. With political campaigns' increasing use of social networking sites such as Facebook, the issue of how social networks condition the effect of campaign communication will be an increasingly important issue for researchers to consider.

Hovland's last several concerns about the external validity of experiments, ephemerality and ego-involvement, are not completely diminished by current research, but several studies do directly address these concerns. Again, there is nothing inherent in the experimental method that demands that scholars design one-shot studies that measure reactions to ads or candidate evaluations immediately after ad exposure. Rather, for practical and financial reasons, many researchers choose not to follow up with subjects later to test how long campaign ads may affect evaluations or intended vote. Like Clinton and Lapinski's (2004) study with Knowledge Networks, studies utilizing online panels allow following up with subjects more easily than bringing people physically back into the lab. Without following up with subjects, it is unclear whether ephemerality of experimentally induced effects should concern political scientists or not. Additionally, we believe that the concern that experiments often involve trivial or less ego involving issues is less serious in campaign experiments, but particularly less worrisome in studies that use real candidates, since partisanship provides information about the candidate and

acts as a strong affective tie between subject and candidate (Kaid 1997; Clinton and Lapinski 2004; Brader 2006; Clinton and Owen 2008).

Campaigns are strategic in their use of resources (Shaw 2007) and are now more able to target particular audiences with different messages or to microtarget likely voters. Without good observational studies that describe when and why campaigns deploy different campaign ad strategies like going negative versus staying positive, experimental studies may produce political situations that do not occur in reality, thus harming external validity. Shaw's study of the 2000 and 2004 presidential campaigns illuminates the strategic logic of where and when presidential candidates advertise, but we know of few studies that outline other strategies employed by candidates. Because campaign strategy is endogenous to the political situation candidates find themselves in, experimental researchers should seriously consider when they would expect to see the results that they demonstrate in their studies.

Experimentation provides researchers significant control over design in order to test causal claims about, for example, whether negative advertising demobilizes voters. Experiments are invaluable in demonstrating the mechanism at work. As with all research, however, experiments require tradeoffs between what types of questions researchers can answer and how to create environments that, on some level, resemble real elections. Yet despite these challenges, experimental research on candidate advertising illuminates a great deal about how voters contend with ads in forming evaluations of candidates and deciding who to choose at election time.

References

Ansolahehere, Stephen, and Shanto Iyengar. 1995. *Going Negative: How Attack Ads Shrink and Polarize the Electorate*. New York: Free Press.

- Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88: 829-38.
- Ansolabehere, Stephen, Shanto Iyengar, and Adam Simon. 1999. "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout." *American Political Science Review* 93: 902-09.
- Arceneaux, Kevin, and Martin Johnson. 2007. "Channel Surfing: Does Choice Reduce Videomalaise?" Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Bartels, Larry. 1991. "Instrument and 'Quasi-Instrumental' Variables." *American Journal of Political Science* 35: 777-800.
- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49: 388-405.
- Chong, Dennis, and James Druckman. 2009. "Dynamic Public Opinion: Framing Effects over Time." Working Paper, Northwestern University.
- Clinton, Joshua, and John Lapinski. 2004. "'Targeted' Advertising and Voter Turnout: An Experimental Study of the 2000 Presidential Election." *Journal of Politics* 66: 69-96.
- Clinton, Joshua, and Andrew Owen. 2009. "An Experimental Investigation of Advertising Persuasiveness: Is Impact in the Eye of the Beholder?" Working Paper.
- Franz, Michael M., Paul Freedman, Kenneth M. Goldstein, and Travis N. Ridout. 2007. *Campaign Advertising and American Democracy*. Philadelphia: Temple University Press.
- Gerber, Alan, James Gimpel, Donald Green, and Daron Shaw. 2007. "The Influence of Television and Radio Advertising on Candidate Evaluations: Results from a Large Scale Randomized Experiment." Working paper, Yale University.
- Haddock, Geoffrey, and Mark Zanna. 1997. "Impact of Negative Advertising on Evaluations of Political Candidates: The 1993 Canadian Federal Elections." *Basic and Applied Social Psychology* 19: 205-23.
- Hovland, Carl. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change." *American Psychologist* 14: 8-17.
- Kaid, Lynda Lee. 1997. "Effects of the Television Spots on Images of Dole and Clinton." *American Behavioral Scientist* 40: 1085-94.
- Kam, Cindy, Jennifer Wilking, and Elizabeth Zechmeister. 2007. "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research." *Political Behavior* 29:

415-40.

- Kinder, Donald R. 2003. "Communication and Politics in the Age of Information." In *Oxford Handbook of Political Psychology*, eds. David O. Sears, Leonie Huddy, and Robert Jervis. New York: Oxford University Press.
- King, David C., and Richard E. Matland. 2003. "Sex and the Grand Old Party: An Experimental Investigation of the Effect of Candidate Sex on Support for a Republican Candidate." *American Politics Research* 31: 595-612.
- Krupnikov, Yanna. 2009. "Who Votes? How and When Negative Campaigning Affects Voter Turnout." Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Lau, Richard R., and Gerald Pomper. 2002. "Effectiveness of Negative Campaigning in U.S. Senate Elections, 1988-98." *American Journal of Political Science* 46: 47-66.
- Lau, Richard R., and Gerald Pomper. 2004. *Negative Campaigning: An Analysis of U.S. Senate Elections*. Lanham, MD: Rowman Littlefield.
- Lau, Richard R., and David Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. New York: Cambridge University Press.
- Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics* 69: 1176-209.
- Lipsitz, Keena, Christine Trost, Matthew Grossman, and John Sides. 2005. "What Voters Want from Campaign Communication." *Political Communication* 22: 337-54.
- Lodge, Milton, Marco R. Steenbergen, and S. Brau. 1995. "The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89: 309-26.
- Mathews, Douglas, and Beth Dietz-Uhler. 1998. "The Black-Sheep Effect: How Positive and Negative Advertisements Affect Voters' Perceptions of the Sponsor of the Advertisement." *Journal of Applied Social Psychology* 28: 1903-15.
- Mendelberg, Tali. 2001. *Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Rapoport, Ronald, Kelly Metcalf, and Jon Hartman. 1989. "Candidate Traits and Voter Inferences: An Experimental Study." *Journal of Politics* 51: 917-32.
- Roddy, Brian, and Gina Garramone. 1988. "Appeals and Strategies of Negative Political Advertising." *Journal of Broadcast & Electronic Media* 32: 415-27.

- Sears, David O. 1986. "College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51: 515-30.
- Shaw, Daron. 2007. *The Race to 270: The Electoral College and the Campaign Strategies of 2000 and 2004*. Chicago: University of Chicago Press.
- Shultz, Cindy, and S. Mark Pancer. 1997. "Character Attacks and their Effects on Perceptions of Male and Female Political Candidates." *Political Psychology* 18: 93-102.

Figure 15-1. Methodological Consequences of Differences Between Observational and Experimental Studies of Candidate Advertisements

Problem	Methodological Consequence
Audience Selectivity	Politicians preach to the choir, minimizing possibilities of further opinion change in observational studies
Unequal/Uncertain Exposure	Harder to document opinion change with observational studies, experiments/lab studies overestimating potential effects
Social Influence	Prior attitudes more resistant to change in the real world, experiments/lab studies overestimating potential effects
Ephemerality	Dependent variable usually measured immediately after treatment in experiments, which maximizes apparent treatment effects from lab studies
Ego-Involvement	Opinions on less ego involving issues are easier to change, but such issues have less real-world importance and the absence of relevant data usually makes them difficult to study with observational methods
Endogeneity of Campaign Strategies	Causal inferences extremely difficult with observational methods, experiments can simulate campaign situations that simply do not occur in actual campaigns

Note: The first five problems listed in Figure 15-1 are discussed by Hovland (1959).

ⁱ We thank Kevin Arceneaux and the book's editors for comments on earlier drafts of this chapter.

ⁱⁱ We will concentrate only on ads from candidates running for office and will ignore interest group ads.

ⁱⁱⁱ Lau and Pomper (2002, 2004) did distinguish between issue- or policy-based statements and person-based

statements (still at best a very gross distinction), but did not find very many differences between the two.

^{iv} While we may be interested in how long the experimental treatments last, it is worth noting that it is unclear how long campaign messages more broadly last. Lodge, Steenbergen, and Brau (1995) suggest that the half life of campaign messages is typically less than a week. We have no reason to believe that the effects of campaign ads are substantially different than other types of campaign messages. One concern is that experiments may overstate how effective ads are in shaping candidate evaluations because subjects often forget which ad they saw even at the end of a relatively short experiment. Yet we agree with Lodge et al.'s contention that "recall is not a necessary condition for information to be influential" (1995, 317-8), meaning that even if experimental subjects later forget their exposure to a particular ad, it may continue to affect candidate evaluation and vote choice.

16. Voter Mobilization

Melissa R. Michelson and David W. Nickerson

Civic participation is an essential component of a healthy democracy. Voting allows citizens to communicate preferences to elected officials and influence who holds public office. At the same time, deficiencies and asymmetries of participation in the United States call into question the representativeness of elected officials and public policies.ⁱ Yet, while political activity is crucial for the equal protection of interests, participation is often seen by individuals as irrational or excessively costly, and it is well known that turnout in the U.S. lags well behind that of other democracies. Scholars have consistently found that participation is linked to socioeconomic variables, psychological orientations, and recruitment. Candidates, parties and organizations thus spend considerable effort mobilizing electoral activity. This chapter highlights contributions made by field experiments to the study of voter mobilization, as well as the problems faced by such work, and opportunities for future study.

1. Observational Studies

Nonexperimental studies have primarily relied on survey research to demonstrate correlations between self-reported mobilization and various civic-minded behaviors, while also controlling for various demographic characteristics (for example, age, education, and income) that are known to be significant predictors of turnout.ⁱⁱ The conclusion usually reached is that mobilization efforts are generally effective (for example, Rosenstone and Hansen 1993; Verba, Schlozman and Brady 1995). However, four major empirical hurdles render this conclusion suspect.

First, campaigns strategically target individuals likely to vote, donate money or volunteer, thereby creating a strong correlation between the behavior or attitude to be studied and campaign contact. Because contacted individuals are more likely to participate than noncontacted individuals – even in the absence of mobilization – strategic targeting causes researchers to overestimate the effect of mobilization. That is, observational samples use an inappropriate baseline for comparison.

Second, individuals who are easier to contact are also more likely to vote. Arceneaux, Gerber, and Green (2006) analyze experimental data as if they were observational by matching contacted individuals in the treatment group to people in the control group with exactly the same background characteristics. They find that matching overestimates the effect of mobilization. Matching fails to account for unobserved differences between treatment and control subjects (for example, residential mobility, health, free time, mortality, or social behavior), leading to inflated estimates of the power of the treatment. Thus, the treatment of interest (that is, mobilization) is likely to be correlated with unobserved causes of participation.

A third drawback to survey-based research is that respondents often exhibit selective recall. Politically aware individuals are more likely to report contact from campaigns and organizations because they pay more attention to political outreach and are more likely to place the event into long-term memory (Vavreck 2007). Since politically interested people are more likely to participate, the correlation between mobilization and behavior could be a function of selective memory. Thus, the key independent variable in observational studies relying on self-reported campaign contact is likely to suffer from measurement error.

Finally, survey questions used to collect self-reported campaign contact offer categories too coarse to estimate treatment effects. Standard survey questions tend to treat all forms of campaign contact as identical. For example, the American National Election Studies (ANES) item on mobilization asks, “Did anyone from one of the political parties call you up or come around and talk to you about the campaign? (IF YES:) Which party was that?” Yet, experiments using various types of outreach clearly show that the method and quality of mobilization matters, generating turnout effects that range from negligible to double digits (Green and Gerber 2008). Coarse, catch-all survey measures obscure the object of estimation by lumping heterogeneous forms of campaign contact together.

Controlled experiments directly solve each of these problems associated with observational studies. Random assignment eliminates selection problems and constructs a valid baseline for comparison. By directly manipulating the treatment provided to the subjects, researchers can avoid relying on overbroad survey questions and the vagaries of self-reported behavior. Field experiments using official lists of registered voters can also maximize external validity by including a wide range of subjects and by using official voter turnout records to measure the dependent variable of interest for both the treatment and the control group.

Pioneering Experiments

The experimental literature on mobilization dates back to Gosnell (1927). Following Gosnell, experiments were used sporadically over the next several decades (Eldersveld 1956; Adams and Smith 1980; Miller, Bositis, and Baer 1981). These small-N studies tested a range of techniques (mail, phone, and door-to-door canvassing) and all reported double-digit increases in voter turnout. Unfortunately, these early studies provided biased estimates of campaign contact

by placing uncontacted subjects assigned to the treatment group in the control group. That is, these pioneering studies undercut the analytic benefits of randomization by focusing only on the contacted individuals and turned the experiments into observational studies.

The real flowering of the experimental study of campaign effects came at the turn of the millennium with the 1998 New Haven experiment (Gerber and Green 2000). A large number of subjects were drawn from a list of registered voters and randomly assigned to various nonpartisan treatments (mail, phone, or door) or to a control group. Gerber and Green then had callers and canvassers carefully record whether each subject in the treatment groups were successfully contacted and referenced official records to verify voter turnout for both the treatment and control groups.ⁱⁱⁱ The failure-to-treat problem was addressed by using random assignment as an instrument for contact, thereby providing an unbiased estimate of the effect of contact. The experimental design and analysis disentangled the effect of mobilization from the effects of targeting and selective memory and was very clear about the nature of contact provided to individuals, thereby avoiding measurement error. They concluded that face-to-face contact raised turnout by nine percentage points, mail boosted turnout by half a percentage point, and phone calls did nothing to increase participation.

It is curious that the logic of experimentation took so long to take root in the study of campaigns. Fisher (1925) laid the intellectual groundwork for experiments during the 1930s. There were few technological hurdles to the process since randomization could be performed manually (for example, coin flip) and the analysis could be done through card sorting. Examples of laboratory experiments studying the effects of television advertisements on attitudes and vote intention had been published in leading journals (see Gadarian and Lau's chapter in this volume).

Regardless of the cause of the delay, the last decade has seen an explosion of interest in the use of field experiments to explore voter mobilization, with increasing attention to other facets of electoral campaigns, such as vote choice and campaign contributions. In a meta-analysis of the more than 100 field experiments replicating their initial study, Green and Gerber (2008) conclude that well conducted door-to-door visits generally increase turnout by six to ten percentage points, volunteer telephone calls by two to five percentage points, and indirect methods such as mail generally not at all (with some notable exceptions).

Extensions to the first experiments

The initial studies have been extended in a large number of ways. Some studies have examined previously tested techniques for heterogeneity. One notable contribution followed up on initial findings that commercial phonebanks were generally ineffective, while volunteer phonebanks were usually successful. Nickerson (2007) trained volunteer callers to behave like commercial phonebank staff, giving them quotas of numbers of individuals to reach each shift, while paying commercial canvassers to behave like volunteers, urging them to take their time and engage voters in conversation. The result was a reversal of the general trend: commercial phonebankers trained to act like volunteers were able to move voters to the polls, while rushed volunteers were ineffective. Thus, Nickerson concluded that it was the quality of the phonebank that mattered, not the identity of the canvasser or whether or not canvassers were paid.

Other experiments have examined other campaign tactics for contacting voters, such as radio and television advertisements, leaflets, email, and text messaging. In general, the pattern has been that personalized outreach is more effective than indirect outreach. But there are notable exceptions. For example, Dale and Strauss (2007) find that text messages are effective at

moving young people to the polls. Whether this is a counterexample or evidence that cell phones are considered personal objects is open to debate, and further research is needed to confirm and further explore their findings. Similarly, the effectiveness of television advertisements (Vavreck 2007; Green and Vavreck 2008) may be evidence that not all indirect methods of reaching out to voters are ineffective, or may say something about the power of visual images. Paradoxically, the same rapid growth in field experiments that allowed for precise estimates of the effectiveness of mobilization techniques has also complicated the theoretical picture, necessitating more experiments.

A third line of extensions from the initial New Haven experiment has focused on subpopulations with below average rates of voter turnout. To the extent that low rates of participation bias the electorate, focusing on groups with the lowest rates of voter turnout is a priority. Research in other areas of political science suggests that civic engagement strategies that are effective with Anglos (non-Latino whites) will not necessarily work for African Americans, Latinos and Asians. However, a lengthy series of recent experiments demonstrate that these subgroups generally respond to requests to vote in a similar manner as do high-propensity voters (Michelson, García Bedolla and Green 2007, 2008, 2009). Each population faces its unique challenges, however. The residential mobility of young and poor voters makes them harder to contact (Nickerson 2006). Campaigns targeting Latinos need to be bilingual in most instances and efforts aimed at Asian Americans need to be multilingual. Despite these challenges, field experiments have proved that all of these groups can be effectively moved to the polls.

A fourth set of analyses extend the experimental project by considering the dynamics of voter mobilization. Contact is found to be more effective as Election Day approaches, and yet thirty to fifty percent of the mobilization effect on turnout in one election is carried into future elections (Gerber, Green, and Shachar 2003). That is, blandishments to vote are less effective when made earlier in an election campaign, suggesting that contact has a limited shelf life, but those individuals who are effectively moved to the polls continue to be more likely to vote in future elections, suggesting the act of voting is transformative. Using data from fourteen experiments targeting low-propensity communities of color conducted previous to the November 2008 election, Michelson, García Bedolla and Green (2009) find a similar habit effect in low-propensity communities of color. Across fourteen separate mobilization experiments conducted during 2008, one third of the mobilization effect generated earlier in the year was transferred to turnout in the general election. Gerber et al.(2003) hypothesize that individuals successfully encouraged to vote may, in the future, feel more self confident about their ability to negotiate the voting process, or may have shifted their self identity to include civic participation rather than abstention.

Mobilization experiments have also explored the effect of social networks. Nickerson (2008) examines two canvassing efforts that spoke with one individual in two-voter households, allowing for measurement of both the effect on contacted voters and their housemates. The study utilized a unique placebo design, wherein individuals assigned to the control group were contacted but received a message encouraging them to recycle. Both experiments found that sixty percent of the propensity to vote was passed along to the other member of the household.

Yet, other experiments using social networks have produced mixed results (see Nickerson's chapter in this volume).

Electoral context is also an important factor; even if all individuals in a treatment group are successfully contacted, not all will be moved to vote. This is a reflection of the ongoing real-world context from which experimental subjects are taken (see Gaines, Kuklinski and Quirk 2007). Arceneaux and Nickerson (2009) argue that mobilization has the strongest effect on voters who are indifferent about turning out, but these indifferent voters are not the same from one election to the next. Only low-propensity voters can be mobilized in high-salience elections, while high-propensity voters are more likely to respond in low-salience elections, and occasional voters are best targeted during mid-level salience elections.

Since the baseline effectiveness of various treatments has been established, voter mobilization experiments provide an excellent real-world setting by which to test social psychological theories. Researchers know how much turnout is elicited using various techniques. By embedding psychological theories into messages encouraging turnout, the strength of the role the psychological constructs play in voter mobilization can be measured. One of the first efforts to link social psychology to voting behavior through field experiments was conducted by a team of researchers at the Ohio State University prior to the 1984 presidential election. Students predicting that they would vote were in fact more likely to do so (Greenwald et al. 1987). Efforts to replicate the finding on a larger scale, however, have failed to uncover reliable treatment effects on representative samples of voters (for example, Smith, Gerber, and Orlich 2003). Gollwitzer's theory of implementation intentions (Gollwitzer 1999), which holds that articulating explicit plans for action increases follow-through, has been found to more than double the effect

of mobilization phone calls by simply asking subjects about when they will vote, where they will be coming from, and how they will get to the polling place (Gerber and Rogers 2009; Nickerson and Rogers 2010).

Psychological theories have also been used to explain apparent paradoxes in the literature. For instance, contacting people more than once, either by phone or in person, does not increase turnout significantly more than a single phone contact. However, an important caveat to this finding is that follow-up calls made to individuals who indicate in an initial contact that they intend to vote has a powerful and large effect on turnout (Michelson, García Bedolla and McConnell 2009). In a series of experiments, Michelson et al. asked youth, Latinos, and Asian Americans that were contacted during an initial round of telephone calls whether or not they intended to vote. Restricting follow-up calls to voters who indicated that they intended to vote resulted in double-digit treatment effects, most of which can be attributed to the second call. To explain this finding, Michelson et al. turn to Sherman's (1980) theory of the self-erasing nature of errors of prediction, which posits that asking people to predict their future behavior increases the likelihood of them engaging in the predicted behavior, and Fishbein and Jaen's (1975) theory of reasoned action, which holds that subjects respond to treatment if a social norm is cued and subjects care what others think.

Monitoring has been found to enhance compliance to social norms in the laboratory (for example, Rind and Benjamin 1994). Consistent with those findings, field experiments have found that the mobilization effect is enhanced by messages that signal to voters that their behavior is being observed. Gerber, Green, and Larimer (2008) sent mailers to targeted individuals that indicated to varying degrees that they were being monitored. Some mailers noted

only that researchers were watching the election, others included the recipient's own voting history or the voting histories of the recipient's neighbors as well, and some also included a promise to send an updated chart after the election. The more intrusive and public the information provided, the larger the effect on turnout. The final treatment arm in the experiment raised turnout by 8.1 percentage points, exceeding the effect of many door-to-door efforts.

Without existing benchmarks to compare the results, the 1998 New Haven experiment simply constituted proof that campaigns can mobilize voters and overcome the collective action problem inherent in political participation. It also provided an invaluable example of how campaigns could be studied experimentally. That template has been expanded on to answer questions of increasing nuance and detail about who can be mobilized, the dynamics of mobilization, and the psychology of mobilization. Despite the wealth of insights gained from the last decade of experiments, the experimental study of mobilization behavior faces a large number of potential problems. The next section discusses the practical problems of carrying out experiments, concerns about the external validity of the findings, and ethical concerns about studying campaign activities.

2. Problems Facing Field Experiments: Implementation

The major attraction of experiments as a methodology is that randomization assures that the treatment and control groups are comparable. By gathering theoretically ideal datasets, researchers can offer transparent and straightforward analysis without the need for control variables or complicated modeling. However, constructing these data is difficult. Not only is the process time-consuming, but problems can arise when working in the field removes control from

the researcher. In particular, treating the correct people and documenting the contact can be difficult.

While the heads of organizations may agree to participate in experiments, faithful execution of the protocol is not always a given. Mistakes can be made by managers when providing lists. Volunteers may make mistakes when knocking on doors, they may speak with everyone encountered on a block in enthusiasm for the campaign, or avoid blocks entirely because they do not think the campaign will be well received. Treating members of the control group is mathematically equivalent to failing to apply the treatment to members of the treatment group and does not necessarily invalidate the experiment. The assignment can still be used as an instrument for actual contact to purge the estimate of the nonrandom determinants of contact, but statistical power will suffer dramatically (Nickerson 2005).

Carefully training managers and canvassers can help to mitigate these problems, as can active involvement by the researcher in providing lists and monitoring the campaign. Randomizing at the precinct level, rather than at the household level, can prevent many errors by managers and volunteers. The serious downside of this strategy is that the power of the experiment decreases. The power of an experiment comes, in large part, from the number of random decisions made. Randomly assigning ten precincts to treatment and control groups, rather than 5,000 households, yields vastly fewer possible outcomes of the process. Statistical power is decreased when subjects in a group share characteristics and tendencies (that is, intracluster correlation is high). Whether this decrease in power offsets problems in implementation depends on the extent of anticipated problems and the degree of subject homogeneity within precincts. Last-minute changes in strategies have caused a number of

experiments to go in the dumpster, as the control group is mobilized just like the treatment group. Nickerson (2005) offers various scalable protocols to conserve statistical efficiency in the face of problems implementing a treatment regime.

Even if an organization makes a good faith effort to adhere to the prescribed protocol, simply applying the treatment to assigned subjects can be objectively difficult. When working door-to-door, canvassers must negotiate unfamiliar streets, and in rural neighborhoods may find themselves in areas without street signs or house numbers. Physical barriers such as locked gates and apartment buildings, the presence of dogs or the lack of sidewalks may prevent canvasses from accessing doors. Even when canvassers have access, targeted voters are often not at home. Young people and low-income individuals are likely to have moved since registering; older individuals are often at work or otherwise away when canvassers are available. All of these factors will cause contact rates to be less than one hundred percent (in fact, contact rates in the high single digits or low teens are not uncommon for a single pass through a neighborhood or call sheet). Low contact rates reduce statistical power and the primary solution is to revisit the neighborhood or phone list repeatedly. This added labor can decrease the number of subjects covered. Thus, researchers should structure their randomizations in such a way that unattempted people can be placed into the control group or omitted from the analysis (Nickerson 2005).

Experiments in some minority areas pose special challenges because of naming conventions. Latino families often use the same first names but with suffixes (for example, Junior, Senior), or with different middle names (for example, Maria A. Garcia and Maria E. Garcia). Hmong all share the same twenty last (clan) names, and have very similar first names as well. Canvassers working in these communities must be particularly attentive to the details of the

names (and perhaps ages or other identifying information) in order to ensure that they are contacting the targeted individual. Those preparing walk lists or call sheets must be attentive to these issues as well; for example, by not deleting the middle or suffix name columns to save space, and by drawing attention to these problems during canvasser training. Furthermore, matching names to voter files after the election can be complicated as multiple matches will be likely. Once again, collecting and retaining as much identifying information as possible will mitigate these problems.

A final logistical problem (and a challenge for internal validity) is defining what constitutes contact from the campaign. Contact is not a problematic definition for impersonal forms of outreach such as mail, leaflets, and email. Incorrect addresses and spam filters may prevent some materials from reaching their intended targets, but most mailed and emailed Get Out The Vote (GOTV) messages can safely be assumed to have been delivered. However, it is difficult to know how much of a script must be completed on the phone or in person to consider a subject treated. This coding decision makes no difference for intent-to-treat analysis that relies solely on assignment to treatment conditions (and is most useful for program evaluation), but it poses a large problem for attempts to measure the effect of a campaign on individuals (the quantity political scientists are typically interested in). If treatment is defined as a respondent listening to the entire script, but there is an effect of listening to half of the script and hanging up, then estimates of the treatment effect will be biased. An alternative is to define treatment more loosely, including any individual with whom any contact is made. This allows for more reasonable adoption of the assumption that noncontact has zero effect, but may also dilute the measured effect of the intended treatment.

A related problem is heterogeneity in the treatment applied by canvassers and callers. Again, variance in the treatment provided is not a problem for indirect tactics, but it is a concern when campaign workers are interacting with subjects. In laboratory settings, variance in treatment is typically solved by limiting oversight and implementation of the experiment to one or two people. This solution is not practical in large voter mobilization experiments where hundreds of thousands of households can be included in the experiment. Conversations that are rushed and impersonal are less effective than those that are measured and conversational (Nickerson 2007; Michelson, García Bedolla and Green 2009; Ha and Karlan 2009). The talent and charisma of individual volunteers will vary in large campaigns and subjects may be given qualitatively different treatments depending on their canvasser or caller.

Researchers can work to minimize variance in treatment by carefully training workers and crafting scripts that anticipate deviations and questions, thereby equipping canvassers to provide consistent answers. However, the researcher should keep in mind that the quantity to be estimated is always an *average* treatment effect. This average conceals variation in how subjects respond and variation in the treatment provided. Researchers can take two steps to capture this variation. Canvassers and callers can be randomly assigned phone numbers or canvassing areas, and researchers can record which canvasser contacts each targeted subject. Combined, these two design principles allow researchers to measure the extent of the variation across canvassers.

3. Problems Facing Field Experiments: External Validity

The chief reason to study campaign effects in the field rather than in the laboratory is to more accurately capture the experience of typical registered voters receiving contact in real-world settings with the associated distractions and outside forces acting on the interaction. That

is, the whole point of field experiments is external validity. However, field experiments themselves can only draw inferences about compliers, campaigns subjecting themselves to experimentation, and the techniques campaigns are willing to execute.

Researchers can attempt to include all registered voters in an experiment and make assignments to treatment and control groups. However, as discussed, the treatment will not be applied to all subjects. Subjects can be usefully divided into those who are successfully treated (compliers and always-takers) and those who are not (noncompliers) (Angrist, Imbens, and Rubin 1996). As an epistemological matter, it is impossible to know the effect of the treatment on noncompliers because they do not accept the assigned treatment by definition. Thus, conditioning on contact provides researchers only with the average treatment effect on those contacted. People who cannot be contacted are likely to be different from people who can be contacted (Arceneaux et al. 2006), so the extent to which the results apply to the uncontacted is an open question. Raising contact rates can address some concerns about external validity, but without one hundred percent compliance it is impossible to know what would happen if all targeted individuals were successfully contacted.

Researchers are also limited by the types of campaigns that agree to cooperate with them. Specifically, campaigns are likely to agree to randomize their contacts only when they have limited resources or if they do not believe the experiment will influence the outcome of the election. Given the high level of uncertainty of most political candidates, as well as the contradictory and expensive advice of campaign consultants, this generally has meant that political parties and candidates have declined to participate in field experiments.^{iv} To date, only one high-profile campaign, that of Rick Perry in the 2006 Texas gubernatorial race, has agreed to

participate in a nonproprietary experimental study (Gerber et al. 2007). The bulk of experiments has been conducted by nonpartisan 501(c)3 civic organizations, many of whom have a strong incentive to cooperate as funders increasingly want such efforts to include experimental evaluation components. If well-funded and highly salient campaigns behave differently and/or voters respond differently to outreach from brand name organizations, then external validity is a real concern for much of the mobilization literature.

A primary tension in the experimental mobilization literature is between theory and authenticity. Working with an actual campaign or organization can expand the scope of an experiment and add verisimilitude, but organizations have competing goals that compromise research design. Because of objections from the organization being studied, theories are rarely tested cleanly. Experiments can be designed to minimize the direct and indirect cost to campaigns, but the tradeoff is nearly unavoidable. For example, groups regularly resist removing a control group from their target pool of potential voters, either because they overestimate their ability to gather enough volunteers and reach all voters in a particular community or because they believe it will hurt their reputation if they do not reach out to all individuals that would expect to be contacted. Organizations also resist trying new techniques proposed by researchers and prefer to use familiar techniques used by the group in past campaigns. Some of these objections can be overcome by offering additional resources in exchange for cooperation, encouraging groups to provide an honest estimate of organizational capacity, and designing experiments to minimize the bureaucratic burden on managers. Still, cutting edge research is difficult to orchestrate with existing campaigns and organizations.

Researchers constructing their own campaign have more freedom, although they are still limited by internal institutional review board (IRB) requirements and federal law, but their efforts may not mimic actual campaign behavior. For example, researchers are likely to be constrained by tax laws preventing research dollars from pursuing partisan aims, thereby limiting much of their research to nonpartisan appeals. Conducting free-standing campaigns also opens researchers to a host of ethical considerations that are largely not present when working with an organization already intervening in the community.

4. Problems Facing Field Experiments: Ethical Concerns

Voter mobilization scholars interacting with real-world politics and political campaigns have the potential to change real-world outcomes. Thus, they face ethical obligations that likely exceed limits that might be imposed by internal IRBs. The first ethical concern is that conducting experiments in actual electoral environments can present a situation where a researcher could swing a close election. Most high-profile elections are decided by large margins, but even here there are well-known exceptions, such as the narrow victories of George W. Bush in Florida in 2000, Christine Gregoire in the Washington gubernatorial election of 2004, and Al Franken in the 2008 Senate race in Minnesota. Local elections are much more frequently decided by small margins, many by only a few dozen votes. Thus, even a nonpartisan voter mobilization campaign could swing an election by increasing voter turnout in one neighborhood but not another. Avoiding experiments that could potentially alter electoral outcomes may well reduce allegations of tampering; however, such a strategy may limit the external validity and usefulness of GOTV research.^v Without research in tightly contested partisan settings, scholars are limited in the

conclusions they can draw about when mobilization works and which types of messages are most persuasive.

On the other hand, working in cooperation with real campaigns does mitigate some ethical concerns. Working with campaigns means that an experiment simply systematizes an activity that would take place in any case. A control group or ineffective experimental treatment to be tested could swing an election, but the decision is ultimately made by the candidate or civic group studied, not by the researcher. Working with organizations engaged in campaigns immunizes researchers to some extent from ethical concerns about election outcomes.

Yet, much as doctors and psychologists face dilemmas on whether to monitor government torture, researchers must consider carefully whether or not they want to be involved in and lend validity to campaigns that pursue illiberal ends, violate privacy, or cause psychological distress. For instance, flyers announcing that elections are held on Wednesday may be an effective campaign tactic, but testing such a tactic violates the democratic norm of broad participation. Similarly, voter files make accessible to scholars massive amounts of personal information. As with all research that involves human subjects, the privacy of individuals must be respected. Scholars should take steps to anonymize data as thoroughly as possible when sharing with other academics and research assistants.

Even when information is used legally, it can cause private citizens to feel that their privacy has been violated and generate adverse consequences, as illustrated by a recent set of experiments conducted by Gerber et al. (2008) and Panagopoulos (2009b). In both cases, researchers indicated to treatment group individuals that their voting history was public knowledge and that they would be broadcasting their election behavior – either via mailings or

newspaper advertisements – to their neighbors. No laws were broken, yet individuals in the treatment groups were horrified to learn that their private voting behavior might be made public, to the extent that in the latter case they contacted their local District Attorneys and the researcher was contacted by law enforcement. Regardless of the legality of such experiments, scholars might think twice about trying to replicate or build upon this sort of work. As data about people become increasingly available for purchase or harvest from the web, researchers should limit what may be considered violations of privacy, even if they are using public data.

5. Future Directions

This chapter has focused on voter turnout in particular because it is the best developed experimental literature with regards to mobilization, yet much remains to be explored. New technologies will need to be tested, such as interactive text messaging, nanotargeting advertisements, and the numerous peer-to-peer activities pioneered by MoveOn.org (see Middleton and Green 2008). More theories from related fields such as psychology (for example, cognitive load), economics (for example, prospect theory), and sociology (for example, social cohesion) can be applied to the voter mobilization setting. More can be learned about the dynamics of information flow in campaigns. The availability of inexpensive mobile computing platforms (for example, Palm Pilots, Blackberries, and iPhones) will afford researchers the luxury of better data and the ability to execute more sophisticated experiments. It is also likely that the effectiveness of tactics will vary over time, and these shifts should be documented.

Moving beyond turnout in general to vote choice is another area where future research is likely to make major inroads. Some nonproprietary research has been done on partisan or persuasive campaigns, but the results from these experiments differ wildly. For example, Gerber

(2004) examines the results of several field experiments conducted in cooperation with actual candidates to estimate the effect of mailings. Preferences are measured by examining ward-level returns for two experiments randomized at the ward level, and with post-election surveys for three experiments randomized at the household level. For the ward-level experiments, mailings sent by the incumbent had a significant effect on vote choice in the primary but not in the general election. By contrast, for the three household-level experiments incumbent mail did not affect vote choice, while challenger mail had statistically significant and politically meaningful effects. Similarly intriguing results are reported by Arceneaux (2007), who found that canvassing by a candidate, or by a candidate's supporters, increased support for that candidate (as measured by a post-election survey) but did not alter voters' beliefs about the candidate. It is not even clear whether partisan or nonpartisan campaigns are better at mobilizing voters. To convincingly answer the question, partisan and nonpartisan messages must be tested head-to-head; such experiments are rare and inconclusive (Michelson 2005; Panagopoulos 2009a). In short, the field is wide open for ambitious scholars to understand what factors influence individual vote choice and whether partisan appeals are more or less effective at stimulating turnout than are nonpartisan appeals.

The reason for the dearth of studies in this area is the difficulty in measuring the dependent variable. Whereas voter turnout is a public record in the U.S., vote choice is private. Thus, researchers must either randomize precincts and measure precinct-level vote choice or they must survey individuals after the election. The precinct-level strategy has two primary downsides. First, treating a sufficient number of precincts to draw valid inference requires very large experiments that are often beyond the budget of experimenters. Second, randomizing at the

precinct level precludes the analysis of subgroups of interest because precinct-level vote totals cannot be disaggregated. Surveying subjects after an election solves the subgroup problem, but introduces problems of its own. Such surveys are expensive and nonresponse rates are often high, leading to problems with external validity and concerns that subject attrition may not be equal across treatment and control groups.

Civic participation is much broader than the act of voting. Citizens (and noncitizens) attend meetings, volunteer for organizations, donate to campaigns, lobby elected officials, and engage in a host of activities. In principle, all of these topics are amenable to experimental study. For example, several experiments have explored charitable giving. Han (2009) randomly changed an appeal to buy a one-dollar bracelet to support Clean Water Action (a national environmental group) by adding two sentences of personal information about the requester, meant to trigger a liking heuristic. Individuals who were asked to donate and who received the appeal with the added personal information were twice as likely to donate. Miller and Krosnick (2004) randomly varied the text of a letter soliciting donations to the National Abortion and Reproductive Rights Action League (NARAL) of Ohio. The control letter included the same sort of language usually found in such fundraising letters, a “policy change threat” letter warned that powerful members of Congress were working hard to make abortions more difficult to obtain, and a “policy change opportunity” letter claimed powerful members of Congress were working hard to make abortions easier to obtain. Recipients of the letters, all Democratic women, were asked to make a donation to NARAL Ohio and to sign and return a postcard addressed to President Clinton. Only the threat letter had a significant effect on financial contributions, while

only the opportunity letter had a significant effect on returned postcards. Future experiments could expand on these results to study campaign donations.

Another emerging area of field experiments explores how citizen lobbying affects roll call votes in state legislatures. Bergan (2007) conducted an experiment in cooperation with two public health-related groups aiming to win passage of smoke-free workplace legislation in the lower house of the New Hampshire legislature. Group members were sent an email asking them to send an email to their legislators; emails intended for legislators selected for the control group were blocked, while emails intended for legislators selected for the treatment group were sent as intended. Controlling for past votes on tobacco-related legislation, the emails had a statistically significant effect on two pivotal votes. This form of political mobilization is increasingly common among grassroots organizations and worthy of further study.

Nearly every civic behavior could be studied using experiments if enterprising researchers were to partner with civic organizations. In exchange for randomly manipulating the appeals to members of the group (or the broader public) and measurement of the outcome of interest (for example, meeting attendance), organizations could learn how to maximize the persuasiveness of their appeals to attract the largest possible set of volunteers, donors, or activists. The work on voter turnout can serve as a useful template for these types of studies.

6. Conclusion

Since the modern launch of the subfield less than a decade ago, hundreds of field experiments have expanded our understanding of when and how voter mobilization campaigns work to move individuals to the polls. Despite real-world hazards such as threatening dogs, contaminated control groups, and uneven canvasser quality, hundreds of efforts have replicated

and extended the initial findings offered by Gerber and Green (2000). Experiments have been conducted in a variety of electoral contexts, with a variety of targeted communities, and exploring a variety of psychological theories. Several chapters in this volume offer additional details about experiments in voter mobilization, including one by Chong on work with minority voters and one by Sinclair on the power of interpersonal communication. Yet, much work remains to be done. We look forward to the next generation of experiments, which in addition to refining existing results will include more new technologies, richer theoretical underpinnings, more work on partisan and persuasive campaigns, and behaviors beyond turnout.

References

- Adams, William C., and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44: 53-83.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-55.
- Arceneaux, Kevin T. 2007. "I'm Asking for Your Support: The Effects of Personally Delivered Campaign Messages on Voting Decisions and Opinion Formation." *Quarterly Journal of Political Science* 2: 43-65.
- Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 1-36.
- Arceneaux, Kevin, and David W. Nickerson. 2009. "Who is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53: 1-16.
- Bergan, Daniel E. 2007. "Does Grassroots Lobbying Work?: A Field Experiment Measuring the Effects of an e-Mail Lobbying Campaign on Legislative Behavior." *American Politics Research* 37: 327-52.
- Dale, Allison, and Aaron Strauss. 2007. "Text Messaging as a Youth Mobilization Tool: An Experiment with a Post-Treatment Survey." Paper presented at the annual meeting of the American Political Science Association, Chicago, IL.

- Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50: 54-65.
- Fishbein, Martin, and Icek Ajzen. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison Wesley.
- Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. London: Oliver & Boyd.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.
- Gerber, Alan S. 2004. "Does Campaign Spending Work? Field Experiments Provide Evidence and Suggest New Theory." *American Behavioral Scientist* 47: 541-74.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653-63.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.
- Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. "Voting May be Habit Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 47: 540-50.
- Gerber, Alan, James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2007. "The Influence of Television and Radio Advertising on Candidate Evaluations: Results from a Large-Scale Randomized Experiment." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Gerber, Alan S. and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You." *Journal of Politics* 71: 178-91.
- Gollwitzer, Peter M. 1999. "Implementation Intentions: Strong effects of Simple Plans." *American Psychologist* 54: 493-503.
- Gosnell, Harold F. 1927. *Getting-Out-the-Vote: An Experiment in the Stimulation of Voting*. Chicago: University of Chicago Press.
- Green, Donald P., and Alan S. Gerber. 2008. *Get Out the Vote: How to Increase Voter Turnout* 2nd Ed. Washington, DC: Brookings Institution Press.
- Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Field Experiments." *Political Analysis* 16: 138-52.

- Greenwald, Anthony G., Catherine G. Carnot, Rebecca Beach, and Barbara Young. 1987. "Increasing Voting Behavior by Asking People if They Expect to Vote." *Journal of Applied Psychology* 72: 315-18.
- Ha, Shang E., and Dean S. Karlan. 2009. "Get-Out-the-Vote Phone Calls: Does Quality Matter?" *American Politics Research* 37: 353-69.
- Han, Hahrie C. 2009. "Does the Content of Political Appeals Matter in Motivating Participation? A Field Experiment on Self-disclosure in Political Appeals." *Political Behavior* 31: 103-16.
- Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99: 283-300.
- Michelson, Melissa R. 2005. "Meeting the Challenge of Latino Voter Mobilization." *Annals of Political and Social Science* 601: 85-101.
- Michelson, Melissa R., Lisa García Bedolla, and Donald P. Green. 2007. "New Experiments in Minority Voter Mobilization: A Report on the California Votes Initiative." San Francisco, CA: The James Irvine Foundation.
- Michelson, Melissa R., Lisa García Bedolla, and Donald P. Green. 2008. "New Experiments in Minority Voter Mobilization: Second in a Series of Reports on the California Votes Initiative." San Francisco, CA: The James Irvine Foundation.
- Michelson, Melissa R., Lisa García Bedolla, and Donald P. Green. 2009. "New Experiments in Minority Voter Mobilization: Final Report on the California Votes Initiative." San Francisco, CA: The James Irvine Foundation.
- Michelson, Melissa R., Lisa García Bedolla, and Margaret A. McConnell. 2009. "Heeding the Call: The Effect of Targeted Two-Round Phonebanks on Voter Turnout." *Journal of Politics* 71: 1549-63.
- Middleton, Joel A., and Donald P. Green. 2008. "Do Community-Based Voter Mobilization Campaigns Work Even in Battleground States? Evaluating the Effectiveness of MoveOn's 2004 Outreach Campaign." *Quarterly Journal of Political Science* 3: 63-82.
- Miller, Joanne A., and Jon A. Krosnick. 2004. "Threat as a Motivator of Political Activism: A Field Experiment." *Political Psychology* 25: 507-23.
- Miller, Roy E., David A. Bositis, and Delise L. Baer. 1981. "Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman." *International Political Science Review* 2: 445-60.

- Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233-52.
- Nickerson, David W. 2006. "Hunting the Elusive Young Voter." *Journal of Political Marketing* 5: 47-69.
- Nickerson, David W. 2007. "Quality is Job One: Volunteer and Professional Phone Calls." *American Journal of Political Science* 51: 269-82.
- Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49-57.
- Nickerson, David W. and Todd Rogers. 2010. "Do You Have a Voting Plan? Implementation Intentions, Voter Turnout, and Organic Plan Making." *Psychological Science* 21: 194-99.
- Panagopoulos, Costas. 2009a. "Partisan and Nonpartisan Message Content and Voter Mobilization: Field Experimental Evidence." *Political Research Quarterly* 62: 70-76.
- Panagopoulos, Costas. 2009b. "Turning Out, Cashing In: Extrinsic Rewards, Extrinsic Rewards, Intrinsic Motivation and Voting." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Rind, Bruce and Daniel Benjamin. 1994. "Effects of Public Image Concerns and Self-Image on Compliance." *Journal of Social Psychology* 134: 19-25.
- Rosenstone, Steven J., and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.
- Sartori, Anne E. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11: 111-38.
- Sherman, Steven J. 1980. "On the Self-Erasing Nature of Errors of Prediction." *Journal of Personality and Social Psychology* 39: 211-21.
- Smith, Jennifer K., Alan S. Gerber, and Anton Orlich. "Self Prophecy Effects and Voter Turnout: An Experimental Replication." *Political Psychology* 24: 593-604.
- Vavreck, Lynn. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2: 287-305.
- Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.

ⁱ For example, California's population has not been majority-Anglo (non-Latino white) for some time, and yet the electorate is over two-thirds Anglo. Thus, elections and ballot measures are decided by an electorate that is not necessarily representative of state opinion.

ⁱⁱ Most of the critique that follows is equally applicable to the few studies using selection models. Selection models acknowledge the problem with strategic targeting by campaigns and attempt to model the process; however, such models rely on strong assumptions that may not be warranted in many instances, so the problem of strategic selection is not fully solved (see Sartori 2003).

ⁱⁱⁱ Ironically, a merging error that did not substantively alter the results cast doubt upon the initial findings with respect to phone calls (Imai 2005).

^{iv} Most partisan experiments conducted to date have been proprietary in nature.

^v Ideally, researchers could work with groups on both sides of the partisan divide to avoid appearances of bias. In practice, partisan organizations are generally suspicious, and researchers are likely to be forced to specialize on one side or the other.

V. Interpersonal Relations

17. Trust and Social Exchange

Rick K. Wilson and Catherine C. Eckelⁱ

Trust and its complement, trustworthiness, are key concepts in political science. Trust is seen as critical for the existence of stable political institutions, and for the formation of social capital and civic engagement (Putnam 1993, 2000; Stolle 1998). It also serves as a social lubricant that reduces the cost of exchange, whether in reaching political compromise (Fenno 1978; Bianco 1994) or in daily market and nonmarket exchange (Lupia and McCubbins 1998; Sztompka 1999; Knight 2001). Researchers in this area face three key challenges. First, the concept of trust has been used in a multiplicity of ways, leaving its meaning unclear. Second, it is used to refer both to trust in government, and trust among individuals (interpersonal trust). Third, it is sometimes seen as a cause and sometimes as an effect of effective political institutions, leaving the causal relationship between trust and institutions unclear.

The definition of trust is muddled by the fact that two distinct research methods have been used to explore it. Early research treats trust as a perception of norms in a society, assessed using survey questions about the trustworthiness or fairness of others. Across forty years the General Social Survey (GSS), World Values Survey (WVS), and American National Election Studies (ANES) have relied on the same questions to evaluate trust. In contrast, recent research has turned to behavioral assessments of trust using incentivized, economics-style laboratory experiments; this work is the focus of this essay. For the most part, behavioral research uses the investment game (Berg, Dickhaut, and McCabe 1995), where one individual decides whether to trust another by making a decision to put his financial wellbeing into the hands of another

person. The relationship between these two concepts of trust – the survey measure of perceived trustworthiness of others, and the decision by a laboratory participant to trust his counterpart – is relatively weak, yet both are used for the same purpose: to assess the levels of trust and reciprocity in a group or society. In Section 1, we present a framework for categorizing concepts of trust, placing our discussion of behavioral trust in a richer context.

Second, there is a difference between trust in government and trust among individuals. Trust in government is addressed in a number of studies, including Miller (1974), Citrin (1974), Hetherington (1998), and Hibbing and Theiss-Morse (2002), who note how political trust varies with assessments of, for example, the presidency and Congress. Levi and Stoker (2000) provide an overview of this work. While trust among citizens is interrelated with political trust, the two are conceptually distinct. Behavioral research has exclusively considered trust between individuals, and has only rarely dealt directly with trust in political or governmental institutions. Because of our focus on behavioral trust, we leave the evaluation of trust in institutions to others.

A third issue is the complex causal relationship between societal levels of trust and the effectiveness and legitimacy of political institutions. While the argument has been made that trust affects and is affected by political institutions, and that it plays an important role in limiting or enhancing the effectiveness of those institutions, the causal relationship has been difficult to disentangle from survey-based data. We argue that behavioral research can play an important role in addressing this important problem.

This essay, then, focuses on the contributions of behavioral research, in particular economics-style experimental studies of trust in a dyadic exchange transaction, to an understanding of trust. We ask, what can laboratory-based studies of dyadic trust contribute to

answering the aforementioned questions? First, we turn to a framework for understanding dyadic trust. We next turn to the canonical trust experiment. Third, we examine the individual level correlates of trust. Fourth, we turn to strategic aspects of trust. Fifth, we detail cross-cultural research on trust. Sixth, we evaluate the link between trust and institutions. Finally, we conclude with what we consider to be the unanswered questions.

1. A Framework for Interpersonal Trust

Interpersonal trust goes by many names: generalized trust, moral trust, particularized trust, encapsulated trust, dyadic trust, and so on. Nannestad (2008) provides a useful framework for considering the broad range of research on trust. He argues that trust can be organized along two dimensions: the first ranging from general to particular, and the second ranging from rational to moral. On the general/ particular dimension, generalized trust is represented by the GSS, WVS, and ANES questions, which ask individuals to assess the degree of fairness and trustworthiness of “most people.” At the other extreme, particularized trust refers to situations where an individual decision to trust has a specific target and content (individual A trusts individual B with respect to X). On the rational/moral dimension, concepts such as Hardin’s (2002) “encapsulated trust” falls close to the rational end of the spectrum. Here, trust is based on a calculation of expected return and depends on the assessed trustworthiness of others. Uslaner (2002), who conceptualizes trust and trustworthiness as moral obligations akin to norms that have been shaped by childhood experiences, falls on the other end of the spectrum.

Survey research is perhaps best suited to assess trust in the general/moral quadrant of Nannestad’s typology, while experimental studies of dyadic trust are in the particular/rational quadrant. This helps clarify why these two approaches to measuring trust are so weakly related.

Glaeser et al. (2000), for example, find little statistical correlation between an individual's answers to the GSS questions and their behavior in the investment game. Both approaches to measuring trust have value, but they address different aspects of the concept. We believe that laboratory experiments are uniquely situated to answer questions in the particular/rational quadrant noted by Nannestad (2008). We do not claim that experiments are a panacea, but rather they provide a useful tool to add to what we already know from the rich literature on trust.

In its simplest form, trust involves a strategic relationship between two actors, where reciprocated trust can improve the wellbeing of both members of the relationship. In political science, this includes negotiations among legislators to trade votes, the decision of a voter to give latitude to a representative, or the willingness of a citizen to comply with the decision of a public official. This concept of dyadic trust is similar to Hardin's (2002) view of "encapsulated" trust, where two actors know something about one another, the context of exchange is clear, and what is being entrusted is well defined. In the case of legislators engaged in vote trading, for example, one legislator faces the problem of giving up a vote with the future promise of reciprocity. Knowing the reputation of the other party is critical to this choice.

Both parties in a dyadic trust relationship have important problems to solve. The truster faces a strategic problem: whether and how much to trust the trustee. His decision depends critically on expectations about the trustee. The problem for the trustee is to decide, if trusted, whether and how much to reciprocate that trust. Her problem is arguably easier, as reciprocity is conditional on the revealed trust of the first mover. Focusing on dyadic trust illustrates an element that is often missing in discussions of interpersonal trust: trust depends on whom (or what) one is dealing with. Individuals do not trust in the abstract, but rather with respect to a

specific target and in a particular context. While the decision about whether and how to reciprocate does not carry the same level of risk for the trustee, their decision also is made in a specific context, with attendant norms of responsibility or obligation. This strategic interaction is difficult to explore in the context of survey-based observational studies, but is well suited to the laboratory.

Since the mid 1990s, more than 150 experimental studies have examined dyadic trust. The standard trust experiment, originally known as the “investment game” (Berg et al. 1995), has proven to be a valuable vehicle for subsequent research. It has given rise to new methodological innovations in experimental protocols, allowed researchers to examine the correlation between behavior and individual characteristics (including neuroscience innovations), provided an environment to study stereotyping and discrimination, and served as a platform for cross-cultural comparison. In addition, it has allowed researchers to examine existing institutional mechanisms and testbed new institutions. In the remainder of this chapter, we examine these aspects of trust experiments and conclude with a set of unanswered questions.

2. The Trust Game

The trust game consists of a sequence of moves between two actors, where both are fully informed about its structure and payoffs. To illustrate, suppose there are two actors, Player A and Player B. Both are endowed with ten dollars by the experimenter. Player A has the right to move first and can choose to keep the ten dollars, or can pass any part of it to the second player. Any amount that is passed is tripled by the experimenter, and then delivered to Player B. (The tripling plays the part of a return on an investment in the game.) Player B now has her original ten dollars and the tripled amount passed to her, and is given the option to send some money back to Player

A. The amount can range from zero dollars to the full tripled value. Player A's move is "trust", in that by sending a positive amount, he entrusts his payoff to Player B; Player B's move is "trustworthiness" or reciprocity. From a game-theoretic perspective, a naïve, payoff-maximizing Player B would retain anything sent to her; Player A, knowing this, will send her nothing. Thus the equilibrium of the game (assuming payoff-maximizing agents) is for Player A to send zero dollars, rightly failing to trust in Player B's trustworthiness.

The canonical implementation of this game has the following characteristics:

- Subjects are recruited from the general student population, and paid a nominal fee for attending, usually five dollars
- Subjects are randomly assigned to the role of Player A or B
- Brief instructions are read aloud, followed by self-paced computerized instructions and a comprehension quiz
- Each player is endowed with an equal amount of money (usually ten dollars)
- Partners are kept anonymous
- A brief questionnaire collects demographic and other information
- Subjects are paid their actual earnings in cash, in private, at the end the experiment

In contrast to the Nash equilibrium, a meta-analysis of results by Johnson and Mislin (2008) shows that, on average, trusters send 50.8 percent of their endowment (based on eighty-four experiments). Trust pays (barely), in that 36.5 percent of what is sent is returned (based on

seventy-five experiments), just over the 33.3% that compensates Player A for what was sent.

Contrary to game-theoretic expectations, trust is widespread and it is reciprocated.

Methodological Issues

As others in this volume note, political scientists are sometimes skeptical of what laboratory experiments can tell us. Experiments seem contrived, the sample is too limited, and the motivations of subjects often seem trivial (see, e.g., Dickson's, Druckman and Kam's, and McDermott's chapters in this volume). A common complaint about laboratory experiments is that, even if subjects are paid, the stakes are insufficient to mimic natural settings. Johansson-Stenman, Mahmud, and Martinsson (2005) ask whether stakes affect behavior in the trust game conducted in Bangladesh, with the highest-stakes game being twenty-five times greater than the lowest-stakes game. The high-stakes setting has the US-dollar price-parity equivalent of \$1,683. They find that, as the size of stakes increases, a somewhat smaller percentage of the money is sent (thirty-eight percent in the high stakes condition compared to forty-six percent in the middle stakes condition). In experiments carried out in Tatarstan and Siberia, subjects are given an endowment equivalent to a full day's wage (Bahry and Wilson 2004); sixty-two percent send half or more of this endowment. While subjects are sensitive to size of stakes, the data clearly show that people trust – and trust pays – even when the stakes are high.

Another common complaint is that students coming into the laboratory are friends and/or anticipate post-game repeated play. High levels of trust may simply be due to subjects investing in reputation. Indeed, Anderhub, Engleman, and Guth (2002) and Engle-Warnick and Slonim (2004) find reputational effects when subjects repeatedly play the trust game. Isolating reciprocity from an investment in reputation is important, and experimenters address this by

taking considerable care to ensure that subjects do not know one another in the same experimental session. To induce complete anonymity, Eckel and Wilson (2006) conduct experiments over the Internet, with subjects matched with others at another site, more than 1000 miles away. The subsequent play of the game is within the range observed in other studies.

A specific complaint is that the game doesn't really measure trust, but rather some other thing such as other-regarding preferences (altruism). Glaeser et al. (2000) ask subjects to report the frequency of small trusting acts – leaving a door unlocked, loaning money to a friend – and find positive correlations between these actions and the trust game. At the same time, they include the standard battery of survey questions and find that they are uncorrelated with trust, but instead are positively associated with trustworthiness. In the same vein, Karlan (2005) finds that the repayment of micro-credit loans is positively correlated with trustworthiness in the trust game, but not with trust. Cox (2004) explicitly tests the role of altruism as a motive for trust and trustworthiness, and finds positive levels of trust, even when controlling for individual level altruism. In sum, the survey measures of trust have weak correlation with behavioral trust, but they seem to predict trustworthiness, indicating that the surveys may be more accurate measures of beliefs about trustworthiness in society. Evidence from variations on the games supports the idea that they constitute valid measures of trust and trustworthiness.

3. Correlates of trust

Observational studies point to heterogeneity in generalized trust within a given population. Uslaner (2002), for example, finds that generalized trust is positively correlated with education, and that African Americans report lower levels of trust. Experimenters also ask what factors are correlated with trust and reciprocity between individuals and corroborate several of

these findings. We first detail results about observable individual characteristics and then turn to a separate discussion of underlying neural mechanisms.

Individual Characteristics

Trust experiments have examined the relationship between personal characteristics, such as gender and ethnicity, and behavior in the games. In a comprehensive survey of gender differences in experiments, Croson and Gneezy (2009) find considerable variation across 20 trust game studies. Many demonstrate no difference in the amount sent, but among the twelve that do, nine show that men trust more than do women. Among the eight studies demonstrating a difference in trustworthiness, six that show women reciprocate more. They argue that the cross-study variation is due to women's greater response to subtle differences in the experimental protocols.

Trust is also rooted in other aspects of socioeconomic status. While experiments with student subjects rarely find an effect of income on behavior (although see Gächter, Herrmann, and Thoni 2004), several recent studies use representative samples and find positive relationships between income and trust behavior (Bellemare and Kröger 2007; Naef et al. 2009). Age is also related to trust and reciprocity. Bellemare and Kröger (2007) find that young and the elderly have lower levels of trust, but higher levels of reciprocity than do middle-aged individuals, a result they attribute to a mismatch between expectations about trust and realized trust. Sutter and Kocher (2007) obtain a similar finding using six age cohorts ranging from eight year old children to sixty-eight year old subjects. They find two clear effects. Trusting behavior is nonlinear with age, with the youngest and oldest cohorts trusting the least, and the twenty-two and thirty-two year old cohorts contributing the most. However, reciprocity is almost linearly related to age,

with the oldest cohort returning the most. These age cohort effects are similar to those reported by Uslaner (2002) using survey data.

Several studies examine religion and trust. Anderson, Mellor, and Milyo (2010) report little effect on trust or trustworthiness of religion, regardless of denomination. This is contrary to the findings by Danielson and Holm (2007) who find that churchgoers in Tanzania reciprocate more than does their student sample. Johansson-Stenman, Mahmud, and Martinsson (2009) match Muslims and Hindus, both within and across religion, and find no difference in any of their matching conditions; they cautiously conclude that religious affiliation does not matter. Together these findings show that there are small effects for standard socioeconomic status variables on behavioral trust.

Some have conjectured that trust is a risky decision, and that observed heterogeneity in trust may be due in part to variations in risk tolerance. In our own work we directly test this conjecture by supplementing the canonical experiment with several different measures of risk tolerance, ranging from survey measures to behavioral gambles with stakes that mirror the trust game (Eckel and Wilson 2004). None is correlated with the decision to trust (or to reciprocate). By contrast, Bohnet and Zeckhauser (2004) focus on the risk of betrayal. They use a simplified trust game with a limited set of choices, and implement a mechanism that elicits subjects' willingness to participate in the trust game, depending on the probability that their partner is trustworthy. They find considerable evidence for betrayal aversion, with trusters sensitive to the potential actions of the population of trustees. Indeed, trusters are less willing to accept a specified risk of betrayal by trustees than to risk a roll of the dice with the same probability of attaining a high payoff. In a later paper, Bohnet et al. (2008) extend this study to six countries

and find variation in betrayal aversion across societies, with greater betrayal aversion associated with lower levels of trust. Whether trust is a risky decision seems to depend importantly on how risk is measured. Clearly this is an area where more work is needed.

In addition to the studies above, laboratory experiments and survey-based studies reach similar conclusions with respect to several factors. Trust is positively associated with level of education and income, a point noted by Brehm and Rahn (1997). Generational differences also emerge, a point that scholars like Inglehart (1997) and Putnam (2000) offer as a cultural explanation for trust.

Contributions from Biology and Neuroscience

A promising arena for understanding individual correlates of trust is linked with biological and neurological mechanisms. In principle, observational studies are equally capable of focusing on these mechanisms. However, most scholars focusing on such issues have a laboratory experimental bent.

Several research teams focus on the neurological basis of trust (for an overview, see Fehr, Kosfeld, and Fischbacher 2005). McCabe et al. (2001) find differences in brain activation in the trust game when subjects play against a human partner as compared to a computer. They speculate what the neural underpinnings might be for trusting behavior. Rilling et al. (2004) examine neural reward systems for a setting similar to the trust game, in which there is the possibility of mutual advantage. Delgado, Frank, and Phelps (2005) focus on both reward and learning systems that follow from iterated play with multiple partners in the trust game. King-Casas et al. (2005) also use an iterated trust game and find not only reward and learning processes, but anticipatory signals in the brain that accurately predict when trust will be

reciprocated. The neural system they isolate is clearly related to processing social behavior and not simply due to internal rewards. Tomlin et al. (2006) report similar results when subjects are simultaneously scanned in an fMRI while playing the trust game.

Several research groups show that the hormone oxytocin (OT) is an important basis for cementing trust. It is proposed that OT is stimulated by positive interactions with a specific partner. Zak, Kurzban, and Matzner (2005) focus on a design to test for changes in OT levels for subjects playing the trust game with another human or playing with a random device. They find elevated levels of OT for trustees assigned to the human condition. Behaviorally, they also observe an increase in reciprocation for those in the human condition. There is no difference in OT for trusters, indicating that the effect is absent for the trust decision. By contrast, Kosfeld et al. (2005) use a nasal spray to administer either OT or a placebo. They find that OT enhances trust, but it is unrelated to trustworthiness.

There is also evidence for a genetic basis of trust. Cesarini et al. (2008) report on trust experiments conducted with monozygotic (MZ) and dizygotic (DZ) twins in the United States and Sweden. While the distribution of trust and trustworthiness is heterogeneous, they find that MZ twins have higher correlations in behavior than do their DZ counterparts. The estimated shared genetic effect ranges from ten to twenty percent in their samples. As the authors admit, it is not all about genes. A significant component of the variation is explained by the twins' environments.

The jury is still out concerning the biological and neural mechanisms that drive trust and trustworthiness. Trust and reciprocation involve complex social behaviors and the capacity to test the mechanisms that cause these behaviors remains elusive.

4. Trust and Stereotypes

Individuals vary systematically in their propensities to trust and to reciprocate trust, but another source of behavioral heterogeneity results from the differences in how individuals are treated by others. Those who have studied campaigns (Goldstein and Ridout 2004; Lau and Rovner 2009) or ethnicity and social identity (Green and Seher 2003; McClain et al. 2009) understand how important it is to control for specific pairings of individuals or groups. Voters respond differently when they have information about a candidate – such as gender, race, or age – than when they have abstract information about a candidate. This is partly because beliefs about others are based on stereotypes. Stereotyping is the result of a natural human tendency to categorize. Two possibilities arise. First, the stereotype may accurately reflect average group tendencies, and so provide a convenient cognitive shortcut for making inferences about behavior. On the other hand, stereotypes can be wrong, reflecting outdated or incorrect information, and can subsequently bias decisions in a way that reduces accuracy. If trust leads to accumulating social capital, then decisions based on stereotypes will advantage some groups and disadvantage others, and negative stereotypes may become self-fulfilling prophecies.

While most of the experimental studies have gone through great efforts to ensure that subjects know nothing about one another, we provide visual information to subjects about their partners. In one design, we randomly assign dyads and allow counterparts to view one another's photograph. This enables us to focus on the strategic implications of the joint attributes of players. To eliminate reputation effects, we use subjects at two or more laboratories at different locations. Photographs are taken of each subject and then displayed to their counterparts. For example, to study the effect of attractiveness on trust and reciprocity, we look at pairings in

which the truster is measured as more (or less) attractive than the trustee. We show that expectations are higher for more attractive trusters and trustees: attractive trusters are expected to send more, and attractive trustees are expected to return more. The attractive truster inevitably fails to live up to high expectations; as a consequence, the truster is penalized and less is reciprocated (Wilson and Eckel 2006).

In another study (Eckel and Wilson 2008), we show that skin shade affects expectations about behavior. Darker skinned trusters are expected to send less, but send more than expected, and they are rewarded for their unexpectedly high trust. The insight we gain is not just from the expectations, but from the response to exceeded or dashed expectations.

Our findings concerning stereotypes are not unusual. For example, trusters prefer to be paired with women, thinking women will be more trustworthy. On average they are, but not to the extent expected (Croson and Gneezy 2009). Trusters send more to lighter skinned partners, trusting them at higher rates, and beliefs about darker skinned partners are weakly supported (see also Fershtman and Gneezy 2001; Haile, Sadrieh, and Verbon 2006; Simpson, McGrimmon, and Irwin 2007; Eckel and Petrie 2009; Naef et al. 2009). These findings are often masked in survey-based studies, as subjects display socially acceptable preferences when it costs them nothing to do so.

In our current work (Eckel and Wilson 2010), we introduce another change to the game by allowing subjects to select their partners. Trusters view the photographs of potential partners after the trust experiment is explained to them, but before any decisions are made. They then rank potential counterparts according to their desirability as a partner, from most to least desirable. To ensure that the ordering task is taken seriously, one truster is randomly drawn and

given her first choice, then a second truster is randomly drawn and given his first choice from those remaining, and so on (following Castillo and Petrie 2009). Not surprisingly, when trusters choose their partners, the overall level of trust is higher. At the same time, when subjects know they have been chosen, they reciprocate at higher rates. Giving subjects some control over the choice of counterpart has a strong positive impact on trust and trustworthiness.

Experiments that focus on the joint characteristics of subjects and that allow for choice among partners are moving toward answering questions about the importance of expectations in strategic behavior, and how expectations are shaped by characteristics of the pairing. Experiments are well suited to answer these questions because of the ability to control information about the pairings of the subjects.

5. Cross-cultural Trust

An ongoing complaint about experiments is that they lack external validity. The concern is that the behavior of American university students is not related to behavior in the general population, within or across different cultures. Several recent experimental studies tackle the question of external validity by looking at population samples, and considerable work has taken place cross-culturally in recent years.

The impetus to behaviorally measure trust across cultures is partly driven by findings by Knack and Keefer (1997) and Zak and Knack (2001) who find that the level of generalized trust in a country is correlated with economic growth. These findings, derived from surveys and aggregate level measures, mirror those by Almond and Verba (1963), who provide evidence that trust is correlated with democratic stability. Researchers using trust experiments have entered this arena as well.

Several studies focus on cross-cultural comparisons of trust using volunteer student subjects. These studies replicate the high levels of trust found among US students, while finding some variability across cultures (see, for example, Yamagishi, Cook, and Watabi 1998; Buchan, Croson, and Dawes 2002; Ashraf, Bohnet, and Piankov 2006).

Bahry and Wilson (2004) are the first to extend these studies to representative population samples in two republics in Russia. They draw from a large sample of respondents who were administered lengthy face-to-face interviews. A subset of subjects was randomly drawn to participate in laboratory-like experiments in the field. While something was gained in terms of confidence in external validity, a price was paid in terms of a loss of control. Sessions were run in remote villages, usually in classrooms or libraries, and the quality and size of the facilities varied (as did the temperature).

Their findings reveal high levels of trust and reciprocity in these republics, despite the fact that the political institutions are regarded with suspicion. On average 51 percent of the truster's endowment was sent and trustees returned 38.3 percent of what was received, a result very close to average behavior among US students. These findings indicate that trust is widespread in an environment where it is unexpected (for example, see Mishler and Rose 2005). More importantly, Bahry and Wilson (2004) point to strong generational differences in norms that lead to distinct patterns of trust and trustworthiness – a finding that would have been unexplored without a population sample.

Others have also generated new insights when conducting trust experiments outside of university laboratories. Barr (2003) finds considerable trust and little variation across ethnic groups in Zimbabwean villages. Carpenter, Daniere, and Takahashi (2004) use a volunteer

sample of adults in Thailand and Vietnam, and find that trust is correlated with formation of social capital measured as owning a home, participating in a social group, and conversing with neighbors. Karlan (2005) obtains a similar finding in Peru. He notes that trust is related to social capital, and shows that trustworthiness predicts the repayment of microcredit loans. Cronk (2007) observes low levels of trust among the Maasai in a neutral (unframed) experimental condition, and even less trust when the decision is framed to imply a long-term obligation. Numerous other studies have focused on trust in Bangladesh (Johansson-Stenman, Mahmud, and Martinsson 2006), Paraguay (Schechter 2007), Kenya (Greig and Bohnet 2005), the US and Germany (Naef et al. 2009), and across neighborhoods in Zurich (Falk and Zehnder 2007). These studies are important in that they aim at linking the trust game to ethnic conflict, repaying loans, and the risk and patience of individuals.

Studies using culturally different groups have moved beyond student samples and begin to assure critics that concerns with external validity are misplaced. As measured by the trust game, trust and trustworthiness permeates most cultures. These studies are beginning to give us insight into cultural variation. How these studies are linked to key questions of support for democratic institutions or increasing political participation have not routinely been addressed.

6. Trust and Institutions

Political scientists have long been concerned with the relationship between interpersonal trust and political institutions. This tradition extends back to Almond and Verba (1963) who claim that there is a strong correlation between citizen trust and the existence of democratic institutions. Subsequent work has examined the nature and directional causality of this relationship. Rothstein (2000) argues that there are two approaches to understanding how trust

among citizens is produced. The first, largely advocated by Putnam (1993), takes a bottom-up approach. Trust emerges when citizens participate in many different environments and, in doing so, experience trust outside their own narrow groups. This, in turn, provides for democratic stability in that citizens develop tolerance for one another, and that ultimately extends to confidence in governmental institutions. The second approach holds that institutions mitigate the risk inherent in a trust relationship, thereby encouraging individuals to trust one another. As Rothstein (2000) puts it, “In a civilized society, institutions of law and order have one particularly important task: to detect and punish people who are ‘traitors’, that is, those who break contracts, steal, murder and do other such noncooperative acts and therefore should not be trusted” (490-1). In this view, institutions serve to monitor relationships, screen out the untrustworthy, and punish noncooperative behavior. For institutions to effectively accomplish this objective, they must be perceived as legitimate.

If citizen trust precedes (or is independent of) institutions, then individual trust ought to be insensitive to the institutions within which they interact. If the causal relationship is such that institutions are crucial for fostering citizen trust, then the legitimacy of the institutions within which individuals make trust decisions should be directly related to the degree of trust and trustworthiness. Laboratory experiments are especially well suited to examine these causal relations.

One type of institutional mechanism introduces punishment into the exchange. We consider two punishment mechanisms. The first allows for punishment by one party of the other in the trust game – *second-party punishment*. Fehr and Rockenbach (2003) allow the first movers in the game to specify an amount that should be returned when making their trust decision. In

one treatment, there is no possibility of punishment; in the other treatment, the first mover has the ability to punish the second mover. The first mover specifies a contract that states the amount he wishes to be returned, and whether he will punish noncompliance. They find that the highest level of trustworthiness is observed when sanctions are possible, but are not implemented by the first mover, and the lowest level of trustworthiness is observed when sanctions are implemented. A similar result is seen in Houser et al. (2008), where the threat of punishment backfires by reducing reciprocity when the first mover asks for too much.

A second punishment mechanism introduces an additional player whose role is to punish the behavior of the players in the trust dyad – *third-party punishment*. Bohnet, Frey, and Huck (2001) focus on whether an ex post mechanism that is designed to reinforce trustworthiness is effective in doing so. In this study, the punishment is implemented automatically, using a computerized robot. The experiment uses a simplified trust game with binary choices, and adds treatments in which failure to reciprocate can trigger a fine, which is imperfectly implemented with a known probability. All actors know the associated thresholds that trigger enforcement and the costs of any associated penalties. Thus the institution is transparent, and is invoked automatically but imperfectly. They find that trust thrives when the institution is weak and the probability of enforcement is low. However, higher levels of punishment crowd out trust. When punishment is punitively large, trust is again observed, but the punishment is so large that it is arguably no longer trustworthiness that motivates the trustee. Trust only thrives when third-party enforcement is absent or low. Using different experimental designs, Kollock (1994) and Van Swol (2003) reach similar conclusions.

Rather than using a robot, Charness and Cobo-Reyes (2008) introduce an explicitly selected third party who is empowered to punish trustees and/or reward trusters. They find more is sent and more is returned when there is the possibility of punishment. Interestingly, even though the third party gains nothing from the exchange between the truster and trustee, that third party is willing to bear the cost of punishing. In contrast, Banuri et al. (2009) adapt the trust game to study bribery, and show that third-party sanctions reduce the incidence of and rewards to trust (as bribery), essentially by reducing the trustworthiness of the bribed official. They argue that high levels of trust and trustworthiness are necessary for bribery to be effective, since parties to the transaction have no legal recourse if the transaction is not completed.

A second type of institutional mechanism uses group decision making for the trust and reciprocity decisions themselves. In Kugler et al. (2007), subjects make trust and reciprocity decisions under group discussion and consensus to decide how much to trust or how much to reciprocate. Compared with control groups using the standard dyadic trust game, they find that trust is reduced when groups decide, but reciprocation is not affected. Song (2008) divides subjects into three-person groups, has individuals make a decision for the group, and then randomly selects one individual's choice to implement the decision for the group. She also finds that levels of trust are reduced and that reciprocity is reduced. Both studies point to trust or reciprocity declining when groups make a decision, a result that echoes similar group polarization results in other game settings (Cason and Mui 1997). Why collective choice mechanisms depress trust is left unexplored.

Finally there has been interest in whether mechanisms that facilitate information exchange, especially cheap talk, enhance trust. Does a trustee's unenforceable promise that trust

will be reciprocated have any effect on trust? In principle it should not. Charness and Dufwenberg (2006) allow trustees an opportunity to send a nonbinding message to the truster. The effect of communication leads to increased amounts of trust, which is then reciprocated: the promise appears to act as a formal commitment by the trustee. As Charness and Dufwenberg explain, trustees who make promises appear “guilt averse,” and so carry out their action. Ben-Ner, Putterman, and Ren (2007) allow both trusters and trustees to send messages. In a control condition, no one was allowed to communicate, in one treatment subjects were restricted to numerical proposals about actions, and in the second treatment, subjects could send text plus numerical proposals. Communication enhanced trust and reciprocity, with the richer form of communication yielding the highest returns for subjects. Sommerfeld, Krambeck, and Milinski (2008) focus on gossip, which is regarded as the intersection of communication and reputation formation. In their experiment, play is repeated within a group and gossip is allowed. The effect is to increase both trust and reciprocation, largely through reputational enhancement (see also Keser 2003 and Schotter and Sopher 2006). Communication matters for enhancing trust and trustworthiness, a finding that is widespread in many bargaining games.

An implication of these findings is that trust is malleable. Perversely, it appears that institutional mechanisms involving monitoring and sanctions can crowd out trust and trustworthiness. While the wisdom of groups might be expected to enhance trust, it does not. Only communication reliably enhances trust and reciprocity, and the more communication, the better. None of these studies pursues why sanctions are sometimes ineffective. One yet unexamined possibility is that the legitimacy of the institutions has not been taken into account. For example, selecting a third-party punisher, as in Charness and Cobo-Reyes (2008), may

endow the sanctions with greater legitimacy, thereby making them more effective. Further work is warranted to determine the basis for effective institutions.

8. The Unanswered Questions

Despite extensive experimental research on trust, there are a number of significant questions left unanswered. In this section we suggest areas for future research.

First, there is still no clear answer to the causal relationship between interpersonal trust and effective political institutions. Because experimental methods have an advantage in testing for causality, we believe it would be fruitful to tackle this question in the lab. As noted above, some work has examined the relationship between trust and specific institutions, but a more general understanding of the characteristics of institutional mechanisms that promote trust, or when trust will breed successful institutions, is paramount.

Second, we do not know enough about why monitoring and sanctioning institutions crowd out trust. Institutions with strong rewards and punishment may indeed act as substitutes for norms of trust and reciprocity. If trust is a fragile substitute for institutional monitoring and sanctioning, then knowing which institutions support, and which undermine, trust is important. It appears that punishment itself is seldom productive, but only the unused possibility of punishment enhances responsible behavior.

Third, little is known about the relationship between trust and social networks. We do not understand whether dyadic trust relationships build communities. Do trusting individuals initiate widespread networks of reciprocated trust? Or do trusting individuals turn inward, limiting the number of partners, thereby segregating communities of trusters? Our own evidence suggests that subjects use skin shade as a basis for discriminating trust. If this persists over time it is easy

to see how segregation can result. In order to understand this dynamic, it is important to study trust as a repeated-play game with large numbers of individuals. A combination of Internet experiments and a longer time period could provide a vehicle for such studies.

Fourth, we still do not know the extent to which the trust game measures political behaviors that are important in natural settings. While we see that the trust game is correlated with some individual characteristics, and predicts small trusting acts such as lending money to friends as well as larger actions such as repayment of micro loans, it is unclear how trust and reciprocity are related to issues of concern to political scientists. For example, how well does the trust game predict leadership behavior? How well does it predict political efficacy? Is it correlated with corrupt behavior of elected or appointed officials?

Fifth, how much insight will biological and neural studies provide into the complex social relationship underlying trusting and reciprocal behaviors? Experimental studies show that trust is sensitive to the context in which the decision is made. Can these biological studies provide explanatory power beyond what we can learn by observing behavior? Their likely contribution will be through a better understanding of the neural mechanisms behind the general perceptual and behavioral biases common to humans.

References

- Almond, Gabriel A., and Sidney Verba. 1963. *The Civic Culture*. Princeton, NJ: Princeton University Press.
- Anderhub, Vital, Dirk Engelmann, and Werner Guth. 2002. "An Experimental Study of the Repeated Trust Game with Incomplete Information." *Journal of Economic Behavior & Organization* 48: 197-216.
- Anderson, Lisa R., Jennifer Mellor, and Jeffrey Milyo. 2010. "Did the Devil Make Them Do It? The Effects of Religion in Public Goods and Trust Games." *Kyklos* 63: 163-75.

- Ashraf, Nava, Iris Bohnet, and Nikita Piankov. 2006. "Decomposing Trust and Trustworthiness." *Experimental Economics* 9: 193-208.
- Bahry, Donna, and Rick K. Wilson. 2004. "Trust in Transitional Societies: Experimental Results from Russia." Presented at the annual meeting of the American Political Science Association, Chicago, IL.
- Banuri, Sheheryar, Rachel Croson, Catherine C. Eckel, and Reuben Kline. 2009. "Towards and Improved Methodology in Analyzing Corruption." CBEES Working Paper, University of Texas, Dallas.
- Barr, Abigail 2003. "Trust and Expected Trustworthiness: Experimental Evidence from Zimbabwean Villages." *Economic Journal* 113: 614-30.
- Bellemare, Charles, and Sabine Kröger. 2007. "On Representative Social Capital." *European Economic Review* 51: 183-202.
- Ben-Ner, Avner, Louis G. Putterman, and Ting Ren. 2007. "Lavish Returns on Cheap Talk: Non-Binding Communication in a Trust Experiment." SSRN Working paper. Retrieved from <http://ssrn.com/abstract=1013582>.
- Berg, Joyce E., John W. Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10: 122-42.
- Bianco, William T. 1994. *Trust: Representatives and Constituents*. Ann Arbor: University of Michigan Press.
- Bohnet, Iris, Bruno S. Frey, and Steffen Huck. 2001. "More Order with Less Law: On Contract Enforcement, Trust, and Crowding." *American Political Science Review* 95: 131-44.
- Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review* 98: 294-310.
- Bohnet, Iris, and Richard Zeckhauser. 2004. "Trust, Risk and Betrayal." *Journal of Economic Behavior & Organization* 55: 467-84.
- Brehm, John, and Wendy Rahn. 1997. "Individual Level Evidence for the Causes and Consequences of Social Capital." *American Journal of Political Science* 41: 999-1023.
- Buchan, Nancy R., Rachel T. A. Croson, and Robyn M. Dawes. 2002. "Swift Neighbors and Persistent Strangers: A Cross-Cultural Investigation of Trust and Reciprocity in Social Exchange." *American Journal of Sociology* 108: 168-206.

- Carpenter, Jeffrey P., Amrita G. Daniere, and Lois M. Takahashi. 2004. "Cooperation, Trust, and Social Capital in Southeast Asian Urban Slums." *Journal of Economic Behavior & Organization* 55: 533-51.
- Cason, Timothy, and Vai-Lam Mui. 1997. "A Laboratory Study of Group Polarization in the Team Dictator Game." *Economic Journal* 107: 1465-83.
- Castillo, Marco, and Ragan Petrie. 2009. "Discrimination in the Lab: Does Information Trump Appearance?" *Games and Economic Behavior*, Forthcoming.
- Cesarini, David, Christopher T. Dawes, James H. Fowler, Magnus. Johannesson, Paul Lichtenstein, and Bjorn Wallace. 2008. "Heritability of Cooperative Behavior in the Trust Game." *Proceedings of the National Academy of Sciences of the United States of America* 105: 3721-26.
- Charness, Gary R., and N. Jimenez Cobo-Reyes. 2008. "An Investment Game with Third Party Intervention." *Journal of Economic Behavior and Organization* 68: 18-28.
- Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica* 74: 1579-601.
- Citrin, Jack. 1974. "Comment: The Political Relevance of Trust in Government." *American Political Science Review* 68: 973-88.
- Cox, James C. 2004. "How to Identify Trust and Reciprocity." *Games and Economic Behavior* 46: 260-81.
- Cronk, Lee. 2007. "The Influence of Cultural Framing on Play in the Trust Game: A Maasai Example." *Evolution and Human Behavior* 28: 352-58.
- Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, Forthcoming.
- Danielson, Anders J., and Hakan J. Holm. 2007. "Do You Trust Your Brethren? Eliciting Trust Attitudes and Trust Behavior in a Tanzanian Congregation." *Journal of Economic Behavior & Organization* 62: 255-71.
- Delgado, Mauricio R., Robert H. Frank, and Elizabeth A. Phelps. 2005. "Perceptions of Moral Character Modulate the Neural Systems of Reward During the Trust Game." *Nature Neuroscience* 8: 1611-18.
- Eckel, Catherine C., and Ragan Petrie. 2009. "Face Value." *American Economic Review*, Forthcoming.
- Eckel, Catherine C., and Rick K. Wilson. 2004. "Is Trust a Risky Decision?" *Journal of Economic Behavior & Organization* 55: 447-65.

- Eckel, Catherine C., and Rick K. Wilson. 2006. "Internet Cautions: Experimental Games with Internet Partners." *Experimental Economics* 9: 53-66.
- Eckel, Catherine C., and Rick K. Wilson. 2008. "Initiating Trust: The Conditional Effects of Sex and Skin Shade among Strangers." Working Paper, Rice University.
- Eckel, Catherine C., and Rick K. Wilson. 2010. "Shopping for Trust." Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Engle-Warnick, Jim, and Robert L. Slonim. 2004. "The Evolution of Strategies in a Repeated Trust Game." *Journal of Economic Behavior & Organization* 55: 553-73.
- Falk, Armin, and Christian Zehnder. 2007. "Discrimination and in-Group Favoritism in a Citywide Trust Experiment." Zurich IEER Working Paper No. 318. Retrieved from <http://ssrn.com/abstract=980875>.
- Fehr, Ernst, Michael Kosfeld, and Urs Fischbacher. 2005. "Neuroeconomic Foundations of Trust and Social Preferences." IEW Working Paper No. 221. Retrieved from <http://ssrn.com/abstract=747884>.
- Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism." *Nature* 422: 137-40.
- Fenno, Richard F., Jr. 1978. *Home Style: House Members in Their Districts*. Boston, MA: Little, Brown and Company.
- Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a Segmented Society: An Experimental Approach." *Quarterly Journal of Economics* 116: 351-77.
- Gächter, Simon, Benedikt Herrmann, and Christian Thoni. 2004. "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence." *Journal of Economic Behavior & Organization* 55: 505-31.
- Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *The Quarterly Journal of Economics* 115: 811-46.
- Goldstein, Kenneth, and Travis N. Ridout. 2004. "Measuring the Effects of Televised Political Advertising in the United States." *Annual Review of Political Science* 7: 205-26.
- Green, Donald P., and Rachel L. Seher. 2003. "What Role Does Prejudice Play in Ethnic Conflict?" *Annual Review of Political Science* 6: 509-31.
- Greig, Fiona, and Iris Bohnet. 2005. "Is There Reciprocity in a Reciprocal-Exchange Economy? Evidence from a Slum in Nairobi, Kenya." Kennedy School of Government Working Paper No. RWP05-044. Retrieved from <http://ssrn.com/abstract=807364>.

- Haile, Daniel, Abdolkarim Sadrieh, and Harrie A. Verbon. 2006. "Cross-Racial Envy and Underinvestment in South Africa." Working Paper, Tilburg University.
- Hardin, Russell. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- Hetherington, Marc J. 1998. "The Political Relevance of Political Trust." *American Political Science Review* 92: 791-808.
- Hibbing, John R., and Elizabeth Theiss-Morse. 2002. *Stealth Democracy: American's Beliefs About How Government Should Work*. New York: Cambridge University Press.
- Houser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith. 2008. "When Punishment Fails: Research on Sanctions, Intention and Non-Cooperation." *Games and Economic Behavior* 62: 509-32.
- Inglehart, Ronald. 1997. *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton, NJ: Princeton University Press.
- Johansson-Stenman, Olof, Minhaj Mahmud, and Peter Martinsson. 2005. "Does Stake Size Matter in Trust Games?" *Economics Letters* 88: 365-69.
- Johansson-Stenman, Olof, Minhaj Mahmud, and Peter Martinsson. 2006. "Trust, Trust Games and Stated Trust: Evidence from Rural Bangladesh." Working Paper, Goteborg University.
- Johansson-Stenman, Olof, Minhaj Mahmud, and Peter Martinsson. 2009. "Trust and Religions: Experimental Evidence from Bangladesh." *Economica* 76: 462-85.
- Johnson, Noel D., and Alexandra Mislin. 2008. "Cultures of Kindness: A Meta-Analysis of Trust Game Experiments." SSRN Working Paper. Retrieved from <http://ssrn.com/abstract=1315325>.
- Karlan, Dean. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review* 95: 1688-99.
- Keser, Claudia. 2003. "Experimental Games for the Design of Reputation Management Systems." *IBM Systems Journal* 42: 498-506.
- King-Casas, Brooks, Damon Tomlin, Cedric Anen, Colin F. Camerer, Steven R. Quartz, and P. Read Montague. 2005. "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange." *Science* 308: 78-83.
- Knack, Stephen, and Philip Keefer. 1997. "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation." *The Quarterly Journal of Economics* 112: 1251-88.

- Knight, Jack. 2001. "Social Norms and the Rule of Law: Fostering Trust in a Socially Diverse Society." In *Trust in Society*, ed. K. S. Cook. New York, NY: Russell Sage Foundation.
- Kollock, Peter. 1994. "The Emergence of Exchange Structures - an Experimental-Study of Uncertainty, Commitment, and Trust." *American Journal of Sociology* 100: 313-45.
- Kosfeld, Michael, Markus Heinrichs, Paul Zak, Urs Fischbacher, and Ernst Fehr. 2005. "Oxytocin Increases Trust in Humans." *Nature* 435: 673-76.
- Kugler, Tamar, Gary Bornstein, Martin G. Kocher, and Matthias Sutter. 2007. "Trust between Individuals and Groups: Groups Are Less Trusting Than Individuals but Just as Trustworthy." *Journal of Economic Psychology* 28: 646-57.
- Lau, Richard R., and Ivy Brown Rovner. 2009. "Negative Campaigning." *Annual Review of Political Science* 12: 285-306.
- Levi, Margaret, and Laura Stoker. 2000. "Political Trust and Trustworthiness." *Annual Review of Political Science* 3: 475-507.
- Lupia, Arthur, and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* New York: Cambridge University Press.
- McCabe, Kevin, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. 2001. "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange." *Proceedings of the National Academy of Sciences* 98: 11832.
- McClain, Paula D, Jessica D. Carew Johnson, Eugene Walton, and Candis S. Watts. 2009. "Group Membership, Group Identity, and Group Consciousness: Measures of Racial Identity in American Politics?" *Annual Review of Political Science* 12: 471-85.
- Miller, Arthur H. 1974. "Political Issues and Trust in Government: 1964-1970." *American Political Science Review* 68: 951-72.
- Mishler, William, and Richard Rose. 2005. "What Are the Political Consequences of Trust? A Test of Cultural and Institutional Theories in Russia." *Comparative Political Studies* 38: 1050-78.
- Naef, Michael, Ernst Fehr, Urs Fischbacher, Jürgen Schupp, and Gert G. Wagner. 2009. "Decomposing Trust: Explaining National and Ethnic Trust Differences." Working Paper, University of Zurich.
- Nannestad, Peter. 2008. "What Have We Learned About Generalized Trust, If Anything?" *Annual Review of Political Science* 11: 413-36.
- Putnam, Robert D. 1993. *Making Democracy Work*. Princeton, NJ: Princeton University Press.

- Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.
- Rilling, James K., Alan G. Sanfey, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2004. "Opposing Bold Responses to Reciprocated and Unreciprocated Altruism in Putative Reward Pathways." *Neuroreport* 15: 2539-43.
- Rothstein, Bo. 2000. "Trust, Social Dilemmas and Collective Memories." *Journal of Theoretical Politics* 12: 477-501.
- Schechter, Laura. 2007. "Traditional Trust Measurement and the Risk Confound: An Experiment in Rural Paraguay." *Journal of Economic Behavior & Organization* 62: 272-92.
- Schotter, Andrew, and Barry Sopher. 2006. "Trust and Trustworthiness in Games: An Experimental Study of Intergenerational Advice." *Experimental Economics* 9: 123-45.
- Simpson, Brent, Tucker McGrimmon, and Kyle Irwin. 2007. "Are Blacks Really Less Trusting Than Whites? Revisiting the Race and Trust Question." *Social Forces* 86: 525-52.
- Sommerfeld, Ralf D., Hans-Juergen Krambeck, and Manfred Milinski. 2008. "Multiple Gossip Statements and Their Effect on Reputation and Trustworthiness." *Proceedings of the Royal Society B-Biological Sciences* 275: 2529-36.
- Song, Fei. 2008. "Trust and Reciprocity Behavior and Behavioral Forecasts: Individuals Versus Group-Representatives." *Games and Economic Behavior* 62: 675-96.
- Stolle, Dietlind. 1998. "Bowling Together, Bowling Alone: The Development of Generalized Trust in Voluntary Associations." *Political Psychology* 19: 497-525.
- Sutter, Matthias, and Martin G. Kocher. 2007. "Trust and Trustworthiness across Different Age Groups." *Games and Economic Behavior* 59: 364-82.
- Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge: Cambridge University Press.
- Tomlin, Damon, M. Amin Kayali, Brooks King-Casas, Cedric Anen, Colin F. Camerer, Steven Quartz, and P. Read Montague. 2006. "Agent-Specific Responses in the Cingulate Cortex During Economic Exchanges." *Science* 312: 1047-50.
- Uslaner, Eric M. 2002. *The Moral Foundations of Trust*. Cambridge: Cambridge University Press.
- Van Swol, Lyn M. 2003. "The Effects of Regulation on Trust." *Basic and Applied Social Psychology* 25: 221-33.
- Wilson, Rick K., and Catherine C. Eckel. 2006. "Judging a Book by Its Cover: Beauty and Expectations in the Trust Game." *Political Research Quarterly* 59: 189-202.

Yamagishi, Toshio, Karen Cook, and Motoki Watabe. 1998. "Uncertainty, Trust, and Commitment Formation in the United States and Japan." *American Journal of Sociology* 104: 165-94.

Zak, Paul, and Stephen Knack. 2001. "Trust and Growth." *Economic Journal* 111: 295-321.

Zak, Paul J., Robert Kurzban, and William T. Matzner. 2005. "Oxytocin Is Associated with Human Trustworthiness." *Hormones and Behavior* 48: 522-27.

ⁱ Seonghui Lee and Cathy Tipton provided research assistance. Comments from Salvador Vázquez del Mercado, David Llanos, Skip Lupia, and Jamie Druckman significantly improved the essay. We gratefully acknowledge support by the National Science Foundation SES-0318116 and SES-0318180.

18. An Experimental Approach to Citizen Deliberation

Christopher F. Karpowitz and Tali Mendelberg¹

Deliberation has become, in the words of one scholar, “the most active area of political theory in its entirety” (Dryzek 2007, 237). Our exploration of the relationship between experiments and deliberation thus begins with normative theory as its starting point. Experiments can yield unique insights into the conditions under which the expectations of deliberative theorists are likely to be approximated, as well as the conditions under which theorists' expectations fall short. Done well, experiments demand an increased level of conceptual precision from researchers of all kinds who are interested in deliberative outcomes. But perhaps most importantly, experiments can shed greater scholarly light on the complex and sometimes conflicting mechanisms that may drive the outcomes of various deliberative processes. In other words, experiments allow researchers to better understand the extent to which, the ways in which, and under what circumstances it is actually *deliberation* that drives the outcomes deliberative theorists expect.

Our strategy for this chapter will be to highlight the strengths of experiments that have already been completed and to point to some aspects of the research that need further improvement and development. We aim to discuss what experiments can do that other forms of empirical research cannot and what experiments need to do in light of the normative theory.

Proceeding from normative theory is not without its difficulties, as deliberative theorists themselves admit (see, for example, Chambers 2003; Thompson 2008). One difficulty is that theories of deliberation offer a wide-ranging, sometimes vague, and not always completely consistent set of starting points for experimental work – as Diana Mutz ruefully observes, “it

may be fair to say that there as many definitions of deliberation as there are theorists” (2008, 525). We recognize, too, the inevitable slippage between theory and praxis that will lead almost every empirical test to be, in some sense, “incomplete” (Fishkin 1995). Finally, we agree that empirical researchers should avoid distorting the deeper logic of deliberative theory in the search for testable hypotheses (Thompson 2008). Experiments cannot “prove” or “disprove” theories of deliberation writ large. The critical question for experimental researchers, then, is not “does deliberation work?” but rather under what conditions does deliberation approach theorists’ goals or expectations?

The literature on deliberation is much too large to allow us to provide full coverage here. A variety of additional research traditions from social psychology and from sociology can usefully inform our attempts to explore deliberative dynamics (see Mendelberg 2002 for an overview), but we focus on *political* discussion.ⁱⁱ We aim to explore deliberation as practiced by ordinary citizens, which means we will set aside the literature on elites (Steiner et al. 2004). We set aside, too, a valuable research tradition focused on dyadic exchanges within social networks (Huckfeldt and Sprague 1995; Mutz 2006) to focus on discussions among groups, not dyads. Our focus on discussion of political issues and topics means that we will not cover the vast and influential research on deliberation in juries (see, for example, Hastie, Penrod, and Pennington 1983; Schkade, Sunstein, and Kahneman 2000; Devine et al 2001), though we emphasize the value and importance of that research. Finally, we cannot do justice to experiments derived from formal theories (see, for example, Meirowitz 2007 or Hafer and Landa 2007).

1. The Substantive Issues: Independent and Dependent Variables

One of the challenges of empirical research on deliberation is the multiplicity of potential definitions (Macedo 1999). Still, many definitions of deliberation share a commitment to a reason-centered, “egalitarian, reciprocal, reasonable, and open-minded exchange of language” (Mendelberg 2002, 153; see also Gutmann and Thompson 1996; Burkhalter, Gastil, and Kelshaw 2002; Chambers 2003). While theories of deliberation do not agree about each of deliberation’s constituent aspects or about all of its expected outcomes (Macedo 1999), it is possible to distill a working set of empirical claims about deliberation’s effects (see Mendelberg 2002; Mutz 2008).

Hibbing and Theiss-Morse (2002) summarize three broad categories of effects: deliberation should lead to “better *citizens*,” “better *decisions*,” and a “better (that is, more legitimate) *system*.” Benefits for individual *citizens* may include increased tolerance or generosity and a more empathetic view of others (Warren 1992; Gutmann and Thompson 1996, 2004); a decrease in the set of pathologies of public opinion documented extensively since Converse’s (1964) seminal work, leading to more political knowledge, an enhanced ability to formulate opinions, greater stability opinions, and more coherence among related opinions (Fishkin 1995); a better understanding of one’s own interests; an increased ability to justify preferences with well-considered arguments (Warren 1992; Chambers 1996, 2003); a better awareness of opponents’ arguments and an increased tendency to recognize the moral merit of opponents’ claims (Habermas 1989, 1996; Chambers 1996; Gutmann and Thompson 1996, 2004); a sense of empowerment, including among those who have the least (Fishkin 1995; Bohman 1997); a greater sense of public-spiritedness (Warren 1992) and an increased willingness to recognize community values and to compromise in the interest of the common good (Mansbridge 1983; Chambers 1996; but see Sanders 1997 and Young 2000); and a

tendency to participate more in public affairs (Barber 1984; Gastil, Deess, and Weiser 2002; Gastil, Deess, and Weiser 2008). Benefits for the quality of *decisions* flow from many of these individual benefits and include the idea that collective decisions or outcomes of deliberating groups will be grounded in increased knowledge, a more complete set of arguments, a fuller understanding of the reasons for disagreement, and a more generous aggregate attitude toward all groups in society, especially those who have the least (Chambers 1996; Gutmann and Thompson 2004).

As to the benefits for democratic *systems*, they center on the rise in support for the system that can follow from deliberation. Increased legitimacy for the system is a complex concept (Thompson 2008) but among other things, it can include a heightened level of trust in democratic processes and a greater sense of confidence that the process has been fairly carried out (see also Manin 1987). This sense becomes particularly important when the ultimate decision does not correspond well to an individual's pre-deliberation preferences or when there is a deep or longstanding conflict at issue (Mansbridge 1983; Benhabib 1996; Chambers 1996).

These laudable outcomes for citizens, decisions, and systems are rooted in a fourth claim: that deliberation is a better decision-making *process*, one that is more public-spirited, more reasonable, more satisfying, and ultimately more just than adversarial and aggregative forms of decision making (Mansbridge 1983; Chambers 1996; Gutmann and Thompson 2004). The process is a crucial mediating variable in deliberation. In other words, normative theory leads empirical investigators to ask what aspects of discourse and linguistic exchange leads to the individual-level outcomes we have described, and what in turn causes those aspects of interaction. Examples of mediating variables include the content and style of interaction, such as

whether deliberators use collective vocabulary such as “us” and ‘we” (Mendelberg and Karpowitz 2007), the number of arguments they make (Steiner et al. 2005), and the extent to which deliberators engage in a collaborative construction of meaning rather than speaking past each other (Rosenberg 2007). Focusing on such mediating variables allows scholars to investigate which aspects of the discourse cause those who have taken part in deliberation to feel that their voices were better heard, that the deliberating group functioned well as a collectivity, or that the process was more collaborative than other forms of interaction. In addition, researchers can investigate the effects of this sort of deliberative interaction for subsequent levels and forms of participation (see, for example, Karpowitz 2006; Gastil et al. 2008).

The variety of approaches to deliberative theory provides a rich set of procedural and substantive conditions, characteristics, and mechanisms that can be explored experimentally in a systematic way. Theorists differ, for example, on the desirability of consensus in deliberative procedures. The requirement of producing consensus is something that can be experimentally manipulated. In this way, empirical researchers can help to specify the relationship between the various potential characteristics of deliberation (the independent variables) and the positive outcomes (in other words, the dependent variables) theorists hope to see.

Of course, key independent variables relevant to deliberative outcomes may not only emerge from theory. They can also be found through careful attention to the real-world context of ongoing deliberative reform effort, where the variety of practices that might be subsumed under the broad heading of deliberation is extraordinary. Mansbridge (1983), Mansbridge et al. (2006), Polletta (2008), Cramer Walsh (2006; 2007), and Gastil (2008) are a few examples of

scholars who contribute to our understanding of deliberation's effects by insightful observation of real-world practices.

2. The Role of Experiments

Though sometimes styled as “experiments in deliberation,” much of the research to date has been purely observational – most often, these are case studies of specific deliberative events (e.g., Mansbridge 1983; Fishkin 1995; Eliasoph 1998; Fung 2003; Gastil and Levine 2005; Karpowitz 2006; Cramer Walsh 2006, 2007; Warren and Pearse 2008). The methods for evaluating and understanding deliberation are also diverse, including participant observation (Eliasoph 1998), survey research (Jacobs, Cook, and Delli Carpini 2009), and content analysis of discussion (see Gamson 1992; Conover, Searing, and Crewe 2002; Hibbing and Theiss-Morse 2002; Schildkraut 2005; Cramer Walsh 2006), just to name a few. Though our emphasis here is on experiments, other research designs are valuable in an iterative exchange with experiments and have a value of their own separate from experiments.

Experimentation is, as Campbell and Stanley put it, “the art of achieving interpretable comparisons” (quoted in Kinder and Palfrey 1993, 7). We follow the standard view that effective experimentation involves a high level of experimenter control over settings, treatments, and observations so as to rule out potential threats to valid inference. Like Kinder and Palfrey, we see random assignment to treatment and control groups as “unambiguously desirable features of experimental work in the social sciences” (1993, 7). In theory, random assignment may not be strictly necessary to achieve experimental control, but it does “provide a means of comparing the yields of different treatments in a manner that rule[s] out most alternative interpretations” (Cook and Campbell 1979, 5).

Much deliberation research is quasi-experimental (Campbell and Stanley 1963): it involves some elements of experimentation – a treatment, a subsequent outcome measure, and comparison across treated and untreated groups – but not random assignment. The lack of random assignment in quasi experiments places additional burdens on the researchers to sort out treatment effects from other potential causes of observed differences between groups. In other words, “quasi-experiments require making explicit the irrelevant causal forces hidden within the *ceteris paribus* of random assignment” (Cook and Campbell 1979, 6; see Esterling, Fung, and Lee 2010 for a sophisticated discussion of nonrandom assignment to conditions in a quasi-experimental setting).

The hallmark of experiments, then, is causal inference through control. Citizen deliberation lends itself quite well to control, since it need not take place within an official or even within a public context. Ordinarily, investigators are quite limited by the intimate link between the behavior we observe and fixed features of the political system. By contrast, citizens can deliberate with a set of strangers with whom they need have no prior connection, in contexts removed from organizational structures, official purviews or public spaces, and can reach decisions directed at no official or public target. In that sense, they present the experimenter a wide degree of control. Ultimately, deliberation often does exist in some dialogue with the political system. But unlike other forms of participation, it also exists outside of public and political spaces and it can therefore be separated and isolated for the study.

Experiments also allow us to measure the outcomes with a good deal more precision than nonexperimental, quasi-experimental, or field experimental studies. All groups assigned to deliberate do in fact deliberate; all groups assigned to deliberate by a particular rule do in fact

use that rule; all individuals assigned to receive information do in fact receive information; and so on. We can measure actual behavior rather than self-reported behavior. Under the fully controlled circumstances afforded by the experimental study of deliberation, we need not worry that people self select into any aspect of the treatment. Intent-to-treat problems largely disappear.

All of this boosts our causal inference considerably. We note that the causal inference is not a consequence of the simulation that control provides. Unlike observational studies or quasi-experimental studies, experiments can and do go beyond simple simulation in two respects: first, the control can help to verify that it is *deliberation* and not other influences creating the consequences that we observe. In addition, the control afforded by experiments allows us to test specific aspects of the deliberation, such as heterogeneous or homogenous group composition; the presence or absence of incentives that generate or dampen conflict of interests; the proportions of women, ethnic and racial groups; group size; the group's decision rule; the presence or absence of a group decision; the presence or absence of facilitators and the use of particular facilitation styles (Rosenberg 2007); the availability of information; or the presence of experts (Myers 2009).

In sum, the goal of experimental research should be to isolate specific causal forces and mediating or moderating mechanisms in order to understand the relationship between those mechanisms and deliberative outcomes as described by normative theorists. In addition, experiments allow us to reduce measurement error. This should lead to insights about how to better design institutions and deliberative settings.

3. Some Helpful Examples

To date, the experimental work on deliberation has been haphazard (Ryfe 2005, 64). Our aim here is not to provide an exhaustive account of every experiment, but to show the strengths and weaknesses of some of the work that has been completed. It is rare for a study to use all the elements of a strong experimental design. We find a continuum of research designs, with some studies falling closer to the gold standard of random assignment to control and treatments, and others falling closer to the category of quasi-experiments.

The best known and arguably the most influential investigation of deliberation is Fishkin's (1995) deliberative poll. In a deliberative poll, a probability sample of citizens is recruited and questioned about their policy views on a political issue. They are sent a balanced set of briefing materials prior to the deliberative event in order to spark some initial thinking about the issues. The representative sample is then brought to a single location for several days of intensive engagement, including small group discussion (with assignment to small groups usually done randomly), informal discussion among participants, a chance to question experts on the issue, and an opportunity to hear prominent politicians debate the issue. At the end of the event (and sometimes again several weeks or months afterward), the sample is asked again about their opinions, and researchers explore opinion change, which is presumed to be the result of the deliberative poll.

The first deliberative polls were criticized heavily on a variety of empirical grounds, with critics paying special attention to whether or not the deliberative poll should qualify as an experiment (Kohut 1996; Merkle 1996). Mitofsky (1996), for example, insists that problems with panel attrition in the response rates in the post-deliberation surveys made causal inferences especially difficult and that the lack of a control group made it impossible to know whether any

change in individual opinion “is due to the experience of being recruited, flown to Austin, treated like a celebrity by being asked their opinions on national television and having participated in the deliberations, or just due to being interviewed twice” (19). Luskin, Fishkin, and Jowell admit that their approach fails to qualify as a full experiment by the standards of Campbell and Stanley “both because it lacks the full measure of control characteristic of laboratory experiments and because it lacks a true, i.e. randomly assigned, control group” (Luskin, Fishkin, and Jowell 2002, 460).

As the number of deliberative polls has proliferated, they have pursued a variety of innovations. For example, subsequent work has included pre- and post-deliberation interviews of both those who were recruited to be part of the deliberating panel but who chose not to attend and post-event interviews of a separate sample of nondeliberators. These additional interviews function as a type of control group, though random assignment to deliberating or nondeliberating conditions is not present. While still not qualifying as a full experiment, these additions function as an “untreated nonequivalent control group design with pretest and posttest” and as “a posttest-only control group design,” as classified by Campbell and Stanley (1963). When such additions are included, the research design of the deliberative poll does have “some of the characteristics of a fairly sophisticated quasi-experiment” (Merkle 1996), characteristics that help eliminate some important threats to valid inference.

Still, the quasi-experimental deliberative poll does not exclude all threats to valid inference, especially when the problem of self-selection into actual attendance or nonattendance at the deliberative event is considered (see Barabas 2004, 692). Like that of recent deliberative polls, Barabas’s analysis of the effects of a deliberative forum about social security is based on

comparing control groups of nonattenders and a separate sample of nonattenders. To further reduce the potential for problematic inferences, Barabas makes use of propensity score analysis. Quasi-experimental research designs that make explicit the potential threats to inference or that use statistical approaches to estimate treatment effects more precisely are valuable advances (see also Esterling, Fung, and Lee 2010). These do not fully make up for the lack of randomization, but they do advance empirical work on deliberation.

We note one additional important challenge related to deliberative polling: the complexity of the deliberative treatment. As Luskin, Fishkin, and Jowell (2002) put it, the deliberative poll is “*one grand treatment*” that includes the anticipation of the event once the sample has been recruited, the exposure to briefing information, small-group discussion, listening to and asking questions of experts and politicians, informal conversations among participants over the course of the event, and a variety of other aspects of the experience (not the least of which is participants’ knowledge that they are being studied and will be featured on television). Perhaps it is the case that deliberation, as a concept, is a “grand treatment” that loses something when it is reduced to smaller facets, but from a methodological perspective, the complexity of this treatment makes it difficult to know what, exactly, is causing the effects we observe. Indeed, it denies us the ability to conclude that any aspect of *deliberation* is responsible for the effects (rather than the briefing materials, expert testimony, or some other nondeliberative aspects of the experience). Experimentation can and should seek to isolate the independent effects of each of these features.

At the other end of the spectrum from the “one grand treatment” approach of the deliberative pollsters are experiments that involve a much more spare conception of deliberation.

Simon and Sulkin (2002), for example, insert deliberation into a “divide-the-dollar” game in which participants were placed in groups of five and asked to divide sixty dollars between them. A total of 130 participants took part in one of eleven sessions, with multiple game rounds played at each session. In the game, each member of the group could make a proposal as to how to divide the money, after which a proposal was randomly selected and voted on by the group. A bare majority was sufficient to pass the proposal. Participants were randomly assigned to one of three conditions – no discussion, discussion prior to proposals, and discussion after proposals. In addition, participants were randomly assigned to either a cleavage condition in which players were randomly assigned to be in either a three-person majority or a two-person minority and proposals were required to divide money into two sums – one for the majority and one for the minority – or a noncleavage condition in which no majority/minority groups were assigned. Simon and Sulkin find that the presence of discussion led to more equitable outcomes for all participants and especially for players who ended up being in the minority.

The experiment employs many of the beneficial features we have highlighted – control over many aspects of the setting and of measurement, and random assignment to conditions. In addition, the researchers ground their questions in specific elements of normative theories. However, the study artificially capped discussion at only 200 seconds of online communication, which detracts from its ability to speak to the lengthier, deeper exchanges that deliberative theory deals with or to the nature of real-world exchanges.

Other experimental approaches have also explored the effects of online deliberation (see, for example, Muhlberger and Weber 2006 and developing work by Esterling, Neblo, and Lazer 2008a, 2008b).ⁱⁱⁱ The most well-known of these so far is the Healthcare Dialogue project

undertaken by Price and Capella (2005, 2007). A year-long longitudinal study with a nationally representative pool of citizens and a panel of healthcare policy elites, this study explored the effectiveness of online deliberations about public policy. The research involved repeated surveys and an experiment in which respondents who completed the baseline survey were randomly assigned to a series of four online discussions or to a nondeliberating control group. In these discussions, participants were stratified as either policy elites, healthcare issue public members (regular citizens who were very knowledgeable about health care issues), or members of the general public. Half of the groups were homogenous across strata for the first two conversations, the other half included discussants of all three types. In the second pair of conversations, half of the participants remained in the same kind of group as in the first wave; the other fifty percent of the participants were switched from homogenous to heterogeneous groups or *vice versa*. Group tasks were, first, to identify key problems related to health care and, second, to identify potential policy solutions (though they did not have to agree on a single solution). To ensure compatibility across groups, trained moderators followed a script to introduce topics and prompt discussion and debate.

Price and Capella find that participation in online discussion led to higher levels of opinion-holding among deliberators and a substantial shift in policy preferences, relative to those who did not deliberate. This shift was not merely the result of being exposed to policy elites, as the movement was greatest among those who did not converse with elites. In addition, participants – and especially nonelites – rated their experience with the deliberation as quite satisfying. Because a random subset of the nondeliberating control group was assigned to read online briefing papers that deliberating groups used to prepare for the discussions, the

experiments also allowed the researchers to distinguish the effects of information from the effects of discussion. While exposure to briefing materials alone increased knowledge of relevant facts, discussion and debate added something more – an increased understanding of the rationales behind various policy positions.

Regardless of whether the findings were positive or negative from the perspective of deliberative theory, the research design employed by Price and Capella highlights many of the virtues of thoughtful, sophisticated experiments. The research includes a large number of participants (nearly 2,500), a significant number of deliberating groups (more than eighty in the first wave and approximately fifty in the second wave), and random assignment from a single sample (those who completed the baseline survey) to deliberation plus information, information only, and no deliberation, no information conditions, with respondents in all groups completing a series of surveys over the course of a calendar year. Price and Capella also leverage experimental control to answer questions that the “grand treatment” approach of deliberative polling cannot. For example, where deliberative polling is unable to separate the independent effects of information, discussion among ordinary citizens, and exposure to elites, Price and Capella are able to show that information has differing effects from discussion and that exposure to elites cannot explain all aspects of citizens’ opinion change. Given large number of groups and the elements of the research design that have to do with differing group-level conditions for deliberation, the Price and Capella design has the potential for even more insight into the ways group-level factors influence deliberators and deliberative outcomes, though these have not been the primary focus of their analyses to date. Still, their design may fail to satisfy some conceptions

of deliberation, as groups simply had to identify potential solutions, not make a single, binding choice.

Experiments relevant to deliberation have also been conducted with face-to-face treatments, though we find considerable variation in the quality of the research design and the direct attention to deliberative theory. Morrell (1999), for example, contrasts familiar liberal democratic decision-making procedures, which include debate using Robert's rules of order, with what he calls "generative" procedures for democratic talk, which include such deliberatively desirable elements as hearing the perspectives of all group members, active listening and repeating the ideas of fellow group members, and considerable small group discussion. His research design includes random assignment to either the liberal democratic condition, the generative condition, or a no discussion condition. Participants answered a short survey about their political attitudes at the beginning and the end of the experimental process. Morrell repeats the study with multiple issues and with differing lengths of discursive interaction. In this research design, participants made a collective decision about an issue, an element that is not present in Fishkin's deliberative polls but that is critical to some theories of deliberation.

In contrast to the comparatively positive outcomes of the experiments in online deliberation we have highlighted, Morrell finds that the deliberatively superior generative procedures do not lead to greater group-level satisfaction or acceptance of group decisions. If anything, traditional parliamentary procedures are preferred in some cases. In addition, in several of the iterations of the experiment, Morrell finds strong mediating effects of the group outcome, contrary to deliberative expectations. Morrell's findings thus call attention to the fact that the

conditions of group discussion, including the rules for group interaction, matter a great deal and that more deliberative processes may not lead to the predicted normative outcomes.

Though we see important strengths in Morrell's approach, we note that the reported results do not speak directly to the value of the presence or absence of deliberation. The dependent variables Morrell reports are nearly all focused on satisfaction with group procedures and outcomes, measures for which the nondeliberating control condition are not relevant. In other words, Morrell's test as reported contrasts only different types of discursive interaction. Moreover, as with our earlier discussion of the "grand treatment", the treatments in both cases are complex, and it is not entirely clear which aspects of "generative" discussion led to lower levels of satisfaction. Finally, we note that Morrell's experiments were based on a very small number of participants and an even smaller number of deliberating groups. All this makes comparison to other, conflicting studies difficult.

Druckman's (2004) study of the role of deliberation in combating framing effects is a good example of the way experiments can speak to aspects of deliberative theory. The primary purpose of Druckman's research is to explore the conditions under which individuals might be less vulnerable to well-recognized framing effects. The relevance to deliberation lies in investigating how deliberation can mitigate the irrationality of ordinary citizens and improve their civic capacities. The study advances the literature on deliberation by assessing the impact of different deliberative contexts.

Druckman presented participants with one of eight randomly assigned conditions. These conditions varied the nature of the frame (positive or negative) and the context in which the participant received the frame. Contexts included: a control condition, in which participants

received only a single, randomly chosen frame; a counter-framing condition, in which participants received both a positive and a negative frame; and two group conditions, in which participants had an opportunity to discuss the framing problems with three other participants. In the homogenous group condition, all members of the group received the same frame, and in the heterogeneous condition, half of the group received a positive frame and half received the negative frame. Participants in the group condition were instructed to discuss the framing problem for five minutes. Druckman recruited a moderate number of participants (580), with approximately 172 taking part in the group discussion conditions. This means that just over forty deliberating groups could be studied.

As with the Simon and Sulkin experiment, exposure to group discussion in this research design is limited and may, therefore, understate the effect group discussion might have. But what is most helpful from the perspective of deliberative theory is a systematic manipulation of both the presence of discussion and the context under which discussion occurred. Druckman finds that the presence of discussion matters – participants in both the homogenous and heterogeneous conditions proved less vulnerable to framing effects than in the control condition. This would seem to be positive evidence for the relationship between deliberation and rationality, but the story is somewhat more complicated than that. Neither discussion condition reduced framing effects as much as simply giving the counterframe to each individual without requiring group discussion. Moreover, homogenous groups appeared to be comparatively more vulnerable to framing effects compared to heterogeneous groups. Results were also strongly mediated by expertise.

Druckman's research design reflects several attributes worthy of emulation. First, the number of groups is sufficiently sizeable for meaningful statistical inference. Second, Druckman has used the key levers of experimental control and random assignment appropriately. This allows him to make meaningful claims about the difference between discussion across different contexts and the difference between discussion and the simple provision of additional information. As we discussed in the previous section, one of the key problems of causal inference in the "grand treatment" design has been whether deliberation is responsible for the observed effects or whether one particular aspect of it – the provision of information – is responsible. Given that information is not unique to deliberation, finding that the effects of deliberation are due primarily to information would considerably lessen the appeal and value of deliberation as a distinct mode of participation. Druckman's results do raise further questions, however, especially with respect to what is actually happening during the discussion period. Druckman does not look inside the "black box" of discussion to understand how the dynamics and the content of discussion vary across the homogenous and heterogeneous conditions.

Finally, we add a few words about our own experimental work on deliberation (Mendelberg and Karpowitz 2007; Karpowitz and Mendelberg 2007; Karpowitz, Mendelberg, and Argyle 2008). We do this in order to highlight a few of the methodological issues that have emerged as we have conducted the research. Our interest in experiments began when we reanalyzed data collected earlier by Frohlich and Oppenheimer (1992). Participants in the experiment were told that they would be doing tasks to earn money; that the money they earned would be based on a group decision about redistribution; but that prior to group deliberation, they would not be told the nature of the work they would be doing. This was meant to simulate

the Rawlsian veil of ignorance, as individuals would not know the specifics of how their decision would affect them personally because they would not know how well or poorly they might perform.

During deliberation, groups were instructed to choose one of several principles of justice to be applied to their earnings, including the option not to redistribute at all. The principle chosen would simultaneously govern the income they earned during the experiment (which was translated into a yearly income equivalent) and apply (hypothetically) to the society at large. Groups were randomly assigned to one of three conditions: imposed, unanimous, and majority rule. In the imposed condition, groups were assigned a principle of justice by the experimenters. In the unanimous and majority conditions, groups had to choose a principle of justice either unanimously or by majority vote, respectively.

The key finding of Frohlich and Oppenheimer's original study was that, when given an opportunity to deliberate behind the veil of ignorance, most groups choose to guarantee a minimum income below which the worst-off member of the group would not be allowed to fall. In our reanalysis of their data (Mendelberg and Karpowitz 2007), we noted that Frohlich and Oppenheimer paid very little attention to the ways in which the group context shaped participants' attitudes and group-level outcomes. Our reanalysis showed that important features of the group context, such as the group's gender composition and its decision rule, interacted to significantly affect group- and individual-level experimental outcomes. But the earlier data were also limited to a significant extent. First, participants in that experiment were not randomly assigned to conditions. In addition, the data did not include a sufficient number of groups of varying gender composition to be entirely confident in our statistical results.

For those reasons, we chose to conduct our own updated version of the experiment, this time with random assignment, a sufficient number of groups (nearly 150), and systematic manipulation of the various gender/decision rule conditions. We also carefully recorded each group discussion in order to explore more fully the dynamics of the group interactions themselves, tying the verbal behavior of each participant during deliberation to their pre- and post-discussion attitudes about the functioning of the group, the need for redistribution, and a host of other variables. Our analysis is still in its initial stages, but we do find evidence that the group-level factors, especially the interaction of group gender and decision rule, affect various aspects of the group's functioning and deliberative dynamics. We also find significant differences between groups that deliberate and control groups that did not.

In sum, in our work we have attempted to advance the study of deliberation methodologically in several ways. We use a larger N, particularly increasing the group N; we employ random assignment; and our design both controls on deliberation itself and isolates the effects of specific aspects of deliberation, some of which derive from empirical studies of citizen discussion (decision rule, the group's heterogeneity or homogeneity), some of which focus on normatively relevant processes of communication (such as equal participation in discussion, use of linguistic terms reflecting a concern for the common good), and some of which supplement these theories by focusing on sociologically important variables such as the group's demographic composition. In conducting these experiments, we have also begun to confront directly some of the practical challenges inherent in attempting to implement random assignment of individuals to group conditions.

Having outlined both positive features and further questions that emerge from several highlighted experiments, we turn next to some of the challenges of effective experimentation about deliberation.

4. Challenges

We have detailed an argument in which we urge more investigations using experimental methods, more care in designing treatments that manipulate various aspects of deliberation, and particularly more frequent use of random assignment to conditions. However, the more control the investigator seeks, the greater the tradeoffs. Control brings artifice and narrow, isolated operationalizations of rich and complex concepts. The behavior of interest is often embedded in the contexts of institutions and social relationships and must therefore ultimately be moved back out of the lab, where every effort was made to isolate it, and studied all over again with attention to these contexts. There are other difficulties involved in the use of experiments – they may be more expensive and effortful than other methods. Here we consider these tradeoffs.

One of the challenges of experimentation is the operationalization of idealized normative theory. Experimental approaches may be particularly vulnerable to the disagreements between theorists and empiricists to the extent that their heightened levels of control bring more stylized and more artificial operationalizations of complex and multifaceted theoretical concepts. We illustrated the issue with a contrast between Fishkin’s “grand treatment” versus Simon and Sulkin’s decision to trade off the complexity of deliberation against the ability to control it. The tradeoff is understandable, but not necessary. It is possible to design a controlled experiment with random assignment with multiple conditions, one of which resembles the “grand treatment”

notion, others of which isolate each of the major elements of the deliberation, and with a control condition identical in every way but lacking all of these elements.

A related second challenge for experimentation is external validity. It is difficult not only to adequately operationalize the key concepts of normative theory – to achieve construct validity (Campbell and Stanley (1963) – but to simulate the causal relationships as they occur in the real world. We need to know ultimately how deliberative efforts interact with real-world actors and institutions. For example, Karpowitz's (2006) study of a local civic deliberation suggests that the deliberators' knowledge that they could pursue their preferences after the deliberation was over by lobbying the city council, writing letters to newspapers, and filing law suits in the courts, significantly affected various aspects of deliberation, including the ability of the deliberation to change minds, enlarge interests, resolve conflicts, and achieve other ends envisioned by normative theorists. This presents a challenge to the external validity of experiments in that their deliberative situation is abstracted from interaction with real institutions. On the other hand, the cumulation of findings such as Karpowitz's from observational studies can, in turn, lead to further hypothesis testing using experimental designs, where the impact of particular institutional contexts can be isolated and studied rigorously.

Mansbridge's (1983) study of a New England town meeting is the classic example of how a careful observational study can lead to theoretically rich insights into group discussion and decision making as practiced in the real world. Mansbridge shows, for example, how residents of the town struggle to navigate their common and conflicting interests in group settings and how the presence or absence of conflicting interests shapes the dynamics of the discussion and patterns of attendance at the town meeting. The textured details of real-world observation found

in Mansbridge's work are often lacking in experimental studies, but her work can also be seen as articulating a set of hypotheses that can be explored much more deeply with the control that experiments provide.

Another aspect of external validity is that, in real-world settings, citizens often have to choose to deliberate. Karpowitz's (2006) analysis of patterns of meeting attendance in a national sample suggests that people who attend meetings are not a random sample of the adult population – they are more opinionated (though not more ideologically extreme) than nonattenders, but also more interested in politics, more knowledgeable about it, and more likely to discuss political issues frequently. In a controlled experiment, people also exercise some level of choice as to whether to participate, but to a much lesser extent. An experiment described up front as focused on deliberation may better approximate a real-world setting in which people choose to participate in deliberation, and people may choose to participate in the experiment for the same reasons they choose to participate in real-world deliberations. But some deliberation experiments may not be described that way. The question then becomes to what extent are the processes and effects of deliberation generalizable from the sample in the experiment to the samples in the real world.

Attending to the relationship between deliberating groups and the wider political context, and to the differences between those who choose to deliberate and those who do not, also raises the question of how deliberation might affect those who do not participate directly, but who view the deliberation of others or who merely read about the work of deliberating groups. Given the problem of scale, deliberation is unlikely to be all-inclusive, and those who sponsor opportunities for deliberation must also communicate their processes and results to the wider public. How

those who were not part of the discussion understand deliberating groups is a topic worth considerable additional study, including with experimental approaches.

A third and final difficulty of randomized experiments is of a practical nature. It is extremely difficult to implement random assignment in the study of deliberation. One variant of this problem comes in the “grand treatment” design. There, the holistic treatment and attempt to approximate the ideal conditions specified by normative theory require a substantial commitment of time and effort by deliberators. A significant percentage of those assigned to a demanding deliberation condition may well refuse treatment, and the decision to drop out of the treatment condition may well be nonrandom, introducing bias into the estimates of causal effects. Lab-based deliberation experiments may face less severe problems because random assignment takes place after subjects come to the lab, so that participants are less likely to opt out of the treatment due to its demanding nature.

Another variant of this problem presents itself when variables of interest are at the group level. This requires a large number of groups, which in turn requires a much larger individual n than lab experiments typically use. The practical challenges of random assignment to group conditions can be significant, especially when potential participants face a variety of different time constraints. For example, in our work on group composition and deliberation, we found that simultaneously accounting for differences in participants’ availability and instituting a random assignment procedure that ensures each recruited participant a roughly equal chance of being assigned to all relevant groups is a complicated exercise.

5. Conclusion – What’s Next?

We began with the notion that empirical research can usefully evaluate the claims of deliberative theorists, and we have developed an argument about the special utility of controlled experiments. The control afforded by experiments allows not only strong causal inference but also the ability to measure, and therefore to study, mediating and outcome variables with a heightened level of precision and accuracy. We have argued that despite a proliferation of self-titled deliberative “experiments,” methodologically rigorous research design with sufficient control and random assignment is still a relative rarity. We are anxious to see experiments with an increased number of participants and especially an increased number of groups. Experimental approaches can also use their high level of control to measure the exchange of language – that is, we can train the analytical microscope more directly on the process of deliberation itself, though this practice is also still rare.

While experimental control allows for unique causal inference, experiments miss some of the richness of real-world deliberative settings. In-depth observational case studies can fill the gap and uncover the meaning of key concepts (e.g., Mansbridge 1983; Eliasoph 1998). Indeed if we were forced to choose between Mansbridge’s classic work and many experiments, we might prefer Mansbridge’s. The ideal research design is an iterative process in which experimentation in the lab is supplemented and informed by observation of real-world settings. Our ecumenism is not, however, a call for a continuation of the hodgepodge of studies that currently characterizes the field. Instead, we need a tighter link between the variables observed in real-world discussions and those manipulated in controlled settings. In addition, the external validity concerns typical of experiments generally apply in the case of deliberation, and may be addressed by supplementing controlled experiments with field experiments that make use of the explosion of deliberative

reform efforts in cities and towns across the United States. Field experiments will be especially helpful if they allow the investigator access to accurate measures of mediating and outcome variables. It is unclear whether they do in fact allow such a degree of access or not, but the effort is worth making.

References

- Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98: 687-701.
- Barber, Benjamin. 1984. *Strong Democracy: Participation politics for a new age*. Berkeley: University of California Press.
- Benhabib, Seyla. 1996. "Toward a Deliberative Model of Democratic Legitimacy." In *Democracy and Difference: Contesting the Boundaries of the Political*, ed. Seyla Benhabib. Princeton, NJ: Princeton University Press.
- Bohman, James. 1997. "Deliberative Democracy and Effective Social Freedom: Resources, Opportunities and Capabilities." In *Deliberative Democracy*, eds. James Bohman, and William Rehg. Boston, MA: MIT Press.
- Burkhalter, Stephanie, John Gastil, and Todd Kelshaw. 2002. "A Conceptual Definition and Theoretical Model of Public Deliberation in Small Face-to-Face Groups." *Communication Theory* 12: 398-422.
- Campbell, Donald T., and Julian Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. New York: Wadsworth Publishing.
- Chambers, Simone. 1996. *Reasonable Democracy*. Ithaca: Cornell University Press.
- Chambers, Simone. 2003. "Deliberative Democratic Theory." *Annual Review of Political Science* 6: 307-26.
- Conover, Pamela J., David D. Searing, and Ivor Crewe. 2002. "The Deliberative Potential of Political Discussion." *British Journal of Political Science* 32: 21-62.
- Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. David E. Apter. New York: Free Press.
- Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. New York: Wadsworth Publishing.

- Cramer Walsh, Katherine. 2006. "Communities, Race, and Talk: An Analysis of the Occurrence of Civic Intergroup Dialogue Programs." *Journal of Politics* 68: 22-33.
- Cramer Walsh, Katherine. 2007. *Talking About Race: Community Dialogues and the Politics of Difference*. Chicago: University of Chicago Press 0073
- Devine, Dennis J., Laura Clayton, Benjamin B. Dunford, Rasmy Seying, and Jennifer Price. 2001. "Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups." *Psychology, Public Policy, and Law* 73: 622-727.
- Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98: 671-86.
- Dryzek, John S. 2007. Theory, Evidence, and the Tasks of Deliberation. In *Deliberation, Participation and Democracy: Can the People Govern?*, ed. Shawn W. Rosenberg. Basingstoke, UK: Palgrave Macmillan.
- Eliasoph, Nina. 1998. *Avoiding Politics How Americans Produce Apathy in Everyday Life*. Cambridge, UK: Cambridge University Press.
- Esterling, Kevin, Archon Fung, and Taeku Lee. 2010. "How Much Disagreement Is Good for Democratic Deliberation? The California Speaks Health care Reform Experiment." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Esterling, Kevin, Michael Neblo, and David Lazer. 2008a. "Estimating Treatment Effects in the Presence of Selection on Unobservables: The Generalized Endogenous Treatment Model." Paper presented at the annual meeting of the American Political Science Association, Chicago, IL.
- Esterling, Kevin, Michael Neblo, and David Lazer. 2008b. "Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment." Paper presented at DG.O2007, the annual meeting of the Digital Government Society, Philadelphia, PA.
- Fishkin, James S. 1995. *The Voice of the People*. New Haven, CT: Yale University Press.
- Frohlich, Norman, and Joe Oppenheimer. 1992. *Choosing Justice*. Cambridge, UK: Cambridge University Press.
- Fung, Archon. 2003. "Deliberative Democracy and International Labor Standards." *Governance* 16: 51-71.
- Gamson, William A. 1992. *Talking Politics*. Cambridge, UK: Cambridge University Press.
- Gaertner, Samuel L., John F. Dovidio, Mary C. Rust, Jason A. Nier, Brenda S. Banker, Christine M. Ward, Gary R. Mottola, and Missy Houlette. 1999. "Reducing Intergroup Bias:

- Elements of Intergroup Cooperation.” *Journal of Personality and Social Psychology* 76: 388–402.
- Gastil, John. 2008. *Political Communication and Deliberation*. Thousand Oaks, CA: Sage.
- Gastil, John, E. Pierre Deess, and Phil Weiser. 2002. “Civic Awakening in the Jury Room: A Test of the Connection Between Jury Deliberation and Political Participation.” *Journal of Politics* 64: 585-95.
- Gastil, John, E. Pierre Deess, Phil Weiser, and Jordan Meade. 2008. “Jury Service and Electoral Participation: A Test of the Participation Hypothesis.” *Journal of Politics* 70: 1-16.
- Gastil, John, and Peter Levine. 2005. *The Deliberative Democracy Handbook: Strategies for Effective Civic Engagement in the Twenty-First Century*. San Francisco, CA: Jossey Bass.
- Gutmann, Amy, and Dennis Thompson. 1996. *Democracy and Disagreement*. Cambridge, MA: Harvard University Press.
- Gutmann, Amy, and Dennis Thompson. 2004. *Why Deliberative Democracy?* Princeton, NJ: Princeton University.
- Habermas, Jürgen. 1989. *The Structural Transformation of The Public Sphere: An Inquiry Into a Category of Bourgeois Society*, trans. by T. Burger with the assistance of F. Lawrence. Cambridge, MA: MIT Press.
- Habermas, Jürgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge, MA: MIT Press.
- Hafer, Catherine, and Dimitri Landa. 2007. “Deliberation as Self-discovery and Institutions for Political Speech.” *Journal of Theoretical Politics* 19: 325-55.
- Hastie, Reid, Steven D. Penrod, and Nancy Pennington. 1983. *Inside the Jury*. Cambridge, MA: Harvard University Press.
- Hibbing, John, and Elizabeth Theiss-Morse. 2002. “The Perils of Voice: Political Involvement’s Potential to Delegitimate.” Presented at the annual meeting of the American Political Science Association, Boston, MA.
- Huckfeldt, Robert, and John Sprague. 1995. *Citizens, Politics and Social Communication: Information and Influence in an Election Campaign*. New York: Cambridge University Press.
- Jacobs, Lawrence, Fay Lomax Cook, and Michael X. Delli Carpini. 2009. *Talking Together: Public Deliberation and Politics in America*. Chicago: University of Chicago Press.

- Karpowitz, Christopher F. 2006. "Having a Say: Public Hearings, Deliberation, and American Democracy." Unpublished dissertation, Princeton University.
- Karpowitz, Christopher F., and Tali Mendelberg. 2007. "Groups and Deliberation." *Swiss Political Science Review* 13: 645-62.
- Karpowitz, Christopher, Tali Mendelberg, and Lisa Argyle. 2008. "Group Effects and Deliberation: The Deliberative Justice Experiment." Paper presented at the International Society for Political Psychology annual scientific meeting, Sciences Po, Paris, France.
- Kinder, Donald R., and Thomas R. Palfrey. 1993. *Experimental Foundations of Political Science*. Ann Arbor, MI: University of Michigan Press.
- Kohut, Andrew. 1996. "The Big Poll That Didn't." *Poll Watch* 4: 2-3.
- Luskin, Robert C., James S. Fishkin, and Roger Jowell. 2002. "Considered Opinions: Deliberative Polling in Britain." *British Journal of Political Science* 32: 455-87.
- Macedo, Stephen. 1999. Introduction. In *Deliberative Politics: Essays on Democracy and Disagreement*, ed. Stephen Macedo. New York: Oxford University Press.
- Manin, Bernard. 1987. "On Legitimacy and Political Deliberation." *Political Theory* 15: 338-68.
- Mansbridge, Jane. 1983. *Beyond Adversary Democracy*. Chicago: University of Chicago Press.
- Mansbridge, Jane, Janette Hartz-Karp, Matthew Amengual, and John Gastil. 2006. "Norms of deliberation: An inductive study." *Journal of Public Deliberation* 2: Article 7.
- Meirowitz, Adam. 2007. "In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values." *Journal of Theoretical Politics* 19: 301-27.
- Mendelberg, Tali. 2002. "The Deliberative Citizen: Theory and Evidence." In *Political Decision Making, Deliberation and Participation: Research in Micropolitics*, Vol. 6, eds. Michael X. Delli Carpini, Leonie Huddy, and Robert Y. Shapiro. Greenwich, CT: JAI Press.
- Mendelberg, Tali, and Christopher Karpowitz. 2007. "How People Deliberate about Justice." In *Deliberation, Participation and Democracy: Can the People Govern?*, ed. Shawn W. Rosenberg. Basingstoke, UK: Palgrave Macmillan.
- Merkle, Daniel M. 1996. "The National Issues Convention Deliberative Poll." *Public Opinion Quarterly* 60: 588-619.
- Mitofsky, Warren J. 1996. "It's Not Deliberative and It's Not a Poll." *Public Perspective* 7: 4-6.
- Morrell, Michael. 1999. "Citizens' Evaluations of Participatory Democratic Procedures: Normative Theory Meets Empirical Science." *Political Research Quarterly* 52: 293-322.

- Muhlberger, Peter, and L. M. Weber. 2006. "Lessons from the Virtual Agora Project: The Effects of Agency, Identity, Information, and Deliberation on Political Knowledge." *Journal of Public Deliberation* 2: Article 6.
- Mutz, Diana C. 2006. *Hearing the Other Side: Deliberative vs Participatory Democracy*. New York: Cambridge University Press.
- Mutz, Diana C. 2008. "Is Deliberative Democracy a Falsifiable Theory?" *Annual Review of Political Science* 11: 521-38.
- Myers, C. Daniel. 2009. "Information Sharing in Democratic Deliberation." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Polletta, Francesca. 2008. "Just Talk: Public Deliberation after 9/11." *Journal of Public Deliberation* 4: 1-24.
- Price, Vince, and Joe N. Cappella. 2005. "Constructing Electronic Interactions Among Citizens, Issue Publics, and Elites: The Healthcare Dialogue Project." Proceedings of the National Conference on Digital Government Research, Atlanta, GA. Digital Government Research Center.
- Price, Vince, and Joe N. Cappella. 2007. "Healthcare Dialogue: Project highlights". Proceedings of the National Conference on Digital Government Research. Philadelphia, PA: Digital Government Research Center.
- Rosenberg, Shawn W., 2007. *Deliberation, Participation and Democracy: Can the People Govern?* Basingstoke, UK: Palgrave Macmillan.
- Ryfe, David M. 2005. "Does Deliberative Democracy Work?" *Annual Review of Political Science* 8: 49–71.
- Sanders, Lynn M. 1997. "Against Deliberation." *Political Theory* 25: 347-76.
- Schildkraut, Deborah. 2005. *Press One for English: Language Policy, Public Opinion, and American Identity*. Princeton, NJ: Princeton University Press.
- Schkade, David, Cass R. Sunstein, and Daniel Kahneman. 2000. "Deliberating about Dollars: The Severity Shift." *Columbia Law Review* 100: 1139-76.
- Simon, Adam, and Tracy Sulkin. 2002. "Discussion's Impact on Political Allocations: An Experimental Approach." *Political Analysis* 10: 402-11.
- Steiner, Jurg, Andre Bachtiger, Markus Spornli, and Marco Steenbergen. 2005. *Deliberative Politics in Action: Analyzing Parliamentary Discourse (Theories of Institutional Design)*. Cambridge, UK: Cambridge University Press.

Thompson, Dennis 2008. "Deliberative democratic theory and empirical political science." *Annual Review of Political Science* 11: 497–520.

Warren, Mark E. 1992. "Democratic Theory and Self-Transformation." *American Political Science Review* 86: 8-23.

Warren, Mark E., and Hilary Pearse. 2008. Introduction. In *Designing Deliberative Democracy: The British Columbia Citizens' Assembly*, eds. Mark E. Warren, and Hilary Pearse. Cambridge, UK: Cambridge University Press.

Young, Iris Marion. 2000. *Inclusion and Democracy*. New York: Oxford University Press.

ⁱ We thank Lee Shaker for invaluable research assistance.

ⁱⁱ See Gaertner et al. (1999) for a particularly informative study.

ⁱⁱⁱ See also Esterling, Neblo, and Lazer (2008a) on estimating treatment effects in the presence of noncompliance with assigned treatments and nonresponse to outcome measures.

19. Social Networks and Political Context

David W. Nickerson

People are embedded in networks, neighborhoods, and relationships. Understanding the nature of our entanglements and how they shape who we are is fundamental to *social* sciences. Networks are likely to explain important parts of personal development and contemporary decision making. Researchers have found social networks to be important in activities as disparate as voting (Berelson, Lazarsfeld, and McPhee 1954), immigration patterns (Sanders, Nee, and Sernau 2002), finding a job (Nordenmark 1999), recycling (Tucker 1999), de-worming (Miguel and Kremer 2004), cardiovascular disease and mortality (Kawachi et al. 1996), writing legislation (Caldeira and Patterson 1987), and even happiness (Fowler and Christakis 2008). A wide range of political outcomes could be studied using social networks; the only limitation is that the outcome be measurable. Ironically, the very ubiquity and importance of social networks make it very difficult to study. Isolating causal effects is always difficult, but when like-minded individuals cluster together, share material incentives, are exposed to common external stimuli, and simultaneously influence each other, the job of reliably estimating the importance of social ties becomes nearly impossible. Rather than offering a comprehensive overview of the wide number of topics covered by social networks, this chapter focuses on the common empirical challenges faced by studies of social networks by: considering the challenges faced by observational studies of social networks; discussing laboratory approaches to networks; and describing how network experiments are conducted in the field. The chapter concludes by

summarizing the strengths and weaknesses of the different approaches and considers directions for future work.

1. Observational studies

Cross-sections

The social networks literature blossomed during the 1940s and 50s with works such as *The People's Choice* (Lazarsfeld, Berelson, and Gaudet 1948) and *Voting* (Berelson et al. 1954). Utilizing newly improved survey technology, the authors administered surveys to randomly selected respondents densely clustered in medium-sized communities. This strategy provided insight into what the neighbors and friends of a respondent believed at the same point in time, allowing correlations in the behaviors and beliefs of friends and neighbors to be measured. The authors found that information flowed horizontally through networks and overturned the opinion leadership model of media effects.

The advent of affordable nationally representative polling largely ended this mode of inquiry. Why study one community when you could study an entire nation? Unfortunately, the individuals surveyed from around the nation had no connection to one another, so the theories developed based on this data generally assumed atomistic voters (for example, Campbell et al. 1964). The political context literature was revived only when Huckfeldt and Sprague returned to the strategy of densely clustering surveys in communities, while adding a new methodological innovation. Huckfeldt and Sprague used snowball surveys, where respondents were asked to name political discussants, who were then surveyed themselves. This technique allowed Huckfeldt and Sprague to measure directly the political views in a person's network rather than infer the beliefs of discussants from neighborhood characteristics. In a series of classic articles

and books, Huckfeldt and Sprague and their many students meticulously documented the degree to which political engagement is a social process for most people (for example, Huckfeldt 1983; Huckfeldt and Sprague 1995). It is fair to say that most contemporary observational studies of political behavior and social networks either rely on survey questions to map social networks or ask questions about politically relevant conversations (for example, Mutz 1998).

These empirical strategies face three primary inferential hurdles making it difficult to account for all plausible alternative causes of correlation. First, people with similar statuses, values, and habits are more likely to form friendships (Lazarsfeld and Merton 1954), so self-sorting rather than influence could drive results. Second, members of a social network are likely to share utility functions and engage in similar behaviors independently of one another. Third, members of social networks are exposed to many of the same external stimuli (for example, media coverage, economic conditions, political events). If the external stimuli influence members of the social network similarly, then observed correlations could be due to these outside pressures rather than the effect of the network. All of these problems can be categorized as forms of omitted variable bias¹ and call into question results based on cross-sectional surveys. Within the framework of a cross-sectional study, it is difficult to conceive of data that could convince a skeptic that the reported effects are not spurious.

Panels

Part of the problem is that influence within a network is an inherently dynamic process. A person begins with an attitude or propensity for a given behavior and the network acts upon this baseline. Observational researchers can improve their modeling of the process by collecting panel data where the same set of individuals is followed over time through multiple waves of a

survey. This strategy allows the researcher to account for baseline tendencies for all respondents and measure movement away from these baselines. Moreover, all measured and unmeasured attributes of a person can be accounted for by including fixed effects for each individual in the sample. In this way, panel data can account for all time-invariant confounding factors.

Panel studies in political science are rare because of the expense involved. Studies focusing on social networks are even less common and nearly always examine families – one of the most fundamental networks in society. Jennings and Niemi's (1974, 1981) classic survey of families over time is the best known panel in political science examining how political attitudes are transferred from parents to children (and vice versa). More often, political scientists are forced to rely on a handful of politically relevant questions in panel studies conducted for other purposes (for example, Zuckerman, Dasovic, and Fitzgerald 2007).

Although they provide a huge advance over cross-sectional data, panel data cannot provide fully satisfactory answers. Even if the type of networks considered could be broadened, dynamic confounding factors, such as congruent utility functions, life-cycle processes, and similar exposure to external stimuli, remain problematic. Furthermore, if the baseline attitudes and propensities are measured with error and that error is correlated with politically relevant quantities, then the chief advantage of panel data is removed because the dynamic analysis will be biased.

Network Analysis

Network analysis is touted as a method to analyze network data to uncover the relationships within a network. Sophisticated econometric techniques have been developed to measure the strength of ties within networks and their effects on various outcomes (Carrington,

Scott, and Wasserman 2005). Instead of assuming the independence of observations, network models adjust estimated coefficients to account for correlations found among other observations with ties to each other. Network analysis is a statistical advancement, but it does not surmount the core empirical challenge facing observational studies of social networks, which is essentially a data problem. Similar utility functions and exposure to external stimuli remain problems, as do selection effects. Selection effects are not only present, but are reified in the model and analysis being used to define the nodes and ties of the network.

To illustrate the challenge facing observational studies of social networks, consider the recent work by Christakis and Fowler (2007, 2008a,b) using the Framingham Heart Study. To supplement health and behavioral data collected since 1948, the Framingham Heart Study began collecting detailed social network data in 1971. Taking advantage of the panel and network structure of the data, Christakis and Fowler found evidence that obesity, smoking, and happiness were contagious. While the claim is entirely plausible, there are three reasons to question the evidence provided and the strength of observed the relationship. First, unobserved factors that influence both alters and egos could drive the results. Second, the strength of the relationship detected violates a few causal models. For instance, Christakis and Fowler (2007) find “geographic distance did not modify the intensity of the effect of the alter’s obesity on the ego” (377). The primary mechanism for jointly gaining and losing weight would presumably be shared meals or calorie-burning activities like walking or perhaps competitive pressure to remain thin. However, none of these mechanisms work for geographically distant individuals, raising the concern that selection bias is driving the results.ⁱⁱ Third, Cohen-Cole and Fletcher (2008) adopt a similar empirical strategy on a similar dataset to Christakis and Fowler and find evidence that

acne and height are also contagious, which constitute failed placebo tests. None of these points disprove the claims by Christakis and Fowler, but they do call into doubt the evidence provided and the strength of the relationships detected.

The Framingham Heart Study is a nearly perfect observational social network dataset. If the answers provided remain unconvincing, then perhaps the observational strategy should be rejected in favor of techniques using randomized experiments. Ordinarily, experiments can get around problems of self-selection and unobserved confounding factors through randomization, but the organic nature of most social networks pose a difficult problem. To test the power of social networks, the ideal experiment would place randomly selected individuals in a range of varying political contexts or social networks. The practical and ethical concerns of moving people around and enforcing friendships are obvious. The time-dependent nature of social networks also makes them inherently difficult to manipulate. Reputation and friendships take a long time to develop and cannot be manufactured and manipulated in any straightforward manner. Thus, the experimental literature testing the effect of social networks on behaviors and beliefs is still in its infancy. Having said that, the next section discusses the laboratory tradition that began in the 1950s.

2. Laboratory Experiments

Assign Context

Many tactics have been used to study social networks in laboratories. The central logic behind all of them is for the researcher to situate subjects in a randomly assigned social context. One of the most famous examples is Asch's (1956) series of classic experiments on conformity. Subjects were invited to participate in an experiment on perception where they had to judge the

length of lines. Control subjects performed the task alone, while subjects in the treatment group interacted with confederates who guessed incorrectly. Subjects in the control group rarely made mistakes, while individuals in the treatment group parroted the errors of the confederates frequently. The initial study was criticized for relying on a subject pool of male undergraduates, who may not be representative of the population as a whole. However, the Asch experiments have been replicated hundreds of times in different settings (Bond and Smith 1996). While the conformity effect persists, it: a) varies across cultures; b) is stronger for women; c) has grown weaker in the United States over time; and d) depends on parts of the experimental design (for example, size of the majority, ambiguity of stimuli) and not others (for example, whether the subject's vote is public or private). Thus, the Asch experiments constitute evidence that peers – even ones encountered for the first time – can shape behavior.

Much of the literature in psychology employs tactics similar to those used by Asch. For instance, social loafing (Karau and Williams 1993) and social facilitation (Bond and Titus 1983) can boast equally long pedigrees and replications.ⁱⁱⁱ While these experiments measure conformity, how strongly the findings apply in real-world setting is unclear. First, the participants are inserted into a peer group with no real connection or bond. These essentially anonymous and ahistorical relationships may accurately characterize commercial interactions, but differ in character from social networks classically conceived. Second, subjects are presented with an artificial task with limited or no outside information on the context (for example, estimating the length of a line). Thus, participants may have little stake in the proceedings and may not take the exercise seriously (that is, subjects want to avoid arguments on trivial matters or think they are playing a joke on the experimenter). Asch-style experiments measure a

tendency to conform, but it is unclear how and under which conditions the results translate to real-world political settings.

Randomly Constructed Network

Creative strategies have been designed to respond to these criticisms about external validity. A recent tactic embraces the isolation of the laboratory and utilizes abstract coordination games with financial incentives for subjects linked to the outcome of the game. The advent of sophisticated computer programs to aid economic games played in the laboratory has facilitated a number of experiments that directly manipulate the social network and the subject's place in it (for example, Kearns, Suri, and Montfort 2006). Researchers can now isolate the factors of theoretical interest within social networks. For instance, researchers can manipulate the degree of interconnectedness, information location, preference symmetry, and external monitoring.

The downside of this strategy is that the networks are not only artificial but entirely abstracted and may not approximate the operation of actual networks. Strategy convergence among players may reflect the ability of students to learn a game rather than measure how social networks operate. Having said that, such experiments serve as a useful "proof of concept" for formal theories of social networks. If people are in networks like X, then people will behave like Y. The challenge is to link real-world phenomena to particular games.

Role Playing

To create more realistic social networks, researchers can have subjects engage in collaborative group tasks to create camaraderie, share information and views about a range of subjects to simulate familiarity, and anticipate future encounters by scheduling post-intervention face-to-face discussion (for example, Visser and Mirabile 2004). These efforts to jump start

genuine social connections or mimic attributes of long-standing relationships are partial fixes. If organic social networks generated over years behave differently than those constructed in the laboratory, it is unclear how relationship building exercises blunt the criticism.

To address some concerns about external validity, some laboratory experiments allow subjects great freedom of action. By randomly assigning subjects roles to be played in scenes, researchers hope to gain insight into real-world relationships. The most famous example of this strategy was *Zimbardo's 1971 Stanford Prison Experiment* (Haney, Banks and Zimbardo 1973) where students were asked to act out the roles of prisoners and guards.^{iv} Most role-playing experiments are not so extreme but the same criticisms often apply. If subjects consciously view themselves as acting, the degree to which the role-play reflects actual behavior is an open question. Behaviors may differ substantially when subjects view participation as a lark and divorced from reality. A common critique of laboratory experiments is that they draw on undergraduates for their subject pool, but the critique has added bite in this setting.^v Whether more mature individuals would behave similarly given the roles assigned is an open question. Many role-playing experiments incorporate features of real-world relationships in order to approximate reality. However, not all details can be incorporated and researchers must make decisions about what features to highlight. The downside of this drive for verisimilitude is that the highlighted attributes (for example, "parents" providing "allowance") may shape the behaviors of subjects, who take cues and conform to expected behaviors. These framing decisions therefore affect experiments and potentially make the results less replicable. Thus, the degree to which social network experiments involving role-playing approximate organic social networks found in the real world is open to question.

Small Groups

Experiments where subjects deliberate in small groups are an important subset of role-playing experiments. For example, the Deliberative Polls conducted by Fishkin (Luskin, Fishkin, and Jowell 2002) invite randomly selected members of a community to discuss a topic for a day. Participants are typically provided with briefing materials and presentations by experts. The experimental component of the exercise is that subjects are randomly placed into small groups to discuss the topic at hand. Thus, subjects could be placed in a group that is ideologically like-minded, hostile, or polarized. By measuring attitudes before and after the small-group deliberation, it is possible to estimate the shift in opinion caused by discussion with liberal, conservative, or moderate citizens. The random assignment to small-group discussion ensures that a subject's exposure to the opposing or supporting viewpoints is not correlated with any characteristics of the individual.^{vi} In this way, researchers can infer how the viewpoint of discussion partners affects an individual.

The evidence of attitudinal contagion from these experiments is mixed (Farrar et al. 2009), but the model is useful to consider. Since these experiments consist of randomly selected citizens talking to other randomly selected citizens, many concerns about external validity are alleviated. The subjects are representative of the community (conditional on cooperation) and the conversation is unscripted and natural (depending on the moderator's instructions). On the other hand, the setting itself does not occur naturally. People discuss political matters with members of their social networks, not randomly selected individuals – much less a set of people who have read common briefing materials on a topic. In fairness, the hypothetical nature of the conversation is precisely Fishkin's goal, because he wants to know the decisions people would

make were they to become informed and deliberate with one another. However, the hypothetical nature of the conversation limits the degree to which the lessons learned from small-group activities can be applied to naturally occurring small groups.

3. Field Experiments

Observational studies examine naturally occurring social networks, but may suffer from selection processes and omitted variable biases. Laboratory experiments of networks possess internal validity, but the social networks studied are typically artificial and possibly too abstract to know how the results apply to real-world settings. Intuitively, conducting experiments in the field could capture the strengths of both research strategies. The reality is more complicated, given the difficulty of conducting experiments in the field, the lack of researcher control, and unique concerns about the external validity of field experiments themselves.

Three strategies can be applied to study social networks experimentally. Researchers can provide an external shock and trace the ripple through the network, control the flow of communication within a network, or randomize the network itself. While the three categories cover most field experiments, the categorization does not apply to lab settings where researchers often manipulate all three analytic levers simultaneously. For instance, in the Asch experiments, subjects are randomly assigned to a network with no confederates, eight confederates providing the wrong answer, or a group with a minority of confederates providing the correct answer. The presence or absence of confederates and their role defines the social network and manipulates the communication within the network. The task of judging the line length is the external shock used to measure the power of social influence. In theory, experiments conducted in the field could also pursue multiple randomization strategies since the categories are not mutually exclusive. In

practice, a researcher will have difficulty manipulating even one aspect of the social network. Organic social networks are difficult to map and manipulate, so researchers have far fewer analytical levers to manipulate compared to the laboratory.

Logistical and Ethical Concerns

Before discussing each of the experimental strategies, it is worth considering a few of the practical hurdles that apply to all three of the research designs. The first difficulty is in measuring the network itself. The researcher has to know where to look for influence in order to measure it and the strategy employed will inherently depend on the setting. For instance, snowball surveys are a good technique for collecting data on social networks in residential neighborhoods or for mapping friendships. Facebook and other social networking sites can be used on college campuses. Cosponsored bills in state legislatures are another possibility. Many studies of interpersonal influence rely on geography as a proxy for social connectedness, assuming that geographically proximate individuals are more likely to interact with one another than with geographically distant people (Festinger, Shachter, and Bach 1950). Each of these measurement techniques defines the network along a single dimension and will miss relationships defined along alternative dimensions. Thus, every study of social networks conducted in the field will be limited to the particular set of ties explicitly measured.

It is important to note that the measurement of the social network cannot be related to the application of the treatment in any way. Both treatment and control groups need to have networks measured in identical manners. In most instances, this is accomplished by measuring social networks first and then randomly assigning nodes to treatment and control conditions. This strategy also has the benefit of preserving statistical power by allowing for prematching

networks to minimize unexplained variance and assuring balance on covariates (Rosenbaum 2005). Given the small size of many networks studied, statistical power is not an unimportant consideration.

While defining the network identically for treatment and control variables may appear obvious, it imposes considerable logistical hurdles. Letting networks be revealed through the course of the treatment imposes a series of unverifiable assumptions and confuses the object of estimation. For instance, in his classic Six Degrees of Separation experiment, Milgram (1967) mailed letters to randomly selected individuals and requested that they attempt to mail letters to a particular individual in a separate part of the country. If the subject did not know the individual (and they would not), they were instructed to forward the letter to a person who would be more likely to know the target. Milgram then counted the number of times letters were passed along before reaching the target destination.

Revealed networks research designs, such as Milgram's, create data where the networks measured may not be representative of the networks of interest. If network characteristics (for example, social distance) correlate with the likelihood of subject treatment regime compliance (that is, forward/return the letters), then inferences drawn about the nature of the network will be biased (that is, Milgram probably overestimated societal connectedness). The treatment could also be correlated with the measurement of the network. Treatments may make certain relationships more salient relative to other relationships, so the networks measured in the treatment group are not comparable to networks assigned to control or placebo conditions. The potential bias introduced by these concerns suggests that researchers should measure the networks to be studied prior to randomization and application of the treatment. The downside of

defining the network in advance is that the analysis will be limited only to the networks the researcher measured ahead of time; less obvious connections and dynamic relationships will be omitted from the analysis. However, the avoidance of unnecessary assumptions and the clarity of analysis that results from clearly defining the network upfront more than compensate for this drawback.

The second major problem facing field experimental studies of networks is the inherent unpredictability of people in the real world where behavior cannot be constrained. This lack of researcher control poses two primary problems for experiments. First, if the behavior of a volunteer network node is part of the experimental treatment (for example, initiating conversations), then planned protocols may be violated. The violation is not necessarily because of noncompliance on the part of the subjects whose outcomes are to be measured (for example, refusing to speak with the experimental volunteer about the assigned topic), but because the person designated to provide the treatment does not dutifully execute the protocol in the way that laboratory assistants typically do. Overzealous volunteers may speak to more people than assigned; undermotivated volunteers may decide to exclude hard to reach members of their network; or, the treatment may deviate substantially from what researchers intend. To contain these problematic participants and prevent biasing the overall experiment, researchers can build safeguards into the initial experimental design (Nickerson 2005). For instance, blocking on the network nodes that provide the treatment can allow the researcher to excise problematic participants without making arbitrary decisions as to what parts of the network to remove.

A second problem that unpredictable behavior creates is that network experiments may be far more contingent and have less external validity. Suppose two people are observed to have

a strong relationship when the network is initially defined. If these people do not interact much during the course of the experiment itself, then the two individuals are unlikely to pass the treatment along to each other and the detected strength of the network will be weak. If the waxing and waning of interactions are random, such differences will balance out across pairs of individuals and the researcher will achieve an unbiased estimate of the average network characteristic to be measured. However, the waxing and waning could be a function of a range of systematic factors. For instance, experiments conducted on student networks are likely to find dramatically different results should the treatments be conducted at the beginning, middle, or end of a semester. Political interest varies during and across elections, so experiments on voting and social networks may be highly contingent. Thus, external validity is a large concern and replication is an especially important aspect of advancing the science of real-world networks.

The final practical hurdle facing researchers conducting experiments on social networks is that special attention must be paid to how the measurement of outcomes can affect the network itself. A researcher may want to see how inserting a piece of information into a network alters beliefs, but the insertion may also spur discussion in its own right. That is, the experiment could provide an unbiased estimate of how the *insertion* of the information affects the network, but cannot say how the *existence* of the information within network alters beliefs. Early social network experimenters were aware of this fact and therefore conducted their research under the label propaganda.^{vii} An extreme example of this dynamic is Dodd's Gold Shield Coffee study (1952) where randomly selected residents of a community were told the complete Gold Shield coffee slogan. The next day, a plane dropped 30,000 leaflets on the town of 300 households. The leaflets said that representatives from the Gold Shield coffee company would give a free pound

of coffee to anyone who could complete the slogan and that 1 in 5 households were already told the slogan. The next day, researchers interviewed everyone in the community to map the spread of the information. The Gold Shield Coffee experiment does not capture how company slogans diffuse through neighborhoods, but it does measure how information diffuses when a plane drops a huge number of leaflets over a very small town.

A more common problem is the measurement of baseline attitudes. Researchers often worry about testing effects among subjects in pre- and post-test designs, but it is possible that administering the pre-test changes the nature of the network. Subjects taking the survey may be more likely to discuss the topics covered in the survey than they would in the absence of the pre-test. Even if no discussion is spurred by the pre-test, subjects may be primed to be especially attentive to treatments related to the topics covered in the pre-test. This increased sensitivity may compromise the external validity of such experiments. Incorporating time-lags between pre-treatment measurements and the application of treatment can alleviate these concerns, as can creating pre-test measures that cover a wide range of topics.

Related to the practical problems in conducting experiments on social networks are the ethical problems. Setting aside obviously unethical practices (for example, forced resettlement), many practices common in political science research are problematic in the context of social networks. The revelation of attitudes about hot button issues (for example, abortion or Presidential approval), the existence of sensitive topics (for example, sexually transmitted diseases, financial distress, or abortion), and holding socially undesirable views (for example, racism, sexism, or homophobia) could fracture friendships and negatively impact communities and businesses. Selectively revealing information to subjects about neighbors can answer many

interesting questions about social networks (for example, Gerber, Green, and Larimer 2008), but should only be practiced using publicly available information or after achieving the explicit consent from subjects. Maintaining strict confidentiality standards is much more important when studying social networks than in atomistic survey conditions. Even revealing the presence or absence of network connections during a snowball survey could affect relationships, so researchers need to think carefully about the presentation of the study and how the assistants administering the survey can assure absolute privacy.

With these hurdles in mind, the three types of field network experiments can now be discussed.

External Shocks to the Network

The first experimental strategy for studying networks is for researchers to provide an external shock to an existing network and track the ripple (for example, Miguel and Kremer 2004; Nickerson 2008). The process involves introducing a change in a behavior or attitude at one node of the network and then examining other points on the network for the change as well. In principle, this strategy is not experimental per se and is like throwing a rock in a lake and measuring the waves. By throwing a large number of rocks into a large number of lakes, good inferences are possible. Randomly sampling nodes in the network only helps generalizability, much like random sampling does not make surveys experimental. To make the strategy truly experimental, multiple networks need to be examined simultaneously and the treatment then be randomly assigned to different networks. This random assignment allows the researcher to account for outside events operating on the networks (for example, the news cycle) and processes working within the network (for example, life-cycle processes). However, remember

that the unit of randomization is the network itself and not the individuals within the networks. Thus, the analysis should either be conducted at the network-level or appropriately account for the clustered nature of the treatment.

Nickerson (2008) provides an example of the strategy by looking for contagion in voter turnout. Households containing two registered voters were randomly assigned to one of three treatments. The first treatment involved face-to-face encouragement to vote in the upcoming election. The second treatment was face-to-face encouragement to recycle that served as a placebo. The final condition was a control group that received no visit from researchers, but could verify that the voter mobilization detected by the experiment was genuine. The placebo condition served to define the network. Voter turnout for the people answering the door in the voting condition would be compared to turnout among people answering the door in the recycling group. Similarly, turnout for the registered voter not answering the door could then be compared across the voting and recycling conditions. The degree to which the canvassing spilled over could then be estimated by comparing the indirect treatment effect (that is, cohabitants of people who opened the door) to the direct treatment effect (that is, for the people who opened the door). The design requires the assumption that subjects do not preferentially open the door for one of the treatments and that only the people opening the door are exposed to the treatment.

A more common strategy was employed by Miguel and Kremer (2004) in their study of a de-worming program in Kenya by using an institution (schools) as the network node to be treated. The order in which rural schools received a de-worming treatment was randomly determined.^{viii} Miguel and Kremer then compared health outcomes, school participation, and school performance for pupils at the treatment and control schools, finding cost-effective gains in

both school attendance and health. The most interesting effect, however, came when the researchers looked beyond the pupils in the experimental schools to villages and schools not included in the study. Untreated villages near treated schools also enjoyed health benefits and increased school attendance, confirming that worms are a social disease. This strategy of relying on institutional nodes of networks can be applied in a wide number of settings. The major hurdle to employing the strategy is collecting a sufficiently large number of networks or institutions to achieve precise and statistically meaningful results.

Controlling the Flow of Communication within a Network

A second strategy is to control the flow of communication within a neighborhood or network. The idea is to recruit participants who apply a treatment to randomly selected members of their social networks. Nickerson (2007) provides an example where volunteers were recruited to encourage friends and neighbors to vote in congressional elections. Volunteers listed people who may need encouragement and who volunteers would be comfortable talking to. The people listed were then randomly assigned to be approached (treatment) or not (control). The same design principle has been applied to proprietary studies of campaign donations and the adoption of consumer products. A major advantage of the design is that the list provided by the volunteers clearly defines the social network to be examined. Since individuals within networks are the unit of randomization, the design can also be much more powerful than designs that randomize across networks.

The biggest problem with controlling the flow of communication within a network is that the experimental interaction may be artificial and not approximate conversations that occur organically. Neighbors, friends, and coworkers rarely make explicitly political appeals to each

other. Most of the hypothesized mechanisms for the diffusion of norms and peer effects are subtle and take time. It is possible that friends have a great deal of influence over each other but recoil from explicit prodding. Thus, such experiments measure the effect of aggressive word-of-mouth campaigns within networks and not the workings of social networks in their natural state.

Inadvertent contamination is a serious problem within social networks that needs to be considered. Volunteers may bring up the experiment in the course of everyday conversation, perhaps following such innocent questions as “What’s new?” Subjects may cross-contaminate themselves by discussing the unusual behavior of the volunteer applying the treatment. These problems can be avoided by randomizing across networks (that is, some volunteers treat everyone and others treat no one), but this comes at the cost of considerable statistical power. The difficulty in controlling communication in social networks makes this type of experiment very difficult to conduct in the field and probably better suited for the laboratory.

Randomizing the Network Itself

The final strategy randomizes the position of people within networks. The steps involve measuring people’s opinions, attributes, or tendencies at Time 0, assigning a place in a network at Time 1, measuring opinions at Time 2, and then modeling Time 2 opinions for one person as a function of opinions at Time 0 of both the subject and the others in the network. Obviously, there are limited settings where subjects can be randomly assigned to places in social networks. The most common use of this strategy has been to examine the effect of roommates among college freshmen looking at outcomes such as grades (Sacerdote 2001) or drug use and sexual behavior (Boisjoly et al. 2006). Less common are experiments where inmates are randomly assigned security levels in prisons (Bench and Allen 2003; Gaes and Camp 2009), which generally find

that prisoners assigned to more secure prisons are no more or less likely to commit crimes within prison but are more likely to commit crimes upon release. The key to these empirical strategies is establishing baseline characteristics prior to assignment to achieve identification. Once the assignment is made and peers are residing together, outside forces could cause conformity independent of any peer effects, thereby creating spurious relationships.

The biggest problem with this research strategy is that researchers rarely have the power to randomly assign the residence of subjects.^{ix} The cases where random assignment is practiced may not generalize to more typical living conditions. The types and intensities of interactions a person has in dormitories or cell blocks may be qualitatively different than interactions that people typically have at work or in their neighborhood. College students and prisoners are often young and may also be more impressionable than older individuals. As a result, these types of studies can tell us a great deal about the dynamics of these particular networks, but how the results apply to other settings is an open question.

A step below randomizing the network itself is randomly providing the opportunity to opt in or out of a network (for example, change neighborhoods or schools). The most famous of these experiments is the Moving to Opportunity (MTO) study (Katz, Kling, and Liebman 2001; Kling, Ludwig, and Katz 2005; Kling, Liebman, and Katz 2007) where randomly selected residents of public housing were provided vouchers to move where they see fit. The experiments then compared the outcomes of families receiving the vouchers to those in the control group with no voucher. The MTO experiments found that subjects electing to move felt safer and healthier, but made few differences with regards to criminal activity, employment, and educational attainment. The same type of experiment has been conducted with regards to schooling, where

randomly selected families are provided vouchers to attend schools of their choosing (Howell et al. 2002).

All experiments provide a complier average treatment effect to some extent, but the dilemma is highlighted in these choice experiments. Many policy analysts would like to know the effect of living in certain types of neighborhoods or attending particular schools on the average person. However, choice experiments can only speak to how the *move* out of one environment and into another affects the *type of person* who would move. Both the treatment and control group also contain people who would stay put given the opportunity and the experiment is uninformative about these subjects. These nonmovers are revealed in the treatment group, but not the control group where randomization only assures that the proportion of nonmovers is the same as the treatment group. Thus, carefully defining the estimand and designing treatment-on-the-treated analysis play a very special role in choice experiments. How best to model the decision-making process is not always obvious and researchers have more discretion than is typically found in the analysis of experiments.

Subject attrition is a special challenge for choice experiments in two ways. First, subjects who take advantage of the voucher program may opt to move out of the area where researchers can easily track behavior. If outcomes for subjects moving out of the area differ from outcomes achieved locally, then the estimated treatment effect will be necessarily biased because movement is inherently correlated with the treatment. Second, subjects not enrolled in the experimental program (that is, the control group) have little reason to comply with researcher's requests for information and may be more likely to drop out of the study. This process could result in a control group that is no longer comparable to the treatment group. Both of these

problems can be solved with sufficient resources to acquire information and incentivize participation, but researchers seeking to conduct choice experiments should take steps to address these two forms of attrition.

4. Conclusion

Social networks have been studied throughout the history of social science, but new analytic tools are providing fresh insights into how people are tied together. Unsurprisingly, no single approach can lay claim to being preferred and all methods have their drawbacks. Observational studies allow researchers to collect large amounts of data and study the real world relationships of interest, but may be plagued by spurious correlations that are impossible to eradicate. Laboratory experiments suffer from no omitted variable bias and can randomly manipulate the theoretically interesting aspects of social networks. The results in the laboratory will generally be theoretically abstract and anonymous networks. The types of real-world networks to which the results apply is an open empirical question that researchers will need to establish. In theory, field experiments should combine the strengths of both the observational and laboratory strategies, but the reality is far messier. The cases where field experiments can be applied to social networks are necessarily limited, so the external validity of the findings is open to question. The amount of control researchers have over the network is limited and many theoretically and practically interesting questions will prove impossible to study.

Thus, a combination of the three approaches is likely to prove the most fruitful. As data become more ubiquitous and available to researchers, observational studies will be able to address an increasing range of issues. Just as lab experiments have helped to guide the theoretical development of game theory, laboratory experiments on social networks will answer increasingly

complicated theoretical questions about network density, information flow, strength of ties, and a host of other factors. As randomized trials become more accepted in a range of policy settings (for example, education, housing, legal enforcement, or environmental protection), the number of opportunities to conduct field experiments on social networks will also increase. Little experimental work on networks has been done to date, but that leaves many fertile avenues for researchers.

References

- Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley, CA: University of California Press.
- Asch, Solomon. E. 1956. "Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority." *Psychological Monographs* 70 (Whole no. 416).
- Bench, Lawrence L., and Terry D. Allen. 2003. "Investigating the Stigma of Prison Classification: An Experimental Design." *The Prison Journal* 83: 367-82.
- Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.
- Boisjoly, Johanne, Greg J. Duncan, Michael Kremer, Dan M. Levy, and Jacque Eccles. 2006. "Empathy or Antipathy? The Impact of Diversity." *American Economic Review* 96: 1890-905.
- Bond, Charles F., and Linda J. Titus. 1983. "Social Facilitation: A Meta-Analysis of 241 Studies." *Psychological Bulletin* 94: 265-92.
- Bond, Rod, and Peter B. Smith. 1996. "Culture and Conformity: A Meta-Analysis of Studies Using Asch's Line Judgment Task." *Psychological Bulletin* 119: 111-37.
- Caldeira, Gregory A., and Samuel C. Patterson. 1987. "Political Friendship in the Legislature." *Journal of Politics* 4: 953-75.
- Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1964. *The American Voter*. Chicago: University of Chicago Press.
- Carrington, Peter J., John Scott, and Stanley Wasserman. 2005. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

- Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network Over 32 Years Background." *New England Journal of Medicine* 357: 370-79.
- Christakis, Nicholas A., and James H. Fowler. 2008a. "The Collective Dynamics of Smoking in a Large Social Network Background." *New England Journal of Medicine* 358: 2249-58.
- Cohen-Cole, Ethan, and Jason M. Fletcher. 2008. "Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analysis." *British Medical Journal* 337: a2533.
- Dodd, Stuart C. 1952. "Testing Message Diffusion from Person to Person." *Public Opinion Quarterly* 16: 247-62.
- Farrar, Cynthia, Donald P. Green, Jennifer E. Green, David W. Nickerson, and Stephen D. Shewfelt. 2009. "Does Discussion Group Composition Affect Policy Preferences? Results From Three Randomized Experiments." *Political Psychology* 30: 615-47.
- Festinger, L., Schachter, S., and Bach, K. 1950. *Social Pressures in Informal Groups: A Study of Human Factors in Housing*. New York: Harper & Row.
- James H. Fowler and Nicholas A. Christakis. 2008. "Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study." *British Medical Journal* 337: a2338.
- Gaes, Gerald G., and Scott D. Camp. 2009. "Unintended Consequences: Experimental Evidence for the Criminogenic Effect of Prison Security Level Placement on Post-Release Recidivism." *Journal of Experimental Criminology* 5: 139-62.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.
- Haney, Craig, Curtis Banks, and Philip Zimbardo. 1973. "Study of Prisoners and Guards in a Simulated Prison." *Naval Research Reviews* 9: 1-17. Washington, DC: Office of Naval Research.
- Howell, William G., Paul E. Peterson, with Patrick J. Wolf, and David E. Campbell. 2002. *The Education Gap: Vouchers and Urban Schools*. Washington, DC: Brookings Institution Press.
- Huckfeldt, Robert. 1983. "The Social Context of Political Change: Durability, Volatility, and Social Influence." *American Political Science Review* 77: 929-44.
- Huckfeldt, Robert, and John Sprague. 1995. *Citizens, Politics, and Social Communication: Information and Influence in an Election Campaign*. New York: Cambridge University Press.

- Jennings, M. Kent, and Richard G. Niemi. 1974. *The Political Character of Adolescence: The Influence of Families and Schools*. Princeton, NJ: Princeton University Press.
- Jennings, M. Kent, and Richard G. Niemi. 1981. *Generations and Politics: A Panel Study of Young Adults and Their Parents*. Princeton, NJ: Princeton University Press.
- Karau, S. J., and Williams, K. D. 1993. "Social Loafing: A Meta-Analytic Review and Theoretical Integration." *Journal of Personality and Social Psychology* 65: 681-706.
- Katz Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Housing Mobility Study." *Quarterly Journal of Economics* 116: 607-54.
- Kawachi, I, G. A. Colditz, A. Ascherio, E. B. Rimm, E. Giovannucci, M. J. Stampfer, and W. C. Willett. 1996. "A Prospective Study of Social Networks in Relation to Total Mortality and Cardiovascular Disease in Men in the USA." *Journal of Epidemiology and Community Health* 1996: 245-51.
- Kearns, M., S. Suri, and N. Montfort. 2006. "An Experimental Study of the Coloring Problem on Human Subject Networks." *Science* 313: 824-27.
- Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75: 83-119.
- Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz. 2005. "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment." *Quarterly Journal of Economics* 120: 87-130.
- Kravitz, David A., and Barbara Martin. 1986. "Ringelmann Rediscovered: The Original Article." *Journal of Personality and Social Psychology*, 50: 936-41.
- Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice*. New York: Columbia University Press.
- Lazarsfeld, Paul F., and Robert K. Merton. 1954. "Friendship as a Social Process: A Substantive and Methodological Analysis." In *Social Control, the Group, and the Individual*, eds. Morroe Berger, Theodore Abel, and Charles H. Page. New York: D. Van Nostrand Company, Inc.
- Luskin, Robert C., James S. Fishkin, and Roger Jowell. 2002. "Considered Opinions: Deliberative Polling in Britain." *British Journal of Political Science* 32: 455-87.
- Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72: 159-217.
- Milgram, Stanley. 1967. "The Small-World Problem." *Psychology Today* 1: 61-7.

- Mutz, Diana C. 1998. *Impersonal Influence*. New York: Cambridge University Press.
- Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233-52.
- Nickerson, David W. 2007. "Don't Talk to Strangers: Experimental Evidence of the Need for Targeting. Presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49-57.
- Nordenmark, Mikael. 1999. "The Concentration of Unemployment within Families and Social Networks: A Question of Attitudes or Structural Factors?" *European Sociological Review* 15: 49-59.
- Rosenbaum, Paul R. 2005. "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies." *The American Statistician* 59:147-52.
- Sacerdote, Bruce I. 2001. "Peer Effects With Random Assignment." *Quarterly Journal of Economics* 116: 681-704.
- Sanders, Jimmy, Victor Nee, and Scott Sernau. 2002. "Asian Immigrants' Reliance on Social Ties in a Multiethnic Labor Market." *Social Forces* 81: 281-314.
- Sunstein, Cass R. Sunstein, Reid Hastie, John W. Payne, David A. Schkade, and W. Kip Viscusi. 2003. *Punitive Damages: How Juries Decide*. Chicago: Chicago University Press.
- Tucker, Peter. 1999. "Normative Influences in Household Waste Recycling." *Journal of Environmental Planning and Management* 42: 63-82.
- Visser, Penny S., and Robert R. Mirabile. 2004. "Attitudes in the Social Context: The Impact of Social Network Composition on Individual-Level Attitude Strength." *Journal of Personality and Social Psychology* 87: 779-95.
- Zuckerman, Alan S., Josip Dasovic, and Jennifer Fitzgerald. 2007. *Partisan Families: The Social Logic of Bounded Partisanship in Germany and Britain*. New York: Cambridge University Press.

ⁱ Selection bias can be even more pernicious than omitted variable bias (see Achen 1986).

ⁱⁱ A similar problem arises when happiness is found to be more contagious for neighbors than coworkers or spouses.

ⁱⁱⁱ Social loafing dates back to at least 1913 when Ringelmann found individuals pulled harder on a rope when working alone than when working in concert with others (Kravitz and Martin 1986).

^{iv} The experiment was halted after six days because of physical and psychological abuse by guards.

^v For a somewhat contradictory view on the external validity of student samples, see Druckman and Kam's chapter in this volume.

^{vi} Many mock jury experiments (for example, Sunstein et al. 2003) share this characteristic.

^{vii} The fact that the military funded much of this research in order to understand the effectiveness of propaganda techniques assisted this decision.

^{viii} This strategy also avoids ethical concerns about denying subjects treatment.

^{ix} People randomly assigned to living quarters typically have limited autonomy (for example, prisoners and soldiers) and, therefore, enjoy additional human subjects protections.

VI. Identity, Ethnicity, and Politics

20. Candidate Gender and Experimental Political Science

Kathleen Dolan and Kira Sanbonmatsu

The largest literature on gender and experimentation in political science concerns voter reaction to candidate gender. One of the earliest and most enduring questions in the study of gender and politics concerns women's election to office. As the number of women candidates and officeholders has increased in the United States over the past several decades, there are more cases of women candidates and officeholders available for empirical analysis. Today, women are a majority of the electorate and women candidates tend to win their races at rates similar to those of men. Yet, the gender gap in candidacy and officeholding remains large and stable. Understanding how voter beliefs about candidate gender shape attitudes and political behavior remains an important area for research.

Experimentation has helped scholars overcome some of the limitations of using observational studies to investigate candidate gender. As Sapiro (1981-2) observed, public opinion surveys may not be able to detect prejudice against women candidates if voters provide socially desirable responses. And if prejudice against women is subconscious, voters may not even be aware of their attitudes. Observational studies are also limited in helping us understand what we cannot observe, namely why far fewer women than men seek office. If women fail to run because they fear a gendered backlash from voters, we are unable to evaluate the experiences of those women.

The women who do run may be “a unique ‘survivor’ group” of candidates because of the recruitment processes that women have had to overcome in order to become candidates (Sapiro

1981-2, 63). Attitudes toward the women who run may not reflect the public's response to "average" women candidates, not unlike the sample selection problem that James Heckman identified with regard to the study of women's wages. Indeed, women candidates are not randomly distributed across districts, states, or types of elective offices because the gender-related attitudes of voters and gatekeepers shape the geographic pattern of where women emerge as candidates and are successful. In addition, when women run for office, they may anticipate voter hostility to their gender and may subsequently work to counteract any negative, gender-based effects in their campaigns. Thus, it may be difficult to observe the effects of gender in electoral politics due to selection effects and strategic decision making by women candidates. For the same reasons, experimentation can be particularly helpful in the study of race/ethnicity (see Chong and Junn's chapter in this volume).

Isolating the effect of candidate gender in observational studies is also difficult precisely because gender contains so much information. Voters may use candidate gender to infer a candidate's personality traits, issue positions, party affiliation, ideology, issue competence, occupation, family role, and qualifications. Too, the impact of candidate gender may interact with other forces in the political environment – candidate political party, ideology, incumbency, prior experience – to create a situation in which candidate gender does not influence voter attitudes and behaviors in the same way for every woman candidate. Indeed, it is precisely the complexity of the typical election environment that makes it difficult for observational studies to accurately capture the effect of gender.

In this chapter, we describe several important studies that capitalize on the benefits of experimentation to expand our understanding of whether voters hold gender stereotypes and

whether voters are biased against women candidates. Experimentation has also been used to understand how candidate gender interacts with factors such as party identification and type of elective office. At the end of this essay, we suggest ways that scholars could use experimental designs to answer remaining and new questions about gender and politics in the future.

1. Gender Bias and Gender Stereotypes

Sapiro (1981-2) conducted one of the pioneering experiments in gender and politics. In a simple design that followed the Goldberg (1968) experiment, Sapiro asked undergraduate students in introductory political science classes to read a speech by a fictional candidate for the U.S. House of Representatives. The purpose of the study – to understand whether subjects reacted differently to the speech based on the gender of the speaker – was not revealed. In the first condition, students were informed that the candidate was “John Leeds” and, in the second condition, students were informed that the candidate was “Joan Leeds.” Because Sapiro wanted to know how voters reacted to candidates in a low-information context, the stimulus was an actual speech given by a U.S. Senator that was ambiguous with respect to political party and most policy issues.

With this design, Sapiro was able to isolate the effect of candidate gender on multiple voter inferences. Prior to the Sapiro study, scholars usually relied on election results and public opinion surveys to gauge voter attitudes toward women. For example, Darcy and Schramm’s (1977) analysis of congressional election results found that female candidates were not at a disadvantage compared to male candidates, once the type of race was taken into account. They concluded that voters were indifferent to candidate gender. However, studies based on aggregate election results do not take into account the selection effects that produce successful women

candidates and cannot explain the low numbers of women candidates. For example, a woman ran in only eight percent of the early 1970s general election races that Darcy and Schramm studied. Meanwhile, public opinion surveys in the early 1970s continued to reveal bias against a hypothetical woman candidate for president, although attitudes were becoming more liberal (Ferree 1974).

Sapiro found no difference in subjects' willingness to support the female candidate. Nor was there a difference in how respondents evaluated the candidate's understanding of policy issues, the clarity of the speech itself, the expected effects of the candidate's proposals, or whether the subject agreed with the policy positions included in the speech. Though Sapiro did not find evidence of prejudice, she found a difference in perceptions of the likelihood the candidate would win the race. A majority of respondents thought that the male candidate would win compared to less than one-half of respondents in the female-candidate condition. Such doubts about women's electability can put women candidates at a disadvantage because voters, donors, interest groups, and political parties may wonder if women candidates are worthy of investment.

Sapiro found that candidate gender affected subjects' evaluations of issue competence on issues that were not specifically mentioned in the stimulus. The female candidate was rated as more likely to be competent on three areas typically associated with women (education, health, and honesty/integrity in government) and less competent on two areas associated with men (military and farm), with no difference in other areas (environment and crime). Thus, issue areas typically associated with women in society potentially provide women with an advantage in the political realm.

One drawback to the Sapiro study is that subjects were only asked to evaluate one candidate and were not provided with the candidate's party affiliation, making the experiment unlike a real-world election (see McGraw's chapter in this volume). Because the study was conducted with a sample of undergraduates, the results might not hold in a general population. On the other hand, as with other gender experiments, student samples make for more stringent tests of the gender bias hypothesis: it should be more difficult to observe gender effects among young voters because age is one of the most consistent predictors of bias against women.

Subsequent studies—both observational and experimental—have confirmed the existence of gender stereotypes as well as the absence of explicit voter opposition to women candidates (e.g., Welch and Sigelman 1982; Rosenwasser and Seale 1988; Alexander and Andersen 1993). For example, Leeper (1991) used a simple experimental design that varied candidate gender and asked undergraduate student subjects to evaluate a single candidate, as in the Sapiro study. Unlike Sapiro, however, Leeper sought to investigate voter reaction to a “masculine” woman candidate by creating a stimulus that emphasized masculine themes and a “tough on crime” message. Because he found no effect of candidate gender on voter evaluations of stereotypically masculine issues, he concluded that the masculine nature of the speech helped the female candidate overcome a traditional disadvantage. Meanwhile, he concluded that “voters may infer that tough, aggressive women still possess latent (stereotypical) warmth” (1991: 254). Voters rated the female candidate as more competent on female issues, such as education and maintaining honesty and integrity in government. The practical advice he offered to female candidates, therefore, was to pursue a “masculine” image without concern for presenting a

“female” side because voters will infer the feminine qualities. Consistent with Sapiro, Leeper found that subjects thought that the female candidate would be less likely to win.

Because of the strong evidence that voters hold stereotypes about issue competence, Huddy and Terkildsen (1993b) set out to determine the source of these political gender stereotypes. Until this point, there was relatively little attention given to the source of the public’s gender stereotypes and little nonexperimental work tried to identify these sources. As the authors suggest, understanding the source of gender stereotypes can help to explain whether stereotypes are widely held and whether they can be overcome. To test their framework, they conducted an experiment with 297 undergraduate students in the Fall of 1990 in which subjects were asked to evaluate a single candidate. They manipulated three between-subjects factors: the sex of the candidate, whether the candidate was running for national or local office, and whether the candidate was described as possessing typically feminine or masculine personality characteristics.¹ Subjects were also asked to judge whether the candidate could be described as having a series of additional traits beyond those mentioned in the description. This allowed Huddy and Terkildsen to determine whether and to what degree people made gender-related inferences about the candidates. Finally, subjects were asked to indicate how well they thought the candidate could handle military, economic, compassion, and women’s issues and how they evaluated the candidate’s party identification, ideology, and position on feminism.

Huddy and Terkildsen’s analysis tested two main hypotheses. The first was that people’s stereotypes about the gender-linked personality traits of women and men (i.e., women are kind, men are aggressive) could lead people to assume gender-based competence in different areas (e.g., women are better at compassion issues, men at military issues). The second hypothesis was

that the political beliefs ascribed to women and men may be the cause: the belief that women are more liberal and Democratic than men could explain why women are perceived to be better at handling compassion issues. In the end, they found significant evidence that inferred traits were more important in determining policy competence stereotypes than were inferred beliefs about candidate partisanship and ideology. Interestingly, their manipulation of gender-linked personality traits did not eliminate the effect of manipulated candidate sex on issue competency ratings.

This study remains influential because it demonstrates the importance of masculine personality traits to evaluations of candidate competence, including competence on “women’s issues.” Their suggestion that women candidates create personas that emphasize their masculine traits has been confirmed by more recent works (Walsh and Sapiro 2003; Bystrom et al. 2004). Huddy and Terkildsen also shed light on the source of voters’ policy competence stereotypes, pointing to the role that perceived traits play in evaluations. Finally, their work moved the subfield forward at a time when nonexperimental data on stereotypes were limited. Their approach took advantage of the ability to manipulate the key variables of candidate sex and gendered personality traits while reducing social desirability concerns.

2. Type and Level of Elective Office

In another extension of our understanding of the role of gender stereotypes in evaluations of candidates, Huddy and Terkildsen (1993a) consider whether the impact of stereotypes is conditional on the context of the offices women seek, specifically the level and type of office. Women are more likely to hold lower level offices, such as school board and state legislative office, rather than higher level offices, such as statewide and congressional office. And, at least

as of yet, a woman has never been elected to the presidency or vice presidency. Work by Rosenwasser et al. (1987) found that college students rated a male candidate as more effective on “masculine” presidential tasks than a female candidate, with the female candidate perceived to be more effective on “feminine” tasks. Furthermore, Rosenwasser and Dean (1989) found that students rated all political offices as more masculine than feminine and that masculine presidential tasks were deemed more important than feminine presidential tasks.

Building on these studies, Huddy and Terkildsen rightly note that most work to that point had focused on the presidency to the exclusion of other national offices and state and local positions. Too, researchers had generally ignored whether the public saw women as better suited for certain types of elective office. Based on earlier findings on the public’s gender stereotypes, they hypothesized that voters will take level and type of office into account when evaluating women candidates. Specifically, they expected that people would see women’s personality traits and policy competency as being better suited for local than national office and nonexecutive over executive positions.

Huddy and Terkildsen (1993a) analyzed data from the aforementioned study (Huddy and Terkildsen 1993b). For this paper, however, Huddy and Terkildsen analyzed the manipulation of candidate gender, candidate gender-linked traits, and level and type of office. They began by providing subjects with a list of nine masculine and seven feminine traits and asking people to evaluate the personality traits of a “good politician” running for president, Congress, mayor, or local council member. This allowed them to determine whether the preferred package of personality traits changed with the level of office. As the authors indicated, one of the strengths of the experimental design they employed was the focus on a “good” hypothetical candidate at

different levels of office. It would be quite difficult to isolate the impact of the office itself if the study consisted of voter evaluations of actual candidates, with their myriad experiences and political identities.

Their initial analysis provided support for their hypothesis that good candidates for national and executive office were expected to hold more masculine characteristics than were candidates for legislative and local office. There were significant main effects for both level (between-subjects) and type (within-subjects) of office. The same general pattern held for people's expectations about policy competence; typical male policy issues like military issues and the economy were considered more central to higher level and executive office. Compassion issues like child care and welfare were seen as more central to legislative and local level office.

Having confirmed that people hold gendered expectations for different kinds of elective office, Huddy and Terkildsen then went on to determine whether candidates lose votes when they do not possess the "appropriate" characteristics for the office they seek. Again their findings conformed to expectation. Masculine traits were most important to candidates for national office, but offered little advantage to those who sought local office. However, they also found that typical feminine traits did not offer an advantage to candidates for local office. Male policies, such as military and policing, were very important to candidates for national office, but the feminine, compassion issues offered no boost to candidates for local office. In general, then, Huddy and Terkildsen demonstrated that people have a clear preference for masculine traits and male policy competence when judging candidates for national office, and appear to consistently devalue feminine traits and female policy competence, even when candidates seek local office. The important influence they identified is level of office, not type; their analysis found that the

distinction between executive and legislative office had no impact on the traits and policy strengths people value.

In addition, their analysis found no real gender differences among subjects in the degree to which subjects valued male traits and male policy competence when evaluating candidates for higher level offices. But male subjects devalued feminine traits as important to these offices and exhibited less willingness to say they would vote for the candidate with more feminine attributes than were female subjects.

3. Gendered Media Effects

Another important area of investigation of gender effects concerns media coverage. Kahn (1994) used an experiment to determine the effect of gender differences in news coverage on candidate impression formation. Kahn's content analysis of newspaper coverage of twenty-six U.S. Senate races and twenty-one gubernatorial races between 1984 and 1988 that featured women candidates revealed gendered patterns of news coverage. She hypothesized that media coverage patterns would vary across the two offices because of the nature of the offices; foreign policy and national security issues that animate Senate politics are more likely to advantage male candidates, whereas statewide issues such as health and education that are more likely to dominate the agendas of gubernatorial candidates are expected to advantage female candidates.

Kahn's content analysis of media coverage revealed that women received more horserace coverage than men and that women senatorial candidates received more negative viability assessments than men. Women also received less issue coverage than men. Turning to an experiment, Kahn recreated these gendered patterns of media coverage in order to measure their effects on impression formation. In all, Kahn identified 14 dimensions of coverage that she used

to form her prototype articles. In order to simulate the news coverage, she created four prototype articles (male/female incumbent, and male/female challenger) for both gubernatorial and senatorial races based on the actual coverage. In two separate studies (one for Senate, one for governor), Kahn used a two-by-four factorial design that varied the four types of coverage and candidate gender. Thus Kahn was able to examine the impact of “female” versus “male” press coverage on both male and female candidates while also taking into account office type and incumbency. Her study remains a useful model for gender scholarship because it used an observational analysis in conjunction with experimentation.

Kahn’s results indicated that gender differences in campaign coverage did shape impression formation, though the effects were strongest for coverage of Senate incumbents. Senate incumbent candidates with female coverage were less likely to be perceived as viable and less likely to be considered strong leaders than Senate incumbents receiving male coverage. Female Senate incumbent candidates were perceived to be more competent on health issues and considered to be more compassionate. The analysis of gubernatorial coverage revealed fewer gender differences than senatorial coverage and the gubernatorial experiment likewise produced fewer effects. Incumbent coverage differences in the experiment were limited to viability assessments, with the candidate in the female incumbent gubernatorial condition considered to be less electable than the candidate who received male incumbent coverage.

Kahn also found that, holding coverage constant, women were perceived as more compassionate and more honest than men, better able to maintain honesty and integrity in government, and more competent in women’s issues and the areas of education and health. Women gubernatorial candidates were perceived to be more knowledgeable than their male

counterparts. Meanwhile, no differences were found on stereotypical male issues (military, leadership, and the economy). Most of these effects were due to the differential evaluation of candidates by female respondents. Finally, Kahn found that the effects of gendered coverage and candidate gender were cumulative.

4. The Intersection of Gender and Party

Scholars of gender politics employing experimental methods have successfully demonstrated that electoral context matters to public evaluations of women candidates. Yet, King and Matland (2003), in a review of previous experimental work on public evaluations of women candidates, suggests one major limitation of experimental work on this topic: the isolation of candidate gender from other important political and social variables that might influence voter reactions to candidates. If voter evaluations are strongly shaped by a candidate's (or their own) party identification, then there may be no room for heuristics such as gender to ultimately have significant influence. Or, as they suggest, it may be the case that party cues interact with gender cues, which could result in women candidates of different political parties being evaluated differently.

King and Matland's data came from an experiment embedded in a random national telephone survey of 820 U.S. adults sponsored by the Republican Network to Elect Women (RENEW). Respondents received a description of a Republican candidate running for Congress. The candidate's gender was manipulated. After the brief description, subjects were asked to evaluate the candidate on a number of traits and state whether they would be likely to vote for this candidate. Because of the sponsor of the survey, only reactions to a Republican candidate were evaluated.

King and Matland's goal was to test the power of gender cues relative to the power of partisan cues on evaluation and willingness to vote for the candidate. They hypothesized that gender cues would be relevant to evaluations of the traits of the candidate but that the cue of party alone would predict willingness to vote for the candidate. Instead, however, the authors found that gender and party cues interacted. Voter party identification was indeed the strongest influence on willingness to vote for the candidates, but there was also a significant interaction with candidate gender. Republican subjects in the pool were more likely to say they would vote for the Republican man than the Republican woman. The opposite was true for Democratic and Independent subjects, who were each more likely to vote for the Republican candidate when she was a woman. Democratic and Independent subjects were no more likely to see the Republican as "conservative" when the candidate was presented as a man or a woman. Republican subjects, on the other hand, perceived the Republican man as more "conservative" than the Republican woman. King and Matland suggested that Republican women pay a price with their own party members for their perceived greater liberalism, but may reap a benefit from this stereotype among Democratic and Independent voters.

They found the same general pattern with evaluation of the traits of the candidates. On each measure (the candidate "shares my concerns," "can be trusted," "is a strong leader," and is "qualified,") the Republican woman candidate received significantly more positive evaluations from Democratic and Independent subjects than from Republican subjects. King and Matland conclude that Republican women candidates may well have to "make up" any votes they lose from their own party's voters with crossover votes from Democratic and Independent voters. However, this interaction of party and gender cues could hurt Republican women in primary

contests. Their findings about the primary election difficulties that the ideology stereotype poses for Republican women are consistent with the conclusions of observational studies (Lawless and Pearson 2008; Sanbonmatsu and Dolan 2009). At the same time, King and Matland acknowledged that the absence of a treatment for Democratic candidates limited their ability to determine whether the interaction of party and gender works the same way in each party.

In addition to advancing knowledge about gender stereotypes by introducing the role of party, King and Matland's experiment was conducted with a random national sample.

Matland and King (2002) criticized past experiments on gender for their almost exclusive use of college student subjects, arguing that college students have less well-developed political ideas and are less likely to participate in politics than older people.ⁱⁱ

5. Future Directions

Over the past forty years, scholars of gender politics have grappled with the myriad ways that gender influences American political life using both observational and experimental methods. However, as this chapter suggests, experimental work has been of particular value to this endeavor, often offering advantages over observational methods. This advantage can be seen in the foundational work that we have reviewed here, as well as in current research. For example, Streb et al. (2008) employed a list experiment to tackle concerns about social desirability issues that can result from directly asking people whether they would support a woman for president. Winter (2008) manipulated media frames around issues of race and gender to determine how these issues shape public opinion, which cannot easily be replicated through observational work. Philpot and Walton (2007) employed an experiment to gauge the simultaneous impact of the intersection of race and sex on support for African American women candidates. Fridkin,

Kenney, and Woodall (2009) manipulated media campaigns to determine the impact of candidate gender on voter reaction to negative advertising. Experimentation can also be used to better understand the gender gap in public opinion. For example, Lizotte (2009) used an experiment to analyze the gender gap in public opinion on the use of force.ⁱⁱⁱ We would urge gender scholars to expand their use of experimental methods because many questions are ripe for experimentation. Future research can continue to use experiments to pursue the study of intersectionality along the lines of the work by King and Matland on party identification and by Philpot and Walton on candidate race. Pinpointing the interaction of gender with features of the electoral context remains an important area for investigation.

Several new lines of inquiry emerge from the 2008 presidential election. First, the sexism and misogyny evident in some of the reaction that Hillary Clinton encountered during the 2008 campaign was unexpected given the thrust of the existing literature about the absence of bias against women candidates, as well as the expectation that gender stereotypes will be attenuated in high-information contexts (Carroll 2009; Lawless 2009; Carroll and Dittmar 2010; Lawrence and Rose 2010). Clinton's experience led Freeze, Aldrich, and Wood (2009) to conduct an experiment in order to understand the persuasiveness of messages about a candidate from a sexist source. The role that gender stereotypes played in voter and media reaction to Hillary Clinton suggests that stereotypes can play a role even in the presence of substantial information about a candidate. Much work remains to be done about how voters form impressions about candidates in high-information contexts. Clinton's bid also calls into question the longstanding finding that voters will infer feminine traits in the absence of their explicit presentation in campaigns. In order to battle stereotypes about the ability of a woman to serve as commander-in-chief, Clinton

may have portrayed an image that was too masculine; voters seem to have penalized her for her failure to appear more feminine.

Clinton's experience also raises questions about whether she would have fared better had she emphasized the historic aspect of her race and her potential to become the country's first female president. Studies that seek to understand internal campaign decision making about candidate gender are few (Fox 1997; Dittmar 2010). Yet, content analysis of the relationship between gender and political advertisements reveals that it is uncommon for women candidates to make gender an issue in their campaigns (Bystrom et al. 2004; Dittmar 2010). Experimental research could help to determine if – and when – women candidates can benefit by making their gender identity an explicit campaign issue.^{iv}

Sarah Palin's 2008 vice presidential campaign provides a different, but equally important, window into stereotypes. Her appearance on the national stage as a socially conservative Republican may alter the dominant stereotypes about political women, which are largely derived from Democratic women. Future studies can probe the extent to which Palin has reshaped voter assumptions about the behavior and traits of women in electoral politics.

Finally, future experimental work on gender could move beyond a reliance, as seen in some early works, on college student populations. Several recent works have successfully employed experiments embedded in surveys of nationally representative samples (King and Matland 2003; Streb et al. 2008; Fridkin et al. 2009). For some analyses about the effects of candidate gender, experiments conducted with a representative sample could significantly strengthen existing findings.

6. Conclusion

An impressive and growing body of gender politics research employing experimental methods points the way to future areas of exploration. The realities of the political world suggest that questions of the impact of gender on candidates, voters, issues, campaigns, and media coverage will continue unabated into the future. The increasing number of women candidates running for a range of political offices and the candidacies of Hillary Clinton and Sarah Palin signal that there is still much to learn about how, when, and why gender is an important political consideration. Increased reliance on the experimental method can expand our understanding of this critical influence.

References

- Alexander, Deborah, and Kristi Andersen. 1993. "Gender as a Factor in the Attribution of Leadership Traits." *Political Research Quarterly* 46: 527-45.
- Bystrom, Dianne, Mary Christine Banwart, Lynda Lee Kaid, and Terry Robertson. 2004. *Gender and Candidate Communication: VideoStyle, WebStyle, NewsStyle*. New York: Routledge.
- Carroll, Susan. 2009. "Reflections on Gender and Hillary Clinton's Presidential Campaign: The Good, the Bad, and the Misogynic." *Politics & Gender* 5: 1-20.
- Carroll, Susan, and Kelly Dittmar. 2010. "The 2008 Candidacies of Hillary Clinton and Sarah Palin: Cracking the 'Highest, Hardest Glass Ceiling.'" In *Gender and Elections: Shaping the Future of American Politics* 2nd Ed., eds. Susan J. Carroll, and Richard L. Fox. New York: Cambridge University Press.
- Darcy, Robert, and Susan Schramm. 1977. "When Women Run Against Men." *Public Opinion Quarterly* 41: 1-13.
- Dittmar, Kelly. 2010. "Inside the Campaign Mind: Gender, Strategy, and Decision-making in Statewide Races." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Ferree, Myra M. 1974. "A Woman for President? Changing Responses, 1958-1972." *Public Opinion Quarterly* 38: 390-99.
- Fox, Richard Logan. 1997. *Gender Dynamics in Congressional Elections*. Thousand Oaks, CA: Sage.

- Freeze, Melanie S., John Aldrich, and Wendy Wood. 2009. "Candidate Evaluations, Negative Messages, and Source Bias." Paper presented at the 32nd Annual Scientific Meeting of the International Society of Political Psychology, Dublin, Ireland.
- Fridkin, Kim, Patrick Kenney, and Gina Serignese Woodall. 2009. "Bad for Men, Better for Women: The Impact of Stereotypes During Negative Campaigns." *Political Behavior* 31: 53-78.
- Goldberg, Philip. 1968. "Are Women Prejudiced Against Women?" *Transaction* 5: 28-30.
- Huddy, Leonie, and Nayda Terkildsen. 1993a. "The Consequences of Gender Stereotypes for Women Candidates at Different Levels and Types of Office." *Political Research Quarterly* 46: 503-25.
- Huddy, Leonie, and Nayda Terkildsen. 1993b. "Gender Stereotypes and the Perception of Male and Female Candidates." *American Journal of Political Science* 37: 119-47.
- Kahn, Kim Fridkin. 1994. "Does Gender Make a Difference? An Experimental Examination of Sex Stereotypes and Press Patterns in Statewide Campaigns." *American Journal of Political Science* 38: 162-95.
- King, David, and Richard Matland. 2003. "Sex and the Grand Old Party." *American Politics Research* 31: 595-612.
- Lawless, Jennifer. 2009. "Sexism and Gender Bias in Election 2008: A More Complex Path for Women in Politics." *Politics & Gender* 5: 70-80.
- Lawless, Jennifer, and Kathryn Pearson. 2008. "The Primary Reason for Women's Underrepresentation? Reevaluating the Conventional Wisdom" *Journal of Politics* 70: 67-82.
- Lawrence, Regina G., and Melody Rose. 2010. *Hillary Clinton's Race for the White House: Gender Politics & the Media on the Campaign Trail*. Boulder, CO: Lynne Rienner.
- Leeper, Mark. 1991. "The Impact of Prejudice on Female Candidates: An Experimental Look at Voter Inference." *American Politics Quarterly* 19: 248-61.
- Lizotte, Mary-Kate. 2009. *The Dynamics and Origins of the Gender Gap in Support for Military Interventions*. Doctoral dissertation, Stony Brook University.
- Matland, Richard, and David King. 2002. "Women as Candidates in Congressional Elections." In *Women Transforming Congress*, ed. Cindy Simon Rosenthal. Norman, OK: University of Oklahoma Press.
- Philpot, Tasha, and Haynes Walton. 2007. "One of Our Own: Black Female Candidates and the Voters Who Support Them." *American Journal of Political Science* 51: 49-62

- Rosenwasser, Shirley, and Norma Dean. 1989. "Gender Role and Political Office: Effects of Perceived Masculinity/Femininity of Candidate and Political Office." *Psychology of Women Quarterly* 13: 77-85.
- Rosenwasser, Shirley, Robyn R. Rogers, Sheila Fling, Kayla Silver-Pickens, and John Butemeyer. 1987. "Attitudes Towards Women and Men in Politics: Perceived Male and Female Candidate Competencies and Participant Personality Characteristics." *Political Psychology* 8: 191-200.
- Rosenwasser, Shirley, and Jana Seale. 1988. "Attitudes Toward a Hypothetical Male or Female Presidential Candidate – A Research Note." *Political Psychology* 9: 591-8.
- Sanbonmatsu, Kira, and Kathleen Dolan. 2009. "Do Gender Stereotypes Transcend Party?" *Political Research Quarterly* 62: 485-94.
- Sapiro, Virginia. 1981-2. "If U.S. Senator Baker were a Woman: An Experimental Study of Candidate Images." *Political Psychology* 2: 61-83.
- Schneider, Monica Cecile. 2007. *Gender Bending: Candidate Strategy and Voter Response in a Marketing Age*. Doctoral dissertation, University of Minnesota.
- Streb, Matthew, Barbara Burrell, Brian Frederick, and Michael Genovese. 2008. "Social Desirability Effects and Support for a Female American President." *Public Opinion Quarterly* 2008: 76-89.
- Walsh, Katherine Cramer, and Virginia Sapiro. 2003. "Marketing Congressional Candidates to Male and Female Audiences: The Performance of Gender in Campaign Television Advertisements." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Welch, Susan, and Lee Sigelman. 1982. "Changes in Public Attitudes Toward Women in Politics." *Social Science Quarterly* 63: 312-22.
- Winter, Nicholas. 2008. *Dangerous Frames: How Ideas about Race and Gender Shape Public Opinion*. Chicago: University of Chicago Press.

ⁱ The office manipulation was analyzed in a separate paper (Huddy and Terkildsen 1993a).

ⁱⁱ Though see Druckman and Kam's chapter in this volume for another perspective on the use of student samples.

ⁱⁱⁱ See also Boudreau and Lupia's chapter in this volume for a discussion of experimental work on the gender gap in political knowledge.

^{iv} See Schneider (2007) for an experimental investigation of women's campaign strategies.

21. Racial Identity and Experimental Methodology

Darren Davis

Interest in how social group attachments translate into political attitudes and behavior motivated much of the early attention to social identity. According to Tajfel (1981), the underlying foundation in the development of political and social beliefs was “the shared perceptions of social reality by large numbers of people and of the conditions leading to these shared perceptions” (15), as opposed to personality and environmental characteristics. Though Tajfel was not referencing the forces that make one’s racial identity relevant to an individual, but instead focused on the development of social identity around the prejudice toward Jews, racial identity (at least in regards to political and social behavior research) has been among the most powerful explanations of behavior. Despite the multitude of identities a person may possess and the events that make such identities more or less salient, *social identity theory* has had special insight and significance into the connection between racial identity and political behavior. In the field of political behavior, no form of identity has received nearly the amount of attention and scrutiny as racial identity (e.g., group consciousness, racial consciousness, linked fate, and race identification).ⁱ

At the same time, however, research on racial identity and political behavior could benefit from adherence to the conceptual foundations of social identity theory as well as from greater reliance on experimental research. Several problems exist in the conceptual development and measurement of racial identity research that beg out for a reexamination of racial identity. These problems are closely tied to a heavily reliance on survey-based research that essentially

ignores the two-stage development of identity and thus overstates the importance of racial identity in influencing political behavior. Controlling the accessibility of racial identity among a multitude of identities a person might possess, as well as the level of identity salience, is critical to the study of racial identity. Otherwise, racial identity may be viewed as somewhat artificial, as researchers impose an identity and assert a certain level of psychological importance. Such artificiality of identity is compounded by the fact racial identity is almost always measured contemporaneously with other political and social attitudes, which makes causal statements tenuous.

In the pages that follow, I review the essence of social identity theory, explore the survey-based approaches to studying racial identity and political behavior, and then propose an experimentally-based research agenda for overcoming such limitations.

1. Social Identity Theory

Social identity theory was originally developed to explain the psychological basis of intergroup discrimination. The core idea is that people tend to simplify the world around them and that a particularly important simplifying device is the categorization of individuals, including themselves, into groups according to their similarities and differences. Describing one's self and others as African American, conservative, a woman, and so forth is a way in which categories are created and maintained. People recognize at an early age their differences and similarities to others.ⁱⁱ Such a simple and automatic classification process is often assumed to be sufficient to produce distinctive group behavior and prejudice. Early on, scholars recognized that the mere categorization or designation of group boundaries could provoke discrimination; however,

Turner (1978, 138-9) would later show that social categorization per se (the “minimal group” paradigm) was, by itself, not sufficient for in-group favoritism.

Once people have categorized themselves and others into distinct groups, self-esteem is enhanced by creating favorable comparisons of their own groups vis-à-vis other groups, thus making their own groups appear superior. Motivating individuals are the need for self-esteem and the desire for a positive self-evaluation.

This process linking categorization and self-esteem to group attachments is simple enough, but it should be clear that not all identities are equally accessible and important at the same time. Identities contribute to our self-concept but, for the most part, they need to be activated and made salient in order to be useful in political and social decision making. For the positive distinctiveness of group identities to become politically and socially important to the individual, a mechanism must exist for activating or making salient the psychological attachment to social categories. Information and political and social events that increase the salience of different identities at different times abound.

However, a different set of assumptions and processes seems to characterize the role of racial identity among African Americans. Without separating out the components of social identity theory within racial identity, the primacy of a racial identity among African Americans is assumed to be a dominant identity and, as a result, an African American racial identity is considered more easily activated and sustained than other identities. This might or might not be true, but it is a testable proposition, just the same. The safest assumption, and one that should guide methodological approach in this area, is that racial identity is highly variable and highly contextual among African Americans. A person might think of him- or herself as African

American and receive positive self-esteem from such a racial identity, but it is important to recognize that the African American identity competes with other identities. It might come as a surprise to some, but African Americans might also think of themselves as Americans, parents, teachers, middle class, and so on.

2. Racial Identity in Political Behavior Research

Most treatments of racial identity in political behavior research, such as with racial consciousness, group consciousness, linked fate, and racial identification, seem to focus on a contrived or artificial identity and fail sufficiently to capture the esteem that comes from preferring one group over another or to account for how one goes from identification to an embodiment of the group.ⁱⁱⁱ The various elements of social identity theory exist independently of each other in the literature. While racial or group consciousness could be considered to capture the salience of racial identity, linked fate, common fate, or group identity could be considered to capture the categorization of racial identity.^{iv}

Consider the public opinion literature that seeks to connect racial and group consciousness to political behavior. Verba and Nie (1972) recognize that racial consciousness or the “self-conscious awareness of one’s group membership” among African Americans could be a potent force in political participation. The authors offer few details about the origins and activation of group consciousness among African Americans, and their measurement of racial consciousness is far removed from the essence of social identity theory. Using responses to public opinion survey questions, they measure black consciousness by whether blacks voluntarily raised the issue of race in response to a series of open ended questions asking about the presence of any conflict within their communities or any problems they perceived in their

personal lives, the community, or the nation. Shingles (1981) subsequently repeated this measure to conclude that black consciousness is grounded in low political efficacy and political mistrust. A problem with this approach to racial identity is that, although African American identity may be related to attributing racial explanations, identity is not required to make such assessments.

Miller, Gurin, and Gurin (1978) define group consciousness as a “politicized awareness, or ideology, regarding the group’s relative positions in society, and a commitment to collective action aimed at realizing the group’s interests” (495). This measure supposedly differs from group identification, which “connotes a perceived self-location within a particular social stratum, along with a psychological feeling of belonging to that particular stratum.” Group consciousness is considered a multidimensional concept integrating group identification, polar affect (i.e., a preference for members of one’s in-group and a dislike for the out-group), polar power (i.e., dissatisfaction with the status of the in-group), and system blame (i.e., a belief that inequities in the system are responsibility for the status of the in-group). Miller et al.’s conceptualization of racial consciousness encompasses many of the consequences of identity, but it leaves unresolved how an individual decides for him or herself which identities are relevant and how group identity evolves from simple attachment to consciousness (or salience). Miller et al. suggest that, through behavior and interactions, individuals learn of the discontent of one’s group position, which makes the group salient or personally meaningful.

Other prominent attempts to assess racial identity among African Americans have been equally assertive in giving individuals an identity. African Americans are assumed to possess a racial identity and it is assumed to take precedence over all other possible identities. No other identities or competing attachments are considered. For instance, Gurin, Hatchett, and Jackson

(1989) initially conceive of identity as a multi-dimensional construct with different behavioral consequences; based on a common fate and an exclusivist identity, racial identity reflected an implicit affiliation with the in-group. Building off this measurement approach, Dawson (1994) intended his “linked fate” to be a simpler construct of racial identity: as African Americans observe an attachment to other African Americans, they also come to believe that their interests, mostly economic, are linked to the economic interests of their racial group (77). Unfortunately, this measure seems to be driven more by available survey-based items than by an understanding of social identity, as the mechanism through which group affiliation or even shared fate becomes salient is not explicit. As a result, it may be premature to suggest that affiliation automatically leads to linked fate (especially along an economic dimension) and that it is always a salient evaluative consideration. Within this same tradition, Tate (1994) equates common fate to racial identification.

Relating objective group membership to psychological attachment, Conover’s (1984) concept of group identification closely mirrors Tajfel’s treatment of identity: a self-awareness of one’s objective membership in the group and a sense of attachment to the group. Beginning with an awareness of their group affiliations, individuals’ salience or attachment to a group (perhaps from past experiences or in response to political events) becomes a component of their self-concept. Group identity, then, becomes a point of reference in organizing and interpreting information and guides how individuals process information concerning others (763). In following Tajfel’s initial conceptualization, Conover is able to show that objective group membership acting in concert with a sense of psychological attachment produces distinctive perceptual viewpoints. This survey-based measure first determines respondents’ objective group

membership from available survey-based measures, such as class, gender, age, and race. Though somewhat artificial, this objective measure does consider a range of identities. Then, Conover determines whether respondents feel especially psychologically attached to the objective groups to which they presumably belong. She accomplishes this by asking respondents which groups they feel particularly close to – people who are “most like you in their ideas, interests, and feelings about things.” Once the respondents finish rating how close they feel to all the groups, they are asked to pick the one group to which they feel closest.

More recent research by Transue (2007) examining identity salience and superordinate identity is also instructive. Using an experiment embedded in a public opinion survey, Transue (2007) examines the salience of multiple identities, both subgroup and superordinate, on policy preferences. Identities are primed through the random assignment of respondents to two different question treatments: one group is asked about their closeness to their ethnic or racial group (subgroup salience) and the other group is asked about their closeness to other Americans (superordinate group salience). Respondents are also assigned to two different dependent variables, willingness to improve education (superordinate treatment) and willingness to improve educational opportunities for minorities (subgroup treatment).

It is clear from these studies that racial identity research reflects more of an afterthought than an intentionally designed research agenda. Racial identity has not been the focus of specialized attention, but rather it has been an idea superimposed on existing data. As is often the case in these circumstances, such an approach creates many problems. Because attitudinal measures occur roughly at the same time in survey research, it is problematic to make causal statements about racial identity. Measures assumed to be influenced by racial identity could

actually prime racial identity. And, because survey-based approaches require questions ahead of time, racial identity or a set of identities are usually imposed on respondents, which might or might not be how they view themselves. In short, this imposed racial identity might well be viewed as contrived or artificial.

Survey-based approaches often are not conducive to studying racial identity. Experimentally-based methodology, in contrast, can provide the control necessary to measure racial identity properly and to make convincing causal statements.

3. The Value of Experiments in the Study of Racial Identity

My argument, so far, has been that the reliance on, or dominance of, the survey research enterprise in political behavior research has had a profound impact on the study of racial identity. Survey research is invaluable, but the approach to studying racial identity requires more attention. I now turn to how an experimental approach can produce more valid measures of racial identity, which would permit stronger assessment of the direction of causality.

Individuals belong to multiple groups and they possess multiple identities. In addition to racial groups, individuals may also identify with their gender, country, schools and universities, organizations and clubs, and occupations. All of the possible identities are too numerous to list and doing so would be futile because the most important groups are those that individuals select for themselves. It is almost impossible to determine which of the identities are important for an individual's self-concept, but this has not prevented those who study the connection between identity and political behavior from doing so. For African Americans, a racial identity and racial consciousness are assumed to be the most prominent identity and the identity from which they receive the most esteem. African Americans are seen as fixating on racial identity as a

consequence of their history, culture, and perceptions of racism and discrimination. I am not suggesting that a racial identity is irrelevant to a person's self-concept, but I am questioning the common assumption that racial group identity is always the most important. The reality is that racial group identity is one of many identities.

Experimental methodology seems more flexible than survey research in allowing a multiple identity approach. Similar to the salience approach, subjects can be presented with multiple identities that might conflict or be incompatible. Subjects would then be expected to identify with their most salient and relevant identities. Because there would be a choice among social groups, individuals would not be forced to respond to a priori social groups with whom they might not have a strong attachment. Such an experimental feature would make it possible for subjects themselves to identify their most salient social group.

4. Experimental Opportunity

The argument that individuals should be allowed to choose the identities they consider relevant and salient is grounded in the political tolerance literature. Beginning with the work by Stouffer (1955), political tolerance was conceived as the willingness to extend democratic rights (i.e., being allowed to speak publicly, teach in public schools, or publish books) to groups on the political right (i.e., suspected communists, atheists, and socialists). As it turned out, Stouffer's measure of tolerance assumed that certain groups in American society, particularly those on the political right, would not be extended democratic rights. By not realizing that many individuals may not find such groups as threatening, the measure of tolerance would be contaminated by ideology. Individuals on the political right would be mistaken as political. To correct this conceptual and measurement issue, Sullivan, Piereson, and Marcus (1979) suggested that

political tolerance implies willingness to permit the expression of those ideas or interests that one opposes or finds objectionable. Following this line of reasoning, Sullivan et al. (1979) proposed a content-controlled measure of tolerance, whereby individuals were allowed to identify functionally equivalent unpopular groups they opposed. Operationally, individuals in a public opinion interview were provided with a list of groups on both sides of the political spectrum (i.e., atheists, pro-abortionists, Ku Klux Klan members) from which they were to select the groups they liked the least. After the selection of their functionally equivalent groups, individuals were then presented with a series of statements about a range of democratic activities in which members of that group might participate.

The take-away from Sullivan et al.'s content-controlled measure is that functionally equivalent groups are important considerations in comparing how individuals perceive groups. Instead of assigning or assuming an identity based on some pre-determined characteristic, such as race or gender, individuals must be allowed to choose their own identities.

An interesting question is of how such an approach would work for racial identity. For starters, it would be important for individuals to select groups with whom they closely identify (of course without using the ambiguous term "identify"). Borrowing from the racial identity literature (Conover 1984), individuals could be asked about the groups "*they feel particularly close to and people who are most like them in their ideas, interests, and feelings about things.*" Similar to the tolerance measure, which was asked about four groups individuals like least, this identity measure could also ask about the top four identity groups.^v

Next, for each of the groups it would be necessary to determine the identity salience or psychological attachment. Assuming that racial identity is among the selected identities, it would

be important to distinguish the salience of racial identity from the salience of other identities. Thus, priming identities by assigning the same treatment to everyone (asking the same follow-up questions across the board) would be problematic because each respondent would have each identity primed or made salient in the same survey context and over a matter of seconds. Because such an approach is taxing on the individual and each identity would be primed temporarily, this is not an ideal approach to assessing the role of identity. Actually, this approach would be worse than imposing a single identity.^{vi}

An interesting approach would entail randomly assigning high salience and low salience primes for each identity an individual selects. In this way, each individual receives only one primed identity (either high or low), which can then be compared to similar identities or compared to a similarly primed alternate identity. Such an approach would be a direct test of the salience of racial identity over an alternate identity. Equally important, such an approach would be a direct test of racial identity against itself and at different levels of salience.

[Figure 21-1 about here]

Consider the example depicted in Figure 21-1, in which individuals are allowed to select a number of identities they consider important (Race, Identity-1, Identity-2, Identity-3) and a randomized assignment of salience for each identity group (High and Low).^{vii} Individuals would be randomly selected and only one identity per person randomly primed (though it would facilitate matters if, across a certain number of individuals, a racial identity could be selected). It is often the case where we are interested in examining racial identity at different levels of salience. The expectation is for high racial identity salience to be more powerful than low racial identity salience in predicting some form of political behavior. If there were no difference

between them, we could conclude that racial identity was unimportant. Another important test involves the extent to which racial identity is more influential than other identities. Thus, instead of assuming that racial identity is more salient than other identities, it could be tested empirically.

5. Conclusion

This essay is about how survey-based approaches can contribute to a flawed conceptualization of racial identity in political behavior research and how experimental methodology might involve a better approach. Perhaps the most serious problem takes the form of imposing an artificial or contrived identity. Individuals possess a multitude of identities that become more or less salient with information and in the interaction with others. Instead of seeking to capture a range of these identities among African Americans, there has been a tendency for researchers to impose a racial identity, regardless of whether or not such an identity is relevant to the individual.

Causal statements are made concerning racial identity when all attitudinal measures are measured contemporaneously. The political and social attitudes that racial identity has been expected to influence are just as likely to determine racial identity. The greatest value of an experimental methodology is its capacity to make stronger claims about the causal relationships of racial identity.

References

- Conover, Pamela Johnston. 1984. "The Influence of Group Identifications on Political Perception and Evaluation." *Journal of Politics* 46: 760-85.
- Dawson, Michael C. 1994. *Behind the Mule: Race and Class in African-American Politics*. Princeton New Jersey: Princeton University Press.

- Gurin, Patricia, Shirley Hatchett, and James S. Jackson. 1989. *Hope and Independence: Blacks' Response to Electoral and Party Politics*. New York: Russell Sage Foundation.
- Miller, Arthur, Patricia Gurin, and Gerald Gurin. 1981. "Group Consciousness and Political Participation." *American Journal of Political Science* 25: 494-511.
- Shingles, Richard D. 1981. "Black Consciousness and Political Participation: The Missing Link." *American Political Science Review* 75: 76-91.
- Stouffer, Samuel. 1955. *Communism, Conformity, and Civil Liberties*. New York: Doubleday.
- Sullivan, John L., James Piereson, and George E. Marcus. 1979. "An Alternative Conceptualization of Political Tolerance: Illusory Increases 1950s-1970s." *American Political Science Review* 73: 781-4.
- Tajfel, Henri. 1981. *Human Groups and Social Categories*. Cambridge: Cambridge University Press.
- Tate, Katherine. 1994. *From Protest to Politics: The New Black Voters in American Elections*. New York: Russell Sage Foundation.
- Transue, John E. 2007. "Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force." *American Journal of Political Science* 51: 78-91.
- Turner, John C. 1978. "Social Categorization and Social Discrimination in the Minimal Group Paradigm." In *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*, ed. Henri Tajfel. London: Academic Press.
- Verba, Sidney and Norman H. Nie. 1972. *Participation in America: Political Democracy and Social Equality*. New York: Harper and Row.

Figure 21-1. Example of Experimental Design for Racial Identity

Selected Identities				
	Race	Identity-1	Identity-2	Identity-3
Randomized Level of Salience	High	High	High	High
	Low	Low	Low	Low

Example of testable hypotheses:

$$H_1: \text{Race}_{\text{High}} > \text{Race}_{\text{Low}};$$

$$H_2: \text{Race}_{\text{High}} = \text{Race}_{\text{Low}} > 0;$$

$$H_0: \text{Race}_{\text{High}} = \text{Race}_{\text{Low}} = 0$$

$$H_3: \text{Race}_{\text{High}} > \text{Identity-3}_{\text{High}};$$

$$H_4: \text{Race}_{\text{High}} = \text{Identity-3}_{\text{High}} > 0;$$

$$H_0: \text{Race}_{\text{High}} = \text{Identity-3}_{\text{High}}$$

ⁱ Research has indeed focused on other identities, such as patriotism, nationalism, gender, and social class. But research on these identities appears to lag behind research on racial identity. Equally, important, social identity theory has not been as readily applied to those identities.

ⁱⁱ Categorization leads to the formation of stereotypes to aid in the processing of information, but the positive and negative attributions underlying discrimination occur when individuals interact with others.

ⁱⁱⁱ By contrived, I mean that it is almost impossible to determine *a priori* the multitude of identities one may possess. But, in the construction of survey research, the researcher has to decide which identities to measure. Thus, these two processes seem somewhat incompatible.

^{iv} Though one can argue that racial identity is different from racial consciousness, group identity, or linked fate, I see those concepts as tapping different aspects of the same multi-component of racial identity. They simply tap different dimensions of racial identity. Whereas group identity and racial identity may be viewed as assessing the identity component of racial identity, group consciousness and racial consciousness may be viewed as assessing the salience component.

^v Another way of measuring this first part of identity could also involve linked fate or common fate measures.

^{vi} The likelihood of individuals selecting the same identities is very low, but this approach requires only that individuals select a racial identity. Because the alternate identities would be used only for comparison, the actual content of those identities are not important.

^{vii} These identities can be any identities, as long as the individual selects them. With the exception of a racial identity, the alternative identities do not have to be identical across individuals. The interest is only in a racial identity.

22. The Determinants and Political Consequences of Prejudice

Vincent L. Hutchings and Spencer Piston

Researchers have been interested in the distribution of prejudice in the population, as well as its effects on policy preferences, vote choice, and economic and social outcomes for racial minorities, since the dawn of the social sciences. One of the earliest scholars to address these questions was W.E.B. DuBois in his classic work, *The Philadelphia Negro* (DuBois 1899). Referring to prejudice in chapter sixteen of his book, DuBois wrote that, “Everybody speaks of the matter, everyone knows that it exists, but in just what form it shows itself or how influential it is few agree” (322). Although these words were written over 100 years ago, this observation still does a good job of summarizing our understanding of prejudice. There have been, to be sure, significant advances in this literature since the time of DuBois but social scientists continue to disagree about the influence of racial prejudice in modern American politics. The aim of this chapter is to explore and adjudicate some of these differences, paying particular attention to the strengths, limitations and contributions provided by the use of experimental methods.

Before examining the ways that scholars have studied prejudice, it is important that we define this term. What is prejudice? Social psychologists were among the first to answer this question. For example, Allport defined (ethnic) prejudice as:

An antipathy based upon a faulty and inflexible generalization. It may be felt or expressed. It may be directed toward a group as a whole, or toward an individual because he is a member of that group (Allport 1979, 9).

As we shall see, subsequent scholars would expand and modify this definition in significant ways. However, the importance of faulty generalizations or stereotypes has remained a central component of virtually all of the definitions that would follow Allport. We will therefore rely upon this broad definition unless otherwise indicated.

1. The Study of Prejudice in the Political Science Literature

The political science literature on prejudiceⁱ has focused primarily on two questions: the role that prejudice plays in structuring policy preferences, and whether or not prejudice influences candidate preferences. This latter subject has centered on contests involving two white candidates, as well as elections between a white candidate and an African American candidate. Each of these areas of study has focused almost exclusively on white attitudes and has sought to determine if prejudice remains a dominant, or at least significant, predictor of preferences in the post-Civil Rights era.ⁱⁱ As we discuss in the next section, observational work on each of these questions has produced a number of contributions but has often failed to isolate the precise role that prejudice plays in public opinion, as well as the circumstances in which its influence is more or less powerful. We highlight in this chapter some of the ways in which experiments have helped to address these questions.

Racial Policy Preferences

Much of the early work on prejudice in the political science literature relied on observational studies. Sears and Kinder (1971), for example, used cross-sectional survey data to explore the impact of prejudicial attitudes on policy preferences and candidate support in Los Angeles. They would go on to develop the theory of symbolic racism, which maintains that a new and subtler form of racism emerged in the aftermath of the Civil Rights movement, spurred

in part by the urban riots of the late 1960s. Sears and Kinder argue that symbolic racism, unlike previous manifestations of anti-black prejudice, did not posit the biological inferiority of African Americans. Rather, the latent antipathy that many whites still felt toward blacks was now combined with the belief that blacks did not try hard enough to get ahead and violated traditional American values such as hard work and individualism (Kinder and Sears 1981; Sears 1988).ⁱⁱⁱ In order to test this theory, Sears and Kinder relied upon an attitudinal scale composed of several different survey items. Respondents are asked to what extent they agree with the following statements, among others: 1) “Irish, Italian, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same.”; 2) “It’s really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.” In general, the symbolic racism scale has been shown to powerfully and consistently predict candidate support and racial policy preferences.^{iv}

Although the symbolic racism scale has frequently emerged as the most powerful correlate of racial policy preferences, a number of critics have challenged this construct (Sniderman and Tetlock 1986; Sniderman and Piazza 1993; Sidanius and Pratto 1999). Their critiques take a variety of forms but, for our purposes, the most relevant questions involve conceptual and measurement issues. Sniderman and his various coauthors offer an alternative, although perhaps not wholly incompatible, view of the role of race in modern American politics. This perspective focuses on the institutional or political forces that structure these views. In other words, they argue that the ways in which politicians *frame* racial issues determine how most Americans express their racial policy preferences.^v Thus, while it is true that Americans disagree

on racial policy questions, this disagreement owes more to partisan or ideological differences than it does to racial attitudes per se (Sniderman and Carmines 1997).

In order to support this alternative view, Sniderman and his colleagues rely heavily upon question wording experiments embedded in national surveys. This approach has the advantage of strong internal and external validity. In one study, Sniderman and Carmines (1997) develop what they refer to as the “Regardless of Race Experiment.” In this experiment, a random half of their sample is asked about their support for job training programs for blacks after being told that, “some people believe that the government in Washington should be responsible for providing job training to [blacks] *because of the historic injustices blacks have suffered*” (italics added). The second half of the sample is also asked about job training programs but with the following rationale: “Some people believe that the government in Washington should be responsible for providing job training to [blacks], *not because they are black, but because the government ought to help people who are out of work and want to find a job, whether they’re white or black.*” Sniderman and Carmines expect the second frame, owing to its more universal character, to be much more popular among whites. As expected, they find that support is higher for race-targeted job training programs when they are justified in universal terms rather than strictly racial ones (thirty-four percent versus twenty-one percent), consistent with their broader argument.

In another experiment, dubbed the “Color-Blind Experiment,” Sniderman and Carmines test their argument about the appeal of universal programs more directly. With this experiment, white respondents were divided into three groups and asked whether the federal government should seek to improve conditions for “blacks who are born into poverty...because of the continuing legacy of slavery and discrimination,” or in order to “make sure that everyone has an

equal opportunity to succeed.” The last group is distinguished from the other two in that, instead of asking about blacks, respondents were asked about efforts to alleviate poverty for “people” because, as in the second condition, it is the government’s role to “make sure that everyone has an equal opportunity to succeed.” They find that support for anti-poverty programs increase by eighteen percentage points when the policy is described in universal terms (condition three) relative to race-specific, and racially justified, terms (condition one).

Sniderman and Carmines acknowledge that their finding, by itself, does not rule out the possibility that it could be the result of anti-black attitudes. However, they argue that if prejudice is driving the lower support for policies targeted at blacks, then whites who are less committed to racial equality should be the most likely to embrace programs that do not mention race. Although plausible, this is not what they find. White respondents who are committed to racial equality are much more likely to support government assistance to the poor no matter how the program is framed. More importantly, moving from racially targeted and racially justified characterizations to a more universal frame increases support for these programs at similar rates for those scoring high or low on their racial equality scale.

Kinder and Sanders (1996) employ a similar experimental manipulation of question wording in their examination of support for government efforts to assist blacks. In 1988, half of the American National Election Studies (ANES) sample was asked about efforts to improve the social and economic position of blacks (condition 1) whereas the other half was asked about blacks and other minorities (condition 2). As with Sniderman and Carmines (1997), Kinder and Sanders find that support for this policy increases by about seven percentage points in the latter condition. They get larger effects in the same direction among blacks, although these results fall

short of statistical significance. They conclude that, “race neutral programs do appear to be more popular among the American public, black and white” (184).

Experiments such as these do provide some conceptual clarity that is often missing when researchers rely primarily on observational studies. Instead of asking the reader to accept a particular interpretation of an attitudinal construct, Sniderman and Carmines simply manipulate the rationale behind the policy (i.e., group-specific or universal) or, in the case of Sniderman and Carmines as well as Kinder and Sanders, whether a policy refers specifically to African Americans, or is applied to some broader group. They typically find diminished support for the racially targeted program, which might be the result of prejudice. However, even when Sniderman and Carmines focus only on racially targeted programs, as in the “Regardless of Race Experiment,” they find that more universal justifications lead to greater support among whites. Similarly, with the “Color-Blind Experiment” they, and Kinder and Sanders, find that the role of prejudice in prompting this greater support is minor, as universal programs are more popular with liberals, conservatives and African Americans. Still, even results derived from an experiment can be open to interpretation. In order to draw an inference that differences across conditions are due to race and only race, the experimental conditions must differ on this dimension and nothing else. When this is not the case, it is impossible to isolate the source of the difference (or lack of difference) across conditions.

For instance, the “Regardless of Race Experiment” purports to show that universal justifications are more compelling than race-specific ones but, in this case, the appropriate inference is uncertain. The two aforementioned conditions do not just differ in terms of whether or not the rationale for the policy is group-specific. The justifications also differ in terms of their

emphasis on the past and in their characterization of the unemployed as eager “to find a job.” They may also differ in terms of their intrinsic persuasiveness in that a reference to “historic injustices” might be too vague and consequently less convincing than references to unemployed workers seeking to find a job. In short, we cannot be sure if justifications framed in universal terms are inherently more persuasive or *if this particular race-specific rationale is simply inferior to the specific universal rationale they employed*. A more convincing race-specific rationale might have asked about support for job training programs “because government ought to help blacks who are out of work and want to find a job.” In this example, the only differences across conditions would be the use of the term black and so any difference could only be attributed to the race-specific nature of the justification.

The Kinder and Sanders experiment is open to a somewhat different criticism. Although the authors conclude that race-neutral programs are more popular, the inclusion of other minority groups to the question (condition 2) arguably does not diminish the role of race and is therefore not race-neutral. It is clear that referencing other minority groups increases the popularity of the program but it does not necessarily follow that the absence of any reference to race would also increase popularity. The discussion of the results also implies that blacks and whites are more supportive of programs to assist blacks *and other minorities* for the same reasons (i.e., universal programs are more popular). It is possible for example, as various studies discussed later will indicate, that those whites who find broader programs more appealing adopt this view because of their negative attitudes about blacks. African Americans, on the other hand, may be motivated by a sense of solidarity with other nonwhite groups and may or may not view truly race-neutral programs more favorably.

Some studies have addressed these concerns about precision in experimental designs. Iyengar and Kinder (1987) provide one example. These researchers were interested in the impact of television news reports on the perceived importance of particular issues. Instead of modifying survey questions, Iyengar and Kinder altered in unobtrusive ways the content of network news accounts on various national issues. In one experiment, subjects viewed one of three stories about an increase in the unemployment rate. The content was constant across conditions except that in the two treatment conditions, the discussion of employment information was followed by an interview with a specific unemployed individual. In one case this person is white and in the other the person is an African American. Given that the only salient difference across the two treatment conditions is the race of the unemployed worker, lower levels of concern with the unemployment problem when the black worker is shown would represent evidence of racial prejudice. This is exactly what the authors find. Moreover, consistent with the prejudice hypothesis, they find that whites who have negative views of blacks are most likely to diminish the importance of unemployment when the worker is an African American. We should note, however, that these results are not entirely unassailable as they achieve only borderline statistical significance raising concerns about the reliability of these findings. Further discussion of this experiment can be found in Iyengar's chapter in this volume.

Reyna et al. (2005) address some of the limitations identified in earlier experimental work on prejudice. These scholars rely upon a question wording experiment in the 1996 General Social Survey. In this within-subjects experiment, respondents were presented with a question about racial preferences with either blacks or women identified as the target group. Reyna and her colleagues find that overall respondents were significantly more supportive of this policy

when women were the target group rather than African Americans, suggesting that not all group-specific policies are created equal. Moreover, following up on the source of these effects, Reyna and her colleagues report that political conservatism among college-educated respondents was significantly related to greater opposition to preferences for blacks. They report that these effects are mediated by “responsibility stereotypes” which turn out to be measured by one of the standard items in the symbolic racism index.^{vi} Providing additional confidence in their results, Reyna and her colleagues replicate their basic findings with a convenience sample of white adults from the Chicago area (N=184).

Increasingly, researchers have begun to focus on the circumstances under which racial prejudice influences the policy positions of whites, rather than whether or not these attitudes play any role in structuring public opinion. In one of the earlier examples of this approach, Nelson and Kinder (1996) had their subjects view and evaluate several photographs of individuals engaged in routine, exemplary, or scandalous activities. Specifically, eighty-four University of Michigan undergraduates were randomly assigned to one of three conditions where they either viewed photos of whites engaged in activities such as gardening (the control group), or pictures of blacks engaged in stereotypic activities such as illegal drug use, or counter-stereotypic imagery of blacks interacting with their family or in school settings. If, as some have argued, negative attitudes about African Americans are often dormant among whites, then exposure to frames that serve to remind them of these stereotypes might enhance the influence of prejudice. Consistent with this view, Nelson and Kinder find that the relationship between attitudes about blacks and support for racial preferences is significantly stronger for subjects exposed to stereotypic depictions of African Americans.

Although the symbolic racism researchers have primarily relied upon observational studies, they have recently called on experimental methods to defend various elements of the theory. One particularly persistent criticism was that symbolic racism theorists had never empirically demonstrated that this new form of racism derives from a blend of anti-black affect and traditional American values. In a 2003 article, Sears and Henry sought to address this criticism. Utilizing a split-ballot design in the 1983 ANES Pilot Study, they adapted the six-item individualism scale so that each item referred specifically to either blacks or women. Respondents received only one of these scales, depending on which ballot they were provided, but all participants received the general individualism scale that made no reference to race or gender. If individualism in the abstract is primarily responsible for white opposition to group-specific preferences, then the individualism scale adapted to apply to women should be as strongly linked to opposition to racial preferences as the individualism scale that was modified to apply to blacks. Similarly, the general individualism scale should also share the same predictive properties as the black individualism scale. If, on the other hand, opposition to racial preferences is primarily driven by individualistic principles applied only to blacks, then one of the core assumptions of the symbolic racism theory would be sustained. Consistent with their theory, they find that only the black individualism scale is significantly correlated with opposition to racial preferences.

Although Sears and Henry (2003) resolve one of the outstanding criticisms of the theory of symbolic racism, it is still unclear whether the construct is confounded with political ideology. Feldman and Huddy (2005) provide some support for this contention. These researchers relied upon a question wording experiment embedded in a representative sample of white New York

state residents (N=760). Respondents were asked whether they supported college scholarships for high-achieving students. The treatment consisted of manipulating whether the students were described as white, black, poor white, poor black, middle-class white, middle-class black, middle-class, or simply poor. The racial group categories alone do not produce any evidence of a double standard but, once class was introduced, Feldman and Huddy find evidence of racial prejudice. Specifically, respondents were much more supportive of college scholarships for white middle-class students (sixty-four percent) than they were for black middle-class students (forty-five percent). Also, they find that the symbolic racism scale is associated with this racial double standard but only for self-identified liberals. Conservatives scoring higher on the symbolic racism scale are more likely to oppose the scholarship program for all groups, not just blacks.^{vii} This suggests that the symbolic racism scale acts more like a measure of political ideology among conservatives. These results are not consistent with the theory of symbolic racism but it is possible that Feldman and Huddy's results might differ if run on a national sample.

Nonracial Policy Preferences

In addition to experiments designed to explore the role of prejudice in shaping attitudes on racial policies, scholars have also used this tool to examine the influence of prejudice on ostensibly nonracial policies. Most of this work has focused on crime or welfare policy. For example, Gilliam and Iyengar (2000) explored whether the race of criminal suspects influenced levels of support for punitive crime policies among viewers of local television news. Unlike most of the experimental work in political science, they included whites and African Americans in their convenience sample (N= 2331). Gilliam and Iyengar present their subjects with one of three

different (modified) newscasts: one featuring a black criminal suspect, one featuring a white suspect, and one version in which there is no photograph or verbal description of the suspect. With the exception of the race of the alleged perpetrator, the newscasts are identical in every way. As anticipated, they find that support for punitive crime policies increase by about six percentage points when subjects view the black suspect, but the effects are much weaker and statistically insignificant when the suspect is white or ambiguous. Interestingly, Gilliam and Iyengar find that this effect only applies to the whites in their study, as the effects are either insignificant or run in the “wrong” direction for blacks. This result provides additional support for their claim that some form of racial prejudice contributes to white support for punitive crime policies. In light of these findings, experimentalists assessing the role of prejudice in shaping policy preferences should include minority subjects whenever possible (see Chong and Junn’s chapter in this volume).

Peffley and Hurwitz (2007) engage in a similar analysis although they focus on the issue of support or opposition to the death penalty. Specifically, they are interested in whether exposure to various arguments against the death penalty would reduce support among blacks and whites. In the first of their two treatment groups, drawn from a national telephone sample, Peffley and Hurwitz present their respondents with the following preamble before asking their views on the death penalty: “some people say that the death penalty is unfair because too many innocent people are being executed.” Their second treatment group is presented with a different introduction: “some people say that the death penalty is unfair because most people who are executed are African Americans.” Finally, respondents in the baseline condition are simply asked their views on this policy without any accompanying frame. They report that both frames are

persuasive among African Americans. In the innocence condition, support for the death penalty drops by about sixteen percentage points and in the racial condition it drops by twelve percentage points. Whites, on the other hand are entirely unaffected by the innocence treatment but, surprisingly, support for the death penalty increases by twelve percentage points in the racial condition. Peffley and Hurwitz attribute this to a priming effect wherein individual attributions of black criminality become much more predictive of support for the death penalty in the racial condition than in either of the other conditions.

The finding that white support for the death penalty increases when respondents are informed that blacks are more likely to be on death row is striking but we should interpret this result with some caution. For starters, unlike many others, Peffley and Hurwitz do not find that anti-black stereotypes contribute to white support for the death penalty. Additionally, their experiment was designed to examine reactions to particular anti-death penalty appeals, rather than to isolate the specific role of anti-black prejudice in shaping death penalty views. If this latter aim were their goal, then perhaps they would have designed conditions noting that, “most people who are executed are men” or “most of the people executed are poor.” Both of the previous statements are true and if support for the death penalty increased among whites in the race condition, but not the others, then this would represent strong evidence of a racial double standard.

The other prominent, and ostensibly nonracial, policy domain that may be influenced by anti-black prejudice is welfare policy. Gilens (1996) examined this issue utilizing a question wording experiment embedded in a 1991 national survey. Additionally, these results were supplemented with a mail-in questionnaire delivered to the respondents who completed the

telephone survey. The respondents in this study were assigned to one of two conditions: in the first condition they were asked their impressions of a hypothetical welfare recipient characterized as a thirty-something black woman with a ten-year old child, who began receiving welfare in the past year. In the second condition, all of these attributes are identical except that the woman in question is described as white. Gilens finds that the white and African American welfare recipients are evaluated similarly but, in the black condition, negative attitudes about welfare mothers are much more likely to influence views of welfare policy. Gilens concludes that whites often have such negative attitudes about welfare because their prototypical recipient is a black woman rather than a white woman.

Experiments and Prejudice Beyond the Black-White Divide in the U.S.

While most of the experimental work in political science on the subject of prejudice has focused on white attitudes about blacks and related policies, some recent work has also begun to focus on attitudes about Latinos. The key policy domain in this literature is typically immigration. For example, Brader, Valentino, and Suhay (2008) employed a two-by-two design manipulating the ethnic focus of an immigration story (Mexican or Russian) as well as the tone (negative or positive). Their subjects participated over the Internet and were drawn from the random digit dial (RDD) selected panel maintained by Knowledge Networks. The authors find that opposition to immigration rises substantially among whites when the negative story highlights immigration from Mexico and that respondent anxiety is the principal mediator of this result.

There is also an emerging literature in political science that utilizes experiments to explore the influence of prejudice outside of the U.S. This research often must confront

challenges that are not present in the American context such as depressed economic and social conditions, as well as multiple national languages. Gibson (2009) has managed to overcome these hurdles in his study about the politics of land reconciliation in South Africa. In one experiment embedded in a nationally representative sample, Gibson exposed black, white, colored, and Asian respondents to one of several vignettes about a conflict over land ownership. In each version of the vignette, one farmer claims that land currently occupied by another farmer was stolen from him and his family during the apartheid era. There were multiple versions of these vignettes, but the key manipulations involved the race of the farmer claiming current ownership of the land and the judicial judgment as to who rightfully owned the property. In some cases, the dispute involves two black South Africans and in other cases the contemporary occupant is white and the farmer making the historical claim is black. In examining respondents' views as to whether the outcome was fair, Gibson finds that the race of the respondent as well as the race of the claimants and the nature of the judgment affected perceptions of fairness. White and black South Africans differed most significantly. Whites were much more likely to judge the outcome as fair if the contemporary owner of the land was awarded ownership and also described as white. Blacks, on the other hand, were considerably more likely to view the outcome as fair if the farmer with the historical claim was awarded the land, and if the losing claimant was white.

The work of Sniderman et al. (2000) represents another intriguing experiment conducted outside of the United States. The goal of the experiment was to determine whether expressed prejudice was more a function of the attitude-holder than the attitude-object. That is, rather than prejudice being “bound up with the specific characteristics of the out-group” (53), the authors

hypothesize that prejudice is a function of an intolerant personality; that is, intolerant people will express prejudice against any out-group. In order to test this hypothesis, the authors assign Italian citizens to conditions in which two out-groups, immigrants from Africa and immigrants from Eastern Europe, are evaluated along two dimensions: the extent to which they have negative personal characteristics, and the extent to which they are responsible for social problems in Italy. Subjects are randomly assigned to evaluate either Eastern European immigrants on both dimensions, African immigrants on both dimensions, or Eastern European immigrants on one dimension and African immigrants on the other. The authors find that prejudice toward Eastern Europeans is as strong a predictor of prejudice toward Africans as is prejudice toward Africans on another dimension. These results lend strong support to the authors' argument that prejudice rests more in the eye of the beholder than in the characteristics of the out-group being evaluated.

2. Prejudice and Candidate Choice

The question of whether white voters discriminate against black candidates is still an open one. Observational work has been suggestive, but suffers from the inability to isolate candidate race, leaving open the possibility that confounding variables are driving the (lack of) results. We briefly review two of the best examples of observational work on this question here. As we will see, their limitations yield an important opportunity for experimental work to make a contribution. We argue, however, that experiments on racial discrimination in the voting booth have not yet taken full advantage of this opportunity. In particular, imprecision in the experimental design has made it difficult to rule out the possibility of alternative explanations. We therefore recommend that future work pay increased attention to this issue and we also argue that, given the mixed results, scholars should turn from the question of whether white racism

hurts black candidates to begin identifying conditions under which prejudice hurts the chances of black candidates.

Highton (2004) examines U.S. House elections in 1996 and 1998. Using exit polls conducted by the Voter News Service, Highton measures discrimination as the difference between white support for white candidates and white support for black candidates, controlling for such factors as incumbency, funding, experience, and demographic characteristics. He finds no difference between white support for white and black candidates and on that basis determines that white voters showed no racial bias in these elections.

However, since Highton uses exit poll data, he lacks a measure of racial attitudes. As a result, he cannot assess whether prejudice is tied to vote choice. To be sure, by itself this may not be much of a problem, as Highton directly examines whether there is a racial double standard. But the process that determines candidate race may be endogenous to the vote choice decision. For example, consider the hypothetical case of black political figure A who decides to run for office. His personality is no more appealing than is the norm for politicians, so he loses the primary due to racial discrimination—that is, he is not sufficiently exceptional to overcome the racial bias of some white voters. He therefore is not considered in Highton's analysis, because Highton counts that contest as one without a black candidate. Now consider the hypothetical case of black political figure B who has an exceptionally appealing personality. He wins the primary because his outstanding personality overwhelms the effect of prejudice. He now counts as a black candidate in Highton's analysis. If this scenario is common, so that only black candidates who are exceptional among politicians make it through the primary and to the general House election, it is possible that discrimination does hurt black candidates in House elections, but that

such discrimination is not evident due to the effects of other candidate characteristics. If Highton had a measure of racial prejudice, and found it to be uncorrelated with vote choice, this might mitigate the aforementioned concern, but he does not.

Further, Highton does not control for competitiveness of the contest; it could be that white voters are voting for black candidates simply because they lack alternative viable options. Finally, Highton does not measure turnout, leaving open the possibility that white discrimination operates through the failure to show up to the voting booth. Lacking a measure of racial attitudes, lacking the ability to assign candidate race, and lacking control over such candidate characteristics as age, name recognition, ideological orientation, and personality, Highton cannot rule out the possibility of confounding variables.

Citrin, Green, and Sears (1990) examine a case study of Democratic candidate Tom Bradley's loss to Republican candidate George Deukmejian in the 1982 gubernatorial contest in California. Unlike Highton, they have access to measures of racial attitudes, making use of data from polls conducted by the *Los Angeles Times* and the Field Institute that were conducted among a statewide sample of white Californians. Importantly, however, the authors do not simply measure the effect of racial attitudes on vote choice, because they recognize that racial attitudes are deeply implicated in policy attitudes. Racial attitudes might have an effect on voting, therefore, not due to the candidate's race but because voters might bring their racial attitudes to bear on their evaluations of the candidate's policy platform. The authors therefore pursue the clever strategy of comparing the influence of racial attitudes on vote choice for Bradley to the influence of racial attitudes on vote choice for other Democrats pursuing such

state offices as lieutenant governor. They find no additional effect of prejudice on vote choice for Bradley, and conclude that Bradley's race did not hurt him among white voters.

As Citrin et al. recognize, their attempt to control for all other relevant factors besides race is by necessity incomplete. Though other candidates for state office may have shared membership in the Democratic Party with Bradley, they surely did not share the exact same policy platform. Further, Bradley's personality, experience, and name recognition were different. He was even running for a different office. We would have increased confidence in the work of Citrin and his colleagues if their controls for other candidate factors were less crude than simply having other white Democrats represent the counterfactual white Bradley. Such is the potential for experiments; to control for factors that observational studies cannot, in order to be sure that any relationships (or lack thereof) between candidate race and vote choice are not an artifact of some other relationship. Indeed, given that very few of the African American members of the House of Representatives hail from majority-white districts, it seems plausible that greater control over candidate characteristics might yield a finding of anti-black discrimination among whites.

Moskowitz and Stroh (1994), for example, presented their undergraduate experimental subjects with a realistic editorial, campaign brochure, and photo, all describing a hypothetical candidate (though subjects were not told that the candidate was hypothetical). Subjects read these materials and then evaluated the candidate. Subjects were randomly assigned to one of two groups wherein the descriptions were equivalent in all respects except one: the race of the candidate. Since the race of the candidate was the only thing that varied, Moskowitz and Stroh can be more confident than can Highton or Citrin and his colleagues that any difference between

groups was a result of candidate race, and thus have the potential to demonstrate stronger evidence of a racial double standard. Experimental control also gives Moskowitz and Stroh the ability to assess the mechanism through which prejudice affects the vote choice, since they measure subjects' perceptions of the candidate's policy positions and personality characteristics. Indeed, Moskowitz and Stroh find that prejudiced white subjects discriminate against black candidates, and that they do so by attributing to black candidates unfavorable character traits and policy positions with which they disagree.

Although Moskowitz and Stroh's work overcomes some of the problems inherent in observational studies, they also encounter a unique set of limitations. For example, the experimental context differed from an actual campaign environment in one potentially devastating way. Subjects were asked questions about their racial attitudes just prior to being presented with material about the hypothetical candidates. As a result, racial considerations may have been primed, causing subjects to bring their racial attitudes to bear on candidate evaluations when they might not otherwise have done so. Since this characteristic of their experiment is artificial, it could be that most of the time the race of the candidate does not influence their vote choice. Moskowitz and Stroh may have identified that candidate race can matter under certain conditions—but it is not certain that those conditions occur in the real world.

Terkildsen (1993) avoids some of the problems of artificiality in the design of her study. First, she measures racial attitudes in the post-test and thus does not run the risk of priming racial attitudes just before asking subjects to evaluate candidates. Second, Terkildsen uses an adult convenience sample selected for jury duty, decreasing generalizability concerns somewhat. However, Terkildsen's analysis also shares a limitation with Moskowitz and Stroh's. Unlike in

an actual election, in which voters typically choose between candidates, subjects were asked to evaluate the candidate in isolation. This is a particularly important limitation, given that previous work suggests that stereotypes may work differently when candidates are evaluated alone instead of in comparison to each other (Riggle et al. 1998).

Sigelman et al. (1995), on the other hand, do ask subjects to choose between two candidates, one of which is black in one condition but white in the other condition. Using a convenience sample composed of adults selected for jury duty in the Tucson, Arizona area, the authors find no evidence of race-based discrimination. The authors use a nine-cell design, and Candidate A is identical across cells: a conservative candidate whose race is unidentified. Candidate B varies by ideology (conservative, moderate, or liberal) and by race (described by the experimenters as “Anglo,” “black,” or “Hispanic”).

Sigelman and her colleagues also claim to find evidence of an interaction effect between race and ideology, in which racial minorities are perceived as more liberal than are white candidates. The authors manipulate ideology by changing the content of each candidate’s speech; subjects are expected to infer the ideology of the candidates by reading their speeches. Unfortunately, the content of the speech varies not just in the ideological principles espoused, but also in the highlighting of racial issues. Whereas the conservative candidate argues that minorities “have become too dependent on government,” and the liberal candidate claims that minorities “have been victims of terrible discrimination in this country,” the moderate candidate does not mention race at all. As a result, it is difficult to tell whether subjects evaluating the ideology of a given candidate were reacting to the candidate’s race, the ideological principles

espoused in the candidate's speech, the extent to which racial issues were highlighted in the speech, or, quite plausibly, interactions among the three.

The experimental design employed by Reeves (1997) overcomes many of these problems. For example, to avoid potential unintended priming effects, racial attitudes are measured six months prior to the implementation of the experiment. Further, respondents, identified in a representative mail survey of Detroit area residents, evaluate the candidates in a comparative context. Evaluations are based on realistic newspaper articles describing a debate between two fictitious candidates. Finally, in a four-cell design, Reeves manipulates the race of one of the candidates (black or white) and the issue area being debated (the environment or affirmative action). Thus, Reeves' design allows him to directly examine the impact of highlighting racial issues.

Unfortunately, Reeves only analyses those subjects who choose to respond to his questionnaire, neglecting to consider the possibility that the decision to respond is endogenous to the effect of the treatments. In Reeves' experiment, unlike with most survey experiments, subjects receive a questionnaire by mail and then can read it before determining whether they want to mail it back to the experimenter. Exposure to the treatment, therefore, may have some impact on the decision to participate in the study, but Reeves does not analyze whether response rates vary across experimental conditions.

Reeves claims to find evidence of racial discrimination when the campaign issue is affirmative action, but what is actually evident in the data is an increase in the number of white subjects who claim to be undecided. To be sure, Reeves finds that the distribution of racial

attitudes among these respondents suggests that they would probably not support a black candidate, but this argument is only suggestive.

A somewhat more recent vintage of studies on the influence of racial prejudice in candidate selection has focused more on indirect effects. With these studies, the emphasis is on contests featuring two white candidates and the prospect that subtle racial cues are employed to the disadvantage of one of the candidates. According to this literature, political candidates in the post-Civil Rights era no longer make direct racial appeals to whites as such efforts would be repudiated by voters across the political spectrum. Instead, covert references are made to race, leading whites to bring their latent anti-black attitudes to bear on candidate preferences. This process has been dubbed “racial priming” (Mendelberg 2001).

Whether or not political campaigns devise subtle racial cues in order to surreptitiously activate the racial views of the electorate is a difficult issue to study. When voters bring their racial attitudes to bear on some voting decisions rather than others, we cannot be sure using observational studies whether this occurred because of specific campaign tactics or some other unrelated event. Experimental manipulation provides perhaps the only way to confidently evaluate this possibility but, as we shall see, even here many questions remain unanswered.

Mendelberg was among the first to examine this question by developing a series of experiments manipulating whether a fictitious gubernatorial candidate’s anti-welfare appeal was racially implicit (i.e., visual references to race but not verbal), explicit (i.e., visual and verbal references to African Americans), or counter-stereotypic (i.e., anti-white, rather than anti-black). Her subjects are drawn from a random sample of New Jersey households. In addition to manipulating racial cues, Mendelberg also manipulated whether participants were told that their

views conformed with or violated society norms on race. She finds that concern with violating norms prevents explicit messages from activating anti-black attitudes with one caveat: racially liberal subjects who are unconcerned when told that their views violate the norm are much more likely to support the racially conservative candidate. This result suggests that further research needs to be done exploring exactly how concern for norms moderates the effects of racial priming.

In spite of this support for the racial priming hypothesis, confirmed and replicated by Valentino, Hutchings, and White (2002), a debate has emerged in the literature regarding the influence of implicit and explicit racial appeals. Employing an experimental design similar to Mendelberg's, but with a nationally representative sample treated over the Internet (N=6,300), Huber and Lapinski (2006) find that implicit appeals are not more effective than explicit messages in priming racial attitudes. In Mendelberg's (2008) response, she questions whether the treatment was delivered successfully and argues that, because subjects' racial predispositions were measured just prior to exposure to the experimental treatments, any differences in priming effects may have been neutralized. Huber and Lapinski (2008) reject these criticisms. More important for our purposes, however, is that the racial priming literature is vulnerable to the charge that measuring racial attitudes immediately prior to or following the treatment may affect the results. When the measurement occurs prior to treatment, researchers run the risk of dampening any priming effect due to "contamination" of the control group. However, when measured after the treatment, the distribution of racial attitudes may be influenced by the manipulation (Linz 2009). Thus, it is not so much that liberals and conservatives are sorting themselves out more appropriately as a consequence of exposure to an implicit racial appeal.

Rather, the candidate preferences may remain firm and the racial attitudes may be changing.

Future work in this area should try to measure racial attitudes some considerable time prior to the treatment in order to avoid this potential problem.

3. Conclusion

In sum, though experiments have contributed to our knowledge of the role that prejudice may play in shaping policy preferences and candidate support, a number of questions remain unanswered. Experiments clearly do provide advantages as they address some of the weaknesses of observational studies. The main weakness of observational research is its inability to control for all possible confounds, and the main strength of experiments is their ability to do just that. Experiments, however, are not a panacea and consequently bring their own set of limitations. The limitations of the work reviewed here include (at least occasionally) an over-reliance on convenience samples that consequently undermine external validity, lack of realism in the treatments, and imprecision in the experimental design, thereby clouding the inferences that can be drawn. Imprecision in the experimental design is especially troubling, given that the main advantage of experiments is the ability of experimenter to eliminate alternative potential confounds. Some of these concerns will always be difficult to address as they represent inherent problems with the use of experiments in the social sciences. However, by devoting our attention to improvement in design we can minimize some of the most glaring weaknesses of this valuable tool.

References

Allport, Gordon W. 1979. *The Nature of Prejudice: 25th Anniversary Edition*. New York: Basic Books.

- Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American Journal of Political Science* 52: 959-78.
- Citrin, Jack, Donald Philip Green, and David O. Sears. 1990. "White Reactions to Black Candidates: When Does Race Matter?" *Public Opinion Quarterly* 54: 74-96.
- DuBois, William Edward Burhardt. 1899. *The Philadelphia Negro*. Philadelphia: University of Pennsylvania Press.
- Feldman, Stanley, and Leonie Huddy. 2005. "Racial Resentment and White Opposition to Race Conscious Programs: Principles of Prejudice?" *American Journal of Political Science* 49: 168-83.
- Gibson, James L. 2009. *Overcoming Historical Injustices: Land Reconciliation in South Africa*. New York: Cambridge University Press.
- Gilens, Martin. 1996. "'Race Coding' and Opposition to Welfare." *American Political Science Review* 90: 593-604.
- Gilliam, Franklin D., and Shanto Iyengar. 2000. "Prime Suspects: The Influence of Local Television News on the Viewing Public." *American Political Science Review* 44: 560-73.
- Highton, Benjamin. 2004. "White Voters and African American Candidates for Congress." *Political Behavior* 26: 1-25.
- Huber, Greg A., and John Lapinski. 2006. "The 'Race Card' Revisited: Assessing Racial Priming in Policy Contests." *American Journal of Political Science* 48: 375-401.
- Huber, Greg A., and John Lapinski. 2008. "Testing the Implicit-Explicit Model of Racialized Political Communication." *Perspectives on Politics* 6:125-34.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.
- Kinder, Donald R., and Lynn M. Sanders. *Divided By Color*. Chicago: University of Chicago Press.
- Kinder, Donald R., and David O. Sears. 1981. "Prejudice and Politics: Symbolic Racism versus Racial Threats to the Good Life." *Journal of Personality and Social Psychology* 40: 414-31.
- Linz, Gabriel. 2009. "Learning and Opinion Change, Not Priming: Reconsidering the Evidence for the Priming Hypothesis." *American Journal of Political Science* 53: 821-837.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm*

- of Equality*. Princeton, NJ: Princeton University Press.
- Mendelberg, Tali. 2008. "Racial Priming Revised." *Perspectives on Politics* 6: 109-123.
- Moskowitz, David, and Patrick Stroh. 1994. "Psychological Sources of Electoral Racism." *Political Psychology* 15: 307-29.
- Nelson, Thomas E., and Donald R. Kinder. 1996. "Issue Frames and Group-Centrism in American Public Opinion." *Journal of Politics* 58: 1055-78.
- Peffley, Mark and Jon Hurwitz. 2007. "Persuasion and Resistance: Race and the Death Penalty in America." *American Journal of Political Science* 51: 996-1012.
- Reeves, Keith. 1997. *Voting Hopes or Fears? White Voters, Black Candidates, and Racial Politics in America*. New York: Oxford University Press.
- Reyna, Christine, P.J. Henry, William Korfmacher, and Amanda Tucker. 2006. "Examining the Principles in Principled Conservatism." *Journal of Personality and Social Psychology* 90: 109-28.
- Riggle, Ellen D., Victor C. Ottati, Robert S. Wyer, James Kuklinski, and Norbert Schwarz. 1998. "Bases of Political Judgments: The Role of Stereotypic and Nonstereotypic Information." *Political Behavior* 14: 67-87.
- Sears, David O. 1988. "Symbolic Racism." In *Eliminating Racism: Profiles in Controversy*, eds. Phyllis A. Katz, and Dalmas A. Taylor. New York: Plenum.
- Sears, David O., and P. J. Henry. 2003. "The Origins of Symbolic Racism." *Journal of Personality and Social Psychology* 85: 259-75.
- Sears, David O., and P. J. Henry. 2005. "Over Thirty Years Later: A Contemporary Look at Symbolic Racism and its Critics." In *Advances in Experimental Social Psychology*, ed. Mark Zanna. New York: Academic Press.
- Sears, David O., and Donald R. Kinder. 1971. "Racial Tensions and Voting in Los Angeles." In *Los Angeles: Viability and Prospects for Metropolitan Leadership*, ed. Werner Z. Hirsch. New York: Praeger.
- Sidanius, Jim, and Felicia Pratto. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. New York: Cambridge Univ. Press.
- Sigelman, Carol K., Lee Sigelman, Barbara J. Walkosz, and Michael Nitz. 1995. "Black Candidates, White Voters." *American Journal of Political Science* 39: 243-65.
- Sniderman, Paul M., and Edward G. Carmines. 1997. *Reaching Beyond Race*. Harvard: Harvard University Press.

- Sniderman, Paul M., and Philip E. Tetlock. 1986. "Symbolic Racism: Problems of Motive Attribution in Political Analysis." *Journal of Social Issues* 42: 129-50.
- Sniderman, Paul M., Pierangelo Peri, Rui J.P. de Figueiredo, and Thomas Piazza. *The Outsider: Prejudice and Politics in Italy*. Princeton, NJ: Princeton University Press.
- Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: Harvard University Press.
- Terkildsen, Nayda. 1993. "When White Voters Evaluate Black Candidates: The Processing Implications of Candidate Skin Color, Prejudice, and Self-Monitoring." *American Journal of Political Science* 37: 1032-53.
- Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White. 2002. "Cues that Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96: 75-90.
- Walton, Hanes, Jr., Cheryl Miller, and Joseph P. McCormick. 1995. "Race and Political Science: The Dual Traditions of Race Relations and African American Politics." In *Political Science in History*, eds. John Dryzek, James Farr, and Stephen Leonard. Cambridge, UK: Cambridge University Press.

ⁱ We do not discuss implicit prejudice here (see Taber and Lodge's chapter in this volume).

ⁱⁱ This focus on attitudes in the post-Civil Rights Era is in part a practical one since political scientists showed little interest in questions of race or racial bias prior to the late 1960s (Walton, Miller, and McCormick 1995).

ⁱⁱⁱ More recently Sears and Henry (2005) have identified four specific themes associated with the theory of symbolic racism: the belief that racial discrimination against blacks has mostly disappeared; that racial disparities in social and economic outcomes are due to blacks not trying hard enough; that blacks are demanding too much too fast; and that blacks have gotten more than they deserve.

^{iv} Some later indexes, such as the modern racism scale and the racial resentment scale, are designed to capture similar or overlapping concepts (McConahay 1982; Kinder and Sanders 1996). In order to limit confusion, we will focus on the symbolic racism scale but the strengths and weaknesses associated with this theory can also be applied to its intellectual progeny.

^v This emphasis on the importance of framing effects can also be found in the work of Nelson and Kinder (1996) and Kinder and Sanders (1996), as discussed later in this chapter.

^{vi} This question reads as follows: "Irish, Italians, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors."

^{vii} Feldman and Huddy (2005) are quick to point out that, although the symbolic racism scale does not behave like a measure of prejudice for this group, conservatives are less supportive of the scholarship program for blacks. Thus, they find evidence of a racial double standard for this group, with opposition to college scholarships particularly low when the target group is middle-class blacks.

23. Politics from the Perspective of Minority Populations

Dennis Chong and Jane Junnⁱ

Experimental studies of racial and ethnic minorities have focused on the influence of racial considerations in political reasoning, information processing, and political participation. Studies have analyzed the types of messages and frames that prime racial evaluations of issues, the effect of racial arguments on opinions, and the impact of racial cues on political choices. Underlying this research is the premise, developed in observational studies (e.g., Bobo and Gilliam 1990; Tate 1994; Dawson 1994; Lien 2001; Chong and Kim 2006; Barreto 2007), that there are racial and ethnic *differences* in how individuals respond to cues and information. For this reason, almost all studies give special attention to the mediating and moderating influences of racial group identification, a core concept in the study of minority politics.

There are too few studies yet to constitute a research program, but the initial forays have successfully featured the advantages of experimental design and distinct perspectives of minority groups. We review the methodology and findings of these experimental studies to highlight their contributions and limitations and to make several general observations and suggestions about future directions in this field. As we shall see, randomization and control strengthen the internal validity of causal inferences drawn in experiments but, of equal importance, the interpretation and significance of results depends on additional considerations including the measurement of variables, the external validity of the experiment, and the theoretical coherence of the research design.

1. Racial Priming

Virtually all racial priming research has examined the public opinion of whites (e.g., Gilens 1999; Gilliam and Iyengar 2000; Mendelberg 2001) and been modeled on prior studies of media priming of voter evaluations (Iyengar and Kinder 1987). The theory of racial priming is that attitudes toward candidates and policies can be manipulated by framing messages to increase the weight of racial considerations, especially prejudice. Among whites, “implicit” racial messages that indirectly address race – using images or code words — are hypothesized to prime racial considerations more effectively than racially explicit messages that violate norms of equality. Explicit statements potentially are less effective because they can raise egalitarian concerns that suppress open expression of prejudice. By contrast, implicit appeals smuggle in racial primes that activate prejudice and turn opinion against a policy or candidate without triggering concerns about equality. This is the racial priming theory as applied to whites (Mendelberg 2001; cf. Huber and Lapinski 2008).

Priming racial considerations among blacks

The dynamics of racial appeals are likely to be different among blacks because racial messages aimed at blacks often promote group interests without raising conflicting considerations between race and equality. Therefore, in contrast to white respondents, blacks should be *more* likely to evaluate an issue using racial considerations when primed with either explicit or implicit racial cues.

To test the idea that explicit and implicit messages affect blacks and whites differently, White (2007) designed an experiment using news articles to manipulate the verbal framing of two issues: the Iraq War and social welfare policy. The sample included black and white college students and adults not attending college. Participants were randomly assigned to one of the

treatments or to a control group that read an unrelated story. For each issue, one of the frames explicitly invoked black group interests to justify the position taken in the article; a second frame included cues that implicitly referred to blacks; and a third frame included nonracial reasons. In the welfare experiment, there were two implicit frame conditions in which the issue was associated with either “inner city Americans” or to “poor Americans” on the assumption these references would stimulate racial resentment among whites and group interests for blacks. Similarly, the implicit racial cue in the Iraq War experiment referred to how the war drained money from social spending.

As in prior studies of white opinion, the experiments confirmed that resentment of blacks among whites was strongly related to support for the war and opposition to welfare spending only in the implicit condition. Among blacks, racial identification was strongly related to support in the explicit condition on both issues but, surprisingly, was unrelated in the implicit condition. Thus, explicitness of the cue has differential effects among blacks and whites, roughly as predicted by the theory.

There are some oddities however. Racial resentment among whites significantly reduces support for the war in the racially explicit condition, when the unequal burden of the war on blacks is emphasized. The racial priming theory predicts that explicit statements should weaken the relationship between resentment and support for the war, but not reverse its direction. In the welfare experiment, egalitarian values are also strongly primed among whites in the implicit racial condition, in addition to out-group resentments. This means egalitarian considerations potentially counteract racial resentment even in the implicit case, contrary to expectations. Finally, in contrast to past research (Kinder and Sanders 1996, Kinder and Winter 2001;), blacks

do not respond racially either to the implicit message that war spending reduces social spending or to the implicit cues used to describe the welfare issue.

The anomalies of an experimental study can sometimes yield as much theoretical and methodological insight as the confirmatory findings. In this case, anomalies force us to reconsider the appropriate test of the priming hypothesis. A possible explanation for the weak effect of implicit cues among blacks is that the treatment affects the overall *level* of support for policies, in addition to the strength of the *relationship* between racial predispositions and policy preferences. A flat slope coefficient between racial identification and policy positions does not eliminate the possibility that levels of support or opposition – reflected in the intercept term – change among both strong and weak identifiers in response to the treatment. The priming hypothesis therefore requires an examination of both intercepts and slopes.

Finally, the imprecise definition and operationalization of explicit and implicit cues raises measurement issues. In the welfare experiment, two implicit cues referring to “inner city Americans” and “poor Americans” were incorporated in arguments made in support of welfare programs. Likewise, the implicit racial condition in the Iraq War experiment refers to the war taking attention away from “domestic issues,” including “poverty,” “layoffs,” “inadequate healthcare,” and “lack of affordable housing.” Without explicitly mentioning blacks, both treatments refer to issues that are associated with blacks in the minds of many Americans. However, the “nonracial” condition in the welfare experiment also refers to “poor” Americans or “working” Americans losing food stamps, Medicaid, and health care, and falling into poverty, which are the same kinds of domestic policy references used in the implicit conditions in the two experiments. As we will elaborate shortly, the imprecise definition of explicit and implicit cues

raises general issues of measurement and pre-testing of treatments that are central to experimentation.

The media's crime beat

A second priming study worth exploring in detail for the substantive contributions and methodological issues it raises is Gilliam and Iyengar's (2000) study of the influence of local crime reporting in the Los Angeles media. Whereas White's study investigated how priming affects the dimensions or considerations people use to evaluate an issue, Gilliam and Iyengar focus on attitude change in response to news stories that stimulate racial considerations underlying those attitudes.

Gilliam and Iyengar hypothesize that the typical crime script used in local television reporting (especially its racial bias against blacks) has had a corrosive effect on viewers' attitudes toward the causes of crime, law enforcement policies, and racial attitudes. They designed an experiment in which participants recruited from the Los Angeles area were randomly assigned to one of four conditions. In the control condition, participants watched a news video that did not include a crime story. In the three other conditions, participants viewed a crime story in which the race of the alleged perpetrator was manipulated. In one of the crime stories, there was no description of the murder suspect. In the other two conditions, digital technology was used to change the race (black or white) of the suspect shown in a photograph.

Gilliam and Iyengar found that whites who were exposed to a crime story (regardless of the race of the suspect) tended to be more likely to give dispositional explanations of crime, prefer harsh penalties, and express racially prejudiced attitudes. Black respondents, in contrast, either were unmoved by the treatments or they were moved in the opposite direction as whites,

toward less punitive and prejudiced attitudes. The limited variance (within racial groups) across treatments is partially explained by one of the more fascinating and disconcerting findings of the study. A large percentage (sixty-three percent) of the participants who viewed the video that did *not* mention a suspect nonetheless recalled seeing a suspect, and most of them (seventy percent) remembered seeing a *black* suspect. This suggests that the strong associations between crime and race in people's minds led participants to fill in missing information using their stereotypes. In effect, the "no suspect" condition served as an implicit racial cue for many participants, reducing the contrast between the "black suspect" and "no suspect" conditions. This finding reinforces the need to pre-test stimuli to check if treatments (and nontreatments) are working as desired.

Gilliam and Iyengar also analyzed Los Angeles County survey data to show that frequent viewers of local news were more likely to express punitive views toward criminals and to subscribe to both overt and subtle forms of racism. This corroboration between the observational and experimental data bolsters the external validity of the experimental effects. But one wonders what impact can be expected from an experimental treatment that is a miniscule fraction of the total exposure to crime stories that participants received prior to joining the experiment. Before reporting their experimental results, Gilliam and Iyengar themselves caution readers that "our manipulation is extremely subtle. The racial cue, for example, is operationalized as a five-second exposure to a mug shot in a ten-minute local news presentation. Consequently we have modest expectations about the impact of any given coefficient" (567).

Yet the results of their treatment prove to be impressively large. For example, exposure to the treatment featuring a black suspect increases scores on the new racism scale by twelve percentage points. Compare that amount to the difference between survey respondents who

hardly ever watched the local news and those who watched the news on a daily basis: the most frequent viewers scored twenty-eight percentage points higher on the new racism scale. It is puzzling how a single exposure to a subtle manipulation can produce an effect that is almost fifty percent of the effect of regular news watching. Perhaps the decay of effects is rapid or the magnitude of the experimental treatment effect varies across participants depending on their pre-treatment viewing habits. Both the experiment and survey indicate the style of local television coverage of crime in Los Angeles has had a detrimental effect on viewers' attitudes toward race and crime, but in order to reconcile the results of the two studies, we need more evidence of how viewers' attitudes are shaped over time when they are chronically primed (with variable frequency) by media exposure.

2. Attitude Change

Important studies by Bobo and Johnson (2004) and Hurwitz and Peffley (2005, 2007) employ survey experimental methods on national samples to study the malleability of black and white attitudes under different framing conditions. A comparison of the results from these studies highlights the variable effects of similar experimental treatments. Bobo and Johnson hypothesize that because blacks are more likely to believe that the criminal justice system is racially biased, they are more likely to be influenced by frames accentuating bias in the system when they are asked their opinion of the death penalty and other sentencing practices. For each survey experiment, respondents were randomly assigned to receive one of several framed versions of a question about the criminal justice system (the treatment groups) or an unframed question (the control group).

Most of the tests revealed surprisingly little attitude change among either blacks or whites in response to frames emphasizing racial biases on death row, racial disparities in the commission of crimes, and wrongful convictions. The only frame that made a slight difference emphasized the greater likelihood that a killer of a white person would receive the death penalty than a killer of a black person. This manipulation significantly lowered support for the death penalty among blacks but not among whites (although the percentage shifts are modest).

Attitudes toward drug offenses proved to be more malleable and responsive, specifically to frames emphasizing racial bias in sentencing. Attempts to change views of capital punishment may yield meager results, but efforts to reframe certain policies associated with the war on drugs may have substantial effects on opinion.

Peffley and Hurwitz (2007) also test whether capital punishment attitudes are malleable among blacks and whites in response to arguments about racial biases in sentencing and the danger of executing innocent people. In contrast to Bobo and Johnson, they find that both arguments reduce support for the death penalty among blacks. But the most shocking result is the racial bias argument causes support to increase significantly among whites. Peffley and Hurwitz explain that the racial bias argument increases support for the death penalty among prejudiced individuals by priming their racial attitudes. This priming effect is made more surprising if we consider the racial bias argument to be an explicit racial argument that might alert white respondents to guard against expressing prejudice.

Peffley and Hurwitz do not reconcile their findings with the contrary results in Bobo and Johnson's survey experiment beyond speculating that the racial bias frames in the other survey may have been harder to comprehend. Among other possible explanations is that Bobo and

Johnson's use of an Internet sample overrepresented individuals with strong prior opinions about the death penalty who were inoculated against framing manipulations. Bobo and Johnson, however, conclude that the frames are resisted irrespective of the strength of prior opinions because they find no differences in the magnitude of framing effects across educational levels.ⁱⁱ

Two other anomalies in the Peffley and Hurwitz study are worth mentioning briefly, as we shall return to them in the general discussion of this body of research. First, "consistent with our expectations, blacks apparently need no explicit prompting to view questions about the death penalty as a racial issue. Their support for the death penalty, regardless of how the issue is framed, is affected substantially by their belief about the causes of black crime and punishment" (1005). Although Peffley and Hurwitz anticipated this result, it might be viewed as being somewhat surprising in light of White's (2007) demonstration that racial attitudes are related to public policies only when they are explicitly framed in racial terms. Second, among both black and white respondents, racial arguments do not increase the accessibility of other racial attitudes, such as stereotypic beliefs about blacks.

Framing affirmative action decisions

Clawson, Kegler, and Waltenburg's (2003) study of the framing of affirmative action illustrates the sensitivity of results to the sample of experimental participants. They used a two-by-two design in which participants received one of four combinations of frames embedded in a media story about a recent Supreme Court decision limiting affirmative action. The decision was described either as a decision barring preferential treatment for any group or as a major blow to affirmative action and social justice; in each media story, there was either a critical comment

about Justice Clarence Thomas, or no comment about Justice Thomas's conservative vote on the issue.

The participants were 146 white and black students from a large Midwestern university. Comparisons of the sample to the NES and NBES samples revealed, as expected, that both black and white participants were younger, better educated, and wealthier than blacks and whites in the national sample. Black participants were also much more interested in politics than the national black sample.

The dominant finding for black participants is they (in contrast to white participants) have firm positions on affirmative action regardless of how a recent conservative court ruling is framed. Among blacks, only their racial attitude toward blacks (measured by racial resentment items) and gender predicted their attitude toward affirmative action; the frames were irrelevant.

The insignificance of framing in this experiment illustrates the difficulty of generalizing beyond the experimental laboratory participants to the general population. Affirmative action is likely to be a more salient issue to African Americans, and attitudes on salient issues are likely to be stronger and more resistant to persuasion. Whether this is true for only a small subset of the black population or for most blacks can only be settled with a more representative sample.

3. Racial Cues and Heuristics

The next set of studies we review involves experimental tests of the persuasiveness of different sources and messages. These studies focus on minority responses to consumer and health messages, but they are relevant for our purposes because their findings on how racial minorities use racial cues in processing information can be extrapolated to political choices.

In the basic experimental design, participants (who vary by race and ethnicity) are randomly assigned to receive a message from one of several sources that vary by race or ethnicity and expertise. The primary hypothesis is that sources that share the minority participant's race or ethnicity will be evaluated more highly along with their message. A second hypothesis is that the impact of shared race or ethnicity will be moderated by the strength of the participant's racial identity. Finally, these studies test whether white participants favor white sources and respond negatively to minority sources.

Appiah (2002) found that black audiences recalled more information delivered by a black source than a white source in a videotaped message. This study also found that white participants' recall of information about individuals on a videotape was unaffected by the race of those individuals. White subjects' evaluation of sources was based on social (occupation, physical appearance, social status) rather than racial features perhaps because race is less salient to individuals in the majority.

Wang and Arpan (2008) designed an experiment to study how race, expertise, and group identification affected black and white audience's evaluations of a public service announcement (PSA). The participants for the experiments were black and white undergraduate students recruited from a university in the southeastern U.S. and from a historically black college in the same city.

Black respondents rated a black source more highly than a white source and reacted more positively toward the PSA when it was delivered by a black source. But the effect of the source on blacks and whites was again asymmetrical. Race did not bias white respondents' evaluations in the same way; instead whites were more affected in their evaluation of the message by the

expertise (physician or nonphysician) of the source than were blacks. Contrary to expectations, strength of racial identity did not moderate the effect of the race of the source.

The favoritism that blacks show toward a black source in a public health message is also demonstrated in an experiment by Herek, Gillis, and Glunt (1998) on the factors influencing evaluation of AIDS messages presented in a video. Blacks evaluated a black announcer as more attractive and credible than a white announcer, but these in-group biases were not manifest among whites. Blacks also favored videos that were built around culturally specific messages, in contrast to multicultural messages. The manipulations in this experiment affected proximate evaluations of the announcer and message, but did not affect attitudes, beliefs, and behavioral intentions regarding AIDS.

Whittler and Spira's (2002) study of consumer evaluations hypothesizes that source characteristics will serve as peripheral cues, but can also motivate cognitive elaboration of messages. Studies have shown that whites sometimes focus more heavily on the content of the message when the source is black (White and Harkins 1994; Petty, Fleming, and White 1999). The sample consisted of 160 black adults from a southeastern city assigned to a 2x2 experimental design. Participants received a strong or weak argument from either a black or white speaker advertising a garment bag.

The evidence in the Whittler and Spira study is mixed: Participants overlooked the quality of arguments and rated the product and advertising more favorably if it was promoted by a black source, but this bias was evident only among participants who identified strongly with black culture. Identification with the black source appeared to generate more thought about the

speaker and the advertisement, but because additional thinking was also biased by identification, greater thought did not lead to discrimination between strong and weak arguments.

Forehand and Deshpande (2001) argue that group targeted ads will be most effective on audiences that have first been ethnically primed. Same ethnicity sources or group-targeted messages may not have a significant impact on audiences unless ethnic self-awareness is initially primed to make the audience more receptive to the source.

The subjects in the Forehand and Deshpande study were Asian American and white students from a west coast university. Advertisements were sandwiched between news segments on video, with ethnic primes preceding advertisements aimed at the ethnic group. Similar results were obtained in both experiments. Exposure to the ethnic prime caused members of the target audience to respond more favorably to the ethnic ad. But the magnitude of the effect of the ethnic prime was not magnified by strong ethnic identification, so the expectation of an interaction with enduring identifications was not met. This is a surprising result because even though strong identifiers are not continuously aware of their ethnicity, their ethnicity should be more chronically accessible, and therefore we would expect strong identifiers to be most sensitive to the ethnic prime. Exposure to the ethnic prime among members of the nontarget market (whites in the experiment) resulted in less favorable responses, but the magnitudes were statistically insignificant. Once again, it does not appear that an ethnic prime has a negative effect on individuals who do not share the same ethnicity.

Extensions to vote choice

An obvious extrapolation from these studies is to examine how variation in the race or ethnicity of a politician influences political evaluations and choices. Kuklinski and Hurley

(1996) conducted one of the few experimental studies in political science along these lines.

African Americans recruited from the Chicago metropolitan region were randomly assigned to one of four treatments or to a control group. Each treatment presented a common statement about the need for self-reliance among African Americans, but the statement was attributed to a different political figure in each of the four conditions: George Bush, Clarence Thomas, Ted Kennedy, or Jesse Jackson. If the statement was attributed to Bush or Kennedy, participants were more likely to disagree with it, but if the observation originated from Jackson or Thomas, they were significantly more likely to agree. As in the case of the aforementioned Whittler and Spira study, some respondents relied entirely on the (peripheral) racial cue to form their judgment, but even those respondents who gave more attention to the substance of the message construed it in light of the source.

Surprisingly, we did not discover any experimental research using this basic design to analyze the effect of race and ethnicity on minority voter choice. An innovative experimental study by Terkildsen (1993) examined the effects of varying the race (black or white) and features (light or dark skin tone) of candidates, but only on the voting preferences of white respondents (who evaluated the white candidate significantly more positively than either of the two black candidates.).

Abrajano, Nagler, and Alvarez (2005) took advantage of an unusual opportunity in Los Angeles County to disentangle ethnicity and issue distance as factors in voting. In this natural experiment using survey data, Abrajano et al. analyzed the electoral choices of voters in two open city races involving Latino candidates running against white candidates. In the mayoral race, the white candidate was more conservative than the Latino candidate, but the white

candidate was the more liberal candidate in the city attorney election. They found Latino voters were more affected by the candidates' ethnicity and much less affected by their issue positions than were white voters.

4. Political Mobilization: Get Out the Vote

Aside from laboratory and survey research on persuasion and information processing, the study of political mobilization is the other area in which there has been sustained experimental research on minority groups.ⁱⁱⁱ Field research on the political mobilization of minorities comes with special challenges, as it requires investigators to go beyond standard methodologies for data collection in the midst of electoral campaigns. Researchers must take care to locate the target populations for study, provide multilingual questionnaires and interviewers in some cases, and design valid and reliable treatments appropriate to minority subjects.

Garcia, Bedolla, and Michelson (2009) report on a field experimental study of a massive effort to mobilize voters through direct mail and telephone calls in California during primary and general election phases of the 2006 election. The content of the direct mail included a get out the vote (GOTV) message but varied in terms of procedural information such as the voter's polling place and a photo included in the mailer that was adjusted "to be appropriate to each national-origin group" (9). The authors found no significant impact of direct mail, and a positive effect of a phone call on voting turnout among the Asian American subjects contacted (with considerable variance across groups classified by national origin). Considering the extremely low base rate of voting in the target population, the treatment had a large proportional impact. The authors' conclusions from this set of experiments and other GOTV studies in California and elsewhere point to the significance of a personal invitation to participate. At the same time, however, they

admitted, “we do not have a well-defined theoretical understanding of why an in-person invitation would be so effective, or why it could counteract the negative effect of low voter resources” (271).

The results from the Garcia Bedolla and Michelson field experiments are consistent with earlier studies of Latinos, Asian Americans, and African Americans that have shown direct mail to be ineffective and personal contact to be effective interventions with minority voters. In a large-scale national field experiment with African American voters during the 2000 election, Green (2004) found no significant effects on turnout with a mailing, and small but statistically significant effects from a telephone call. In another large field experiment during the 2002 election, Ramirez (2005) analyzed results from attempted contact with nearly a half-million Latinos by the National Association of Latino Elected and Appointed Officials (NALEO). Neither the robo-calling nor the direct mail had a discernible and reliable influence on voter turnout among Latinos, but the live telephone calls did have a positive effect on mobilizing voters.

Trivedi (2005) attempted to discern whether distinctive appeals to ethnic group solidarity among Indian Americans would increase voter turnout. Despite three alternative framings with racial and ethnic cues, there were no significant effects on voter turnout of any of the three groups that received the mailing. Wong’s (2005) field experiment during the 2002 election included a postcard mailing or a phone call for randomly assigned Asian American registered voters in Los Angeles County who resided in high-density Asian American areas. In contrast to other studies that show no effect from direct mail, Wong found positive effects for both mobilization stimuli on Asian American voter turnout.

Finally, Michelson (2005) reports on a series of field experiments with Latino subjects in central California. Her results show that Latino Democrats voted at a higher rate when contacted by a coethnic canvasser. Michelson suggests possible social and cognitive mechanisms that explain why “coethnic” contacts may stimulate higher levels of participation: “Latino voters are more likely to be receptive to appeals to participate when those appeals are made by coethnics and copartisans. . . . In other words, if the messenger somehow is able to establish a common bond with the voter – either through shared ethnicity or through shared partisanship – then the voter is more likely to hear and be affected by the mobilization effort” (98-99).

Taken together, none of the field experiments shows strong or consistent positive effects from direct mailings, regardless of content, format, or presentation of the information in a language assumed to be most familiar to the subject. Second, personal contact with a live person in a telephone call has a modest absolute effect, and potentially a greater effect if the contact occurs in person, especially if that individual is a “coethnic” canvasser. Third, effective methods for mobilizing specifically minority voters are essentially the same as those found to work on majority group populations. Personal contacts, unscripted communications, and face-to-face meetings provide more reliable boosts to turnout than do more automated, remote techniques (Gerber and Green 2000).

To the extent that field experimental research on the mobilization of minority voters is distinct from studies of white voters, it is due to the nascent hypothesis that identifications and social relationships based on race and ethnicity ought to moderate the impact of mobilization treatments. This intuition or hunch lies behind experimental manipulations that try to stimulate group identification using racial or ethnic themes or imagery in communications. A Latino voter,

for example, is told his vote will help to empower the ethnic community, or an Asian American is shown a photograph of Asian Americans voting out of duty as U.S. citizens.

The prediction that treatments will elevate racial and ethnic awareness and thereby increase one's motivation to vote is based on several assumptions about both the chronic accessibility of social identifications and the durability of treatment effects. Some observations drawn earlier from our review of political psychology research should lower our expectations about the impact of such efforts to manipulate racial identities. In communications experiments on racial priming, participants are typically unusually attentive to experimental treatments; volunteers who agree to participate are eager to cooperate with the experimenter's instructions and pay close attention to any materials they are asked to read or view. This combination of higher motivation and capacity means the laboratory subject receives what amounts to an especially large dosage of the treatment. Finally, if we assume the treatment to be fast acting but also fast fading – like a sugar rush – then its effects will only be detected if we measure them soon after it is administered.

The typical GOTV field experiment deviates from these conditions in each instance. Attention to the treatment and interest in politics are low (indeed, sometimes the participants are selected on this basis); comprehension may be impaired because of language difficulties or inability. Furthermore, the ad hoc design of the ethnic cues makes them equivalent to an untested drug whose effects have not heretofore been demonstrated in pre-testing. Unlike laboratory experiments in which effects are measured promptly, the GOTV treatments are expected to influence behavior (not simply attitudes) days or weeks later, so it is perhaps not surprising they have proved to be anemic stimulants.

5. General Observations and Future Directions

In our review, we considered not only whether a treatment had a significant effect in a particular study, but also the interpretation of those effects (does the explanation accurately reflect what occurred in the experiment?) and the consistency of experimental effects across related studies (are the results of different studies consistent with a common theory?). We elaborate in this section by discussing how to strengthen the internal and external validity of experiments through improvements in measurement, design, and theory building.

Generalizing from a single study

The variability of results across studies recommends careful extrapolation from a single study. For example, the failure of a framed argument to move opinion may be explained by the imperviousness of participants on the issue or to the weakness of the frames. Bobo and Johnson (2004) chose the former interpretation in arguing that “with respect to the death penalty, our results point in the direction of the relative fixity of opinion” (170). However, they also concluded from their survey experiment that reframing the war on drugs as “racially biased” might significantly reduce support for harsh sentencing practices among both blacks and whites.

Peffley and Hurwitz’s (2007) contrary finding that whites increase their support for the death penalty when told it is racially biased led them to conclude that “direct claims that the policy discriminates against African Americans are likely to create a backlash among whites who see no real discrimination in the criminal justice system” (1009). Whether the frames used in the two studies varied or the participants varied in the strength of their existing attitudes cannot be resolved without systematically comparing the frames and prior attitudes of the participants; the same ambiguity between the strength of arguments versus prior opinions hovers over the null

findings in Clawson et al.'s (2003) study of affirmative action attitudes among blacks. At a minimum, replication of results using both identical and varied treatments and different samples of respondents would increase our confidence in either the stability or malleability of opinion on these issues.

The importance of pre-testing measures

The internal validity of a study depends on reliable and valid measures. A general lesson drawn from the experiments on persuasion and information processing, as well as the GOTV studies, is the need to pre-test stimuli to establish that treatments have the characteristics attributed to them. These pre-tests will be sample dependent and should be administered to individuals who do not participate in the main experiment.

Our review provided several instances where progress on the effects of racial priming and framing would be aided by more clearly defined measures of explicit and implicit messages. Mendelberg (2001) classifies visual racial cues as implicit messages and direct verbal references to race as explicit messages. White (2007) distinguishes between explicit cues that mention race and implicit verbal cues that allude to issues and terms that are commonly associated with race; but a domestic issue cue that is assumed to be an implicit racial cue in one experiment is defined as a nonracial cue in another experiment. Bobo and Johnson (2004) and Peffley and Hurwitz (2007) introduce frames that refer directly to the disproportionate treatment of blacks under the criminal justice system, but do not interpret their respondents' reactions to these frames using the implicit-explicit theoretical framework.

This conceptual task is made more difficult because the dividing line between explicit and implicit varies across audiences. Different audiences, owing to differences in past learning

experiences, will draw different connotations from the same message. Certain messages are so blunt that they obviously draw attention to racial considerations. Other messages allude indirectly to race and can be interpreted in racial terms only by those who are able to infer racial elements from ostensibly nonracial words or symbols because of common knowledge that such symbols or words connote racial ideas. Of course, if the common knowledge is so widespread as to be unambiguous, then even the implicit message becomes explicit to everyone in the know. Thus there is supposedly a sweet spot of ambiguity wherein lies messages that cause people to think in racial terms either without their knowing it or without their having to admit it because there is a plausible nonracial interpretation of the message.

We do not have a ready solution for distinguishing between explicit and implicit messages. One possibility is that the location of a message on a continuum ranging from more to less explicit will correspond to the balance of racial and nonracial interpretations and thoughts that are spontaneously mentioned when interpreting the message (Feldman and Zaller 1992). However, individuals who are careful to monitor their public behavior may not candidly report their spontaneous thoughts, especially if those thoughts are racial in nature.

A covert method of eliciting the same information uses subliminal exposure to a given message followed by tests of reaction times to racial and nonracial stimuli. More explicit messages may be expected to produce quicker reactions to racial stimuli than implicit messages. Lodge and Taber (in press) provide a convincing demonstration of how implicit testing of competing theoretical positions can shed light in the debate over the rationales underlying support for symbolic racial political values (for a review, see Sears, Sidanius, and Bobo 2000). A racial issue (e.g., affirmative action) is used as a prime (presented so briefly on the screen that it

registers only subconsciously), and the words automatically activated by this issue (determined by the speed of recognizing them) are interpreted to be the considerations raised by an issue. Coactivation of concepts is said to represent habits of thought reflecting how individuals routinely think about the issue. Using this method, Lodge and Taber show that, among supporters of affirmative action, ideology and racial considerations were activated – i.e., facilitated by the affirmative action issue prime; but among opponents, only racial words were activated – e.g., “gang” and “afro”. Therefore, it appears the liberal position on the issue drew on more principled considerations while the conservative position rested on racial considerations.

Adding realism through competition and over-time designs

In the framing studies we examined, participants were exposed to arguments on only one side of the issue under investigation. In contrast, competition between frames and arguments reflects the reality of political debate in democratic systems. Multiple frames increase the accessibility of available considerations and competition between frames can motivate more careful deliberation among alternatives (Chong and Druckman 2007). Framing effects produced by a one-sided frame often are not sustained in competitive environments (Sniderman and Theriault 2004; Chong and Druckman 2007; cf. Chong and Druckman 2008).

The theory of implicit and explicit messages, in particular, would benefit from a competitive experimental design because racial priming describes an inherently dynamic process in which strategic political messages are transmitted and countered and subject to claims and counterclaims about the meaning of the message and the intent of the messenger. The essence of an intrinsic racial message is that it can be defended against attacks that it *is* a racial (and perhaps racist) message. How the originator of the message parries these attacks undoubtedly has much

to do with the success of the original strategy. Despite the theory, all testing has been essentially static and limited to a one-time administration of one-sided information.

Another way to increase the realism of designs is to examine communications processes over time. The persuasion and information processing studies we reviewed were one-shot studies in which the magnitudes of communication effects were measured immediately following exposure to the treatment. The design of these laboratory experiments contrasts with conditions in the real world, where individuals typically receive streams of messages and act upon them at the end of a campaign. The interpretation of a one-shot experiment therefore should take account of the previous experiences of participants and the subsequent durability of any observed effects. A treatment may have a larger impact if it is received early rather than late in a sequence of communications, because the effect of a late treatment may be dulled by past messages. In addition, we want to measure the durability of effects in the post-treatment period. A significant treatment effect may decay rapidly either on its own accord or under the pressure of competing messages. Ultimately, the effect of a treatment will be time dependent.

The importance of taking account of participants' pre-treatment experiences is evident in the study of death penalty frames. The effectiveness of arguments against the death penalty likely depends on whether participants have previously heard and factored these arguments into their attitudes on the issue (Gaines, Kuklinski, and Quirk 2007). A paradox in this regard is the sizable effects generated by Gilliam and Iyengar's (2000) "subtle" media crime news intervention, which led us to wonder why experimental participants who have been "pre-treated" with everyday exposure to crime news coverage would nevertheless remain highly sensitive to the experimental treatment. If one-shot experimental exposure to crime stories produces an effect

that is fifty percent of the effect of chronic real-world exposure, how much of this short-term effect decays and how much endures in the long-term effect? To disentangle these processes of learning and decay of opinion, we need to move from one-shot experimental designs to panel experiments, in which we measure attitude change in response to a series of exposures to treatments over time (Chong and Druckman 2008). A panel design would allow us to determine how the size and durability of effects are moderated by past experiences, the passage of time, and subsequent exposure to competing messages.

Integrating theory and design

Although experiments are well suited to testing whether an arbitrary treatment has an impact on an outcome variable (in the absence of a theoretically derived hypothesis), ultimately such a theory is required to explain and bring coherence to disparate results. Otherwise, experiments are at risk of being a series of one-off exercises.

The theoretical challenge of specifying the meaning and measurement of racial group identification and its relationship to political behavior and attitudes is one of the most significant hurdles in research on the political psychology and behavior of minorities. GOTV studies have tested whether stimulating racial identification can increase turnout in elections, but this research has not been guided explicitly by a theory of the mechanisms that activate racial identification nor of the factors that convert identification to action. As noted, the failure of efforts to motivate voter turnout by using racial or ethnic appeals can be explained in large part by the superficial nature of the treatments. GOTV experiments might draw on the results of past survey research on racial identification, which have shown the connection between group identification and political participation is mediated by perceptions of group status, discontent with the status quo, and

beliefs about the origins of group problems and efficacy of group action (Shingles 1981; Miller et al. 1981; Marschall 2001; Chong and Rogers 2005). This model of racial identification suggests experimental manipulation of racial awareness by itself will have little effect on political participation without the constellation of intervening cognitive factors that motivate individuals to participate.

A fruitful theory of racial identification provides testable hypotheses of the conditions in which voters can be more easily mobilized on the basis of their race or ethnicity. Minority voters should be more responsive to racial cues when electoral candidates and issues place group interests at stake and collective action is an effective means to obtain group goals (Chong 2000). The selection of future sites of GOTV field experiments therefore might exploit political contexts in which minority voters are likely to be especially susceptible to messages and contacts that prime their racial identification.^{iv} Contact by coethnic organizers, which has proved effective in past studies, may have even greater impact in these circumstances, especially if campaign workers are drawn from the voter's social network. When political opportunities for gain present themselves, monitoring within the group -- verified to be one of the most powerful influences on voting in GOTV research (Gerber, Green, and Larimer 2008) -- would apply added social pressure on individuals to contribute to public goods.

The importance of theoretical development to experimental design applies to all areas of research we have covered. Although we grouped the persuasion, priming, and framing research as a "set" of studies that address common issues, they are not unified theoretically. This inhibits development of a research program in which there is consensus around certain theoretical concepts and processes that serve as a framework for designing new experimental studies.

All of the communications research we have discussed here can be interpreted in terms of existing dual process theories of information processing (Fazio and Olson 2003; Petty and Cacioppo 1986; Petty and Wegener 1999). According to this theory, individuals who have little motivation or time to process information will tend to rely on economical short cuts or heuristic rules to evaluate messages. Attitude change along this “peripheral” route, however, can be transitory, leading to short-term reversion to past beliefs.

Conversely, individuals who are motivated and able to think more deeply about a subject, because of incentives or predispositions to do so, will process information along a “central” route by giving closer attention to the quality of arguments in the message. Individuals who hold strong priors on the subject are more likely to ignore or resist contrary information and to adhere to their existing attitudes. In some cases, the cognitive effort they expend will end up bolstering or strengthening their attitudes. But if the arguments are judged to be strong and persuasive, they can lead to attitude change that is enduring.

A variety of additional studies can be built around the dynamics of dual process theories as they pertain to racial identities and attitudes. The motivation and opportunities of participants can be manipulated to determine the conditions that increase or reduce the salience of race. We can experiment with manipulating information processing modes by varying the speed of decision making, the stakes of the decision, and increasing personal accountability to see whether central or peripheral routes are followed. Explicit racial messages, for example, should exert less influence on people who have had sufficient opportunity and motivation to engage in self-monitoring of their responses to the racial cue (Terkildsen 1993). Cognitive elaboration, of course, should not be expected to ensure attitude change. As the Whittler and Spira (2002) study

discovered, racial identities can anchor viewpoints (like party identification) through motivated reasoning and biased information processing.

6. Conclusion

Experimental research on the political perspectives of minorities holds much promise for advancing our understanding of U.S. politics. Indeed, the insights generated by experimental studies of framing, persuasion, racial priming, and political mobilization in both majority and minority populations have made it difficult to think of these topics outside of the experimental context. Observational studies of these subjects are hampered by selection biases in the distribution and receipt of treatments and lack of control over the design of treatments. One of the most important advantages of experimental over traditional observational or behavior methods is the promise of greater internal validity of the causal inferences drawn in the experiment. We can test the impact of alternative treatments without strong priors about the mechanism that explains why one treatment will be more effective than another.

The range of possible studies is exciting. For example, we can randomly manipulate the background characteristics of hypothetical candidates for office in terms of their partisanship, race, or ideology and estimate the impact of these differences on candidate preference among voters. Similarly, experiments could be designed to highlight or frame specific features of candidates or issues to observe the effect of such manipulations on voter preferences. The salience of the voter's racial and ethnic identity can be heightened or reduced to see how group identity and campaign messages interact to change voter preferences or increase turnout. Efforts to embed research designs and studies in a theory of information processing may yield the most fruitful set of results. In experimental designs, the dynamics of dual process theories can be

exploited to manipulate participants' motivation and opportunity to evaluate information in order to reveal the conditions that systematically influence the salience of race.

At the same time, experimentation cannot be regarded as a substitute for theory building. Results across studies are often conflicting, illustrating the sensitivity of results to variations in measurement and the sample of experimental participants. In the context of research on racial and ethnic minorities, we discussed how theory is essential for conceptual development, designing treatments, interpreting results, and generalizing beyond particular studies. We also identified what we believe to be promising directions to address the external validity of experimental designs, including the incorporation of debate and competition, use of over-time panel designs, and greater attention to the interaction between treatments and political contexts.

References

- Abrajano, Marisa A., Jonathan Nagler, and R. Michael Alvarez. 2005. "A Natural Experiment of Race-Based and Issue Voting: The 2001 City of Los Angeles Elections." *Political Research Quarterly* 58: 203-18.
- Appiah, O. 2002. "Black and White Viewers' Perception and Recall of Occupational Characters on Television." *Journal of Communications* 52: 776-93.
- Barreto, Matt A. 2007. "Si Se Puede! Latino Candidates and the Mobilization of Latino Voters." *American Political Science Review* 101: 425-41.
- Bobo, Lawrence, and Frank Gilliam. 1990. "Race, Sociopolitical Participation and Black Empowerment." *American Political Science Review* 84: 379-93.
- Bobo, Lawrence D., and Devon Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." *DuBois Review* 1: 151-80.
- Chong, Dennis. 2000. *Rational Lives: Norms and Values in Politics and Society*. Chicago: University of Chicago Press.
- Chong, Dennis, and James N. Druckman. 2007. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101: 637-55.

- Chong, Dennis, and James N. Druckman. 2008. "Dynamic Public Opinion: Framing Effects Over Time." Paper presented at the annual meeting of the American Political Science Association, Boston, MA.
- Chong, Dennis, and Dukhong Kim. 2006. "The Experiences and Effects of Economic Status among Racial and Ethnic Minorities." *American Political Science Review* 100: 335-51.
- Chong, Dennis, and Reuel Rogers. 2005. "Racial Solidarity and Political Participation." *Political Behavior* 27: 347-74.
- Clawson, Rosalee A., Elizabeth R. Kegler, and Eric N. Waltenburg. 2003. "Supreme Court Legitimacy and Group-Centric Forces: Black Support for Capital Punishment and Affirmative Action." *Political Behavior* 25: 289-311.
- Dawson, Michael. 1994. *Behind the Mule Race and Class in African American Politics*. Princeton, NJ: Princeton University Press.
- Fazio, R. H., and Olson, M. A. 2003. "Attitudes: Foundations, functions, and consequences." In *The Handbook of Social Psychology*, eds. Michael A. Hogg, and Joel M. Cooper. London: Sage.
- Feldman, Stanley, and John Zaller. 1992. "The Political Culture of Ambivalence: Ideological Responses to the Welfare State." *American Journal of Political Science* 36: 268-307.
- Forehand, Mark R., and Rohit Deshpande. 2001. "What We See Makes Us Who We Are: Priming Ethnic Self-Awareness and Advertising Response." *Journal of Marketing Research* 38: 336-48.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "Rethinking the Survey Experiment." *Political Analysis* 15: 1-21.
- Garcia Bedolla, Lisa, and Melissa R. Michelson. 2009. "What Do Voters Need to Know? Testing the Role of Cognitive Information in Asian American Voter Mobilization." *American Politics Research* 37: 254-74.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653-63.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.
- Gilens, Martin. 1999. *Why Americans Hate Welfare: Race, Media, and the Politics of Anti-Poverty Policy*. Chicago: University of Chicago Press.

- Gilliam, Frank D., Jr., and Shanto Iyengar. 2000. "Prime Suspects: The Impact of Local Television News on the Viewing Public." *American Journal of Political Science* 44: 560-73.
- Green, Donald P. 2004. "Mobilizing African-American Voters Using Direct Mail and Commercial Phone Banks: A Field Experiment." *Political Research Quarterly* 57: 245-55.
- Herek, Gregory M., J. Roy Gillis, and Erik K. Glunt. 1998. "Culturally Sensitive AIDS Educational Videos for African American Audiences: Effects of Source, Message, Receiver, and Context." *American Journal of Community Psychology* 26: 705-43.
- Huber, Gregory A., and John S. Lapinski. 2008. "Testing the Implicit-Explicit Model of Racialized Political Communication." *Perspectives on Politics* 6: 125-34.
- Hurwitz, Jon, and Mark Peffley. 2005. "Explaining the Great Racial Divide: Perceptions of Fairness in the U.S. Criminal Justice System." *Journal of Politics* 67: 762-83.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.
- Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. Chicago: University of Chicago Press.
- Kinder, Donald R., and Nicholas Winter. 2001. "Exploring the Racial Divide: Blacks, Whites, and Opinions on National Policy." *American Journal of Political Science* 45: 439-56.
- Kuklinski, James H., and Norman L. Hurley. 1996. "It's a Matter of Interpretation." In *Political Persuasion and Attitude Change*, eds. Diana C. Mutz, Paul M. Sniderman, and Richard A. Brody. Ann Arbor: University of Michigan Press.
- Lien, Pei-te. 2001. *The Making of Asian America Through Political Participation*. Philadelphia: Temple University Press.
- Lodge, Milton, and Charles Taber. In press. *The Rationalizing Voter*. New York: Cambridge University Press.
- Marschall, Melissa. 2001. "Does the Shoe Fit? Testing Models of Participation for African-American and Latino Involvement in Local Politics." *Urban Affairs Review* 37: 227-48.
- Mendelberg, Tali M. 2001. *The Race Card: Campaign Strategy, Implicit Messages and the Norm of Equality*. Princeton: Princeton University Press.
- Michelson, Melissa R. 2005. "Meeting the Challenge of Latino Voter Mobilization." *Annals of the American Academy of Political and Social Science* 601: 85-101.

- Miller, Arthur, Patricia Gurin, Gerald Gurin, and Oksana Malanchuk. 1981. "Group Consciousness and Political Participation." *American Journal of Political Science* 25: 494-511.
- Peffley, Mark, and Jon Hurwitz. 2007. "Persuasion and Resistance: Race and the Death Penalty in America." *American Journal of Political Science* 51: 996-1012.
- Petty, Richard E., and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.
- Petty, Richard E., Fleming, Monique A., and White, P. H. 1999. "Stigmatized sources and persuasion: Prejudice as a determinant of argument scrutiny." *Journal of Personality and Social Psychology* 76: 19-34.
- Petty, Richard E., and Wegener, D. T. 1999. "The Elaboration Likelihood Model: Current Status and Controversies." In *Dual Process Theories in Social Psychology*, eds. Shelly Chaiken and Yaacov Trope. New York: Guilford Press.
- Ramirez, Ricardo. 2005. "Giving Voice to Latino Voters: A Field Experiment on the Effectiveness of a National Nonpartisan Mobilization Effort." *Annals of the American Academy of Political and Social Science* 601: 66-84.
- Sears, David O., Jim Sidanius, and Lawrence Bobo, Eds. 2000. *Racialized Politics: The Debate About Racism in America*. Chicago: University of Chicago Press.
- Shingles, Richard D. 1981. "Black Consciousness and Political Participation: The Missing Link." *American Political Science Review* 75: 76-91.
- Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*, eds. Willem E. Saris and Paul M. Sniderman. Princeton: Princeton University Press.
- Tate, Katherine. 1994. *From Protest to Politics: The New Black Voters in American Elections*. Cambridge: Harvard University Press.
- Terkildsen, Nayda. 1993. "When White Voters Evaluate Black Candidates: The Processing Implication of Candidate Skin Color, Prejudice, and Self-Monitoring." *American Journal of Political Science* 37: 1032-53.
- Trivedi, Neema. 2005. "The Effect of Identity-Based GOTV Direct Mail Appeals on the Turnout of Indian Americans." *Annals of the American Academy of Political and Social Science* 601: 115-22.

- Wang, Xiao, and Laura M. Arpan. 2008. "Effects of Race and Ethnic Identity on Audience Evaluation of HIV Public Service Announcements." *The Howard Journal of Communications* 19: 44-63.
- White, Ismail K. 2007. "When Race Matters and When It Doesn't: Racial Group Differences in Response to Racial Cues." *American Political Science Review* 101: 339-54.
- White, Paul H., and Stephen G. Harkins 1994. "Race of Source Effects in the Elaboration Likelihood Model." *Journal of Personality and Social Psychology* 67: 790-807.
- Whittler, Tommy E., and Joan Scatone Spira. 2002. "Model's Race: A Peripheral Cue in Advertising Messages?" *Journal of Consumer Psychology* 12: 291-301.
- Wong, Janelle S. 2005. "Mobilizing Asian American Voters: A Field Experiment." *Annals of the American Academy of Political and Social Science* 601: 102-14.

ⁱ We are grateful to Xin Sun, Caitlin O'Malley, and Thomas Leeper for research assistance on this project.

ⁱⁱ Their explanation assumes the strength of attitudes toward the death penalty is positively correlated with education.

ⁱⁱⁱ A full review of GOTV experiments is provided by Michelson and Nickerson's chapter in this volume.

^{iv} There is a risk the individuals in these more racialized political contexts may have already been activated by the ongoing campaign prior to the experimental manipulation. This pre-treatment of respondents may dampen the impact of any further experimental treatment that duplicates what has already occurred in the real campaign.

VII. Institutions and Behavior

24. Experimental Contributions to Collective-Action Theory

Eric Coleman and Elinor Ostrom¹

Collective-action problems are difficult problems that pervade all forms of social organization, from within the family, to the organization of production activities within a firm, and to the provision of public goods (PG) and the management of common-pool resources (CPRs) at local, regional, national, and global scales. Collective-action problems occur when a group of individuals could achieve a common benefit if most contribute needed resources. Those who would benefit the most, however, are individuals who do *not* contribute to the provision of the joint benefit and free-ride on the efforts of others. If all free-ride, however, no benefits are provided.

Political scientists trying to analyze collective-action problems have been influenced by a narrow, short-term view of human rationality combining an all-powerful computational capacity, on the one hand, with no capability to adapt or acquire norms of trustworthiness and fair contributions to the provision of collective benefits, on the other. To provide public goods, it is thought that governments must devise policies that change incentives to coerce citizens to contribute to collective action.

Formal analysis of collective-action problems has been strongly affected by the pathbreaking work of Olson (1965) on *The Logic of Collective Action* and the use of game theory (e.g., R. Hardin 1982; Taylor 1987), which improved the analytical approach to these problems. By replacing the naive assumptions of earlier group theorists (Bentley 1949; Truman 1958) that individuals will always pursue common ends, these modes of analysis force analysts to recognize the essential tensions involved in many potential social interactions. Using the same model of

individual behavior used to analyze production and consumption processes of private goods to examine collective-action problems was an essential first step toward providing a firmer foundation for all types of public policy. The empirical support for these predictions within a competitive market setting gave the enterprise an initial strong impetus.

Homo economicus has turned out to be a special analytical tool rather than the general theory of human behavior. Models of short-term material self-interest have been highly successful in predicting marginal behavior in competitive situations in which selection pressures screen out those who do not maximize external values, such as profits in a competitive market (Alchian 1950; Smith 1991), or the probability of electoral success in party competition (Satz and Ferejohn 1994). Thin models of rational choice have been less *successful* in explaining or predicting behavior in one-shot or finitely repeated social dilemmas in which the theoretical prediction is that no one will cooperate (Poteete, Janssen, and Ostrom 2010, ch. 6; Ostrom in press). Research using observational data also shows that some groups of individuals do engage in collective action to provide local public goods or manage common-pool resources without external authorities offering inducements or imposing sanctions (see NRC 2002).

1. Cooperation in the Prisoner's Dilemma

One way to conceptualize a collective-action dilemma is the Prisoner's Dilemma (PD) game shown in Figure 24-1. Imagine two states engage in a nuclear arms race. There are two players (the two states) who choose simultaneously to either cooperate (reduce armaments) or defect (build armaments). If only one state cooperates, it will be at a strategic disadvantage. In this case, the outcome (cooperate, defect) has a payout of -1 to the cooperating state. The defecting state receives a payout of 2 because of strategic gains. If both states cooperate, each would receive a payout of 1 because they do not bear the costs of building armaments. If neither cooperates, both

receive a payout of zero because neither has gained a strategic advantage. In short, states are best off if they can dupe others into cooperating while they defect, moderately well off if they both cooperate, worse off if they both defect, and in very bad trouble if they are the sole cooperator.

[Figure 24-1 about here]

Camerer (2003) stresses that *games* are not equivalent to *game theory*. Games denote the players, strategies, and rules for making decisions in particular interactions. Game theory, on the other hand, is a “mathematical derivation of what players with different cognitive capabilities are likely to do in games” (Camerer 2003, 3). Collective-action game theory was dominated until recently by the view of short-term rational self-interest expounded by Olson (1965).

In the PD game, if there are no possibilities of enforceable, binding contracts, this view predicts defection as it is a strictly dominant strategy (Fudenberg and Tirole 1991, 10). This prediction can change if repeated play is allowed. A long stream of political science research, led by Axelrod (1984), has examined repeated PD games and found that if play is repeated indefinitely, cooperative strategies are theoretically possible. However, if there is a predetermined end to repetition, the strictly dominant theoretical strategy remains to defect in each round (Axelrod and Hamilton 1981, 1392).

This view of human behavior led many to conclude that without external enforcement of binding contracts, or some other change in institutional arrangements, humans will not cooperate in collective-action dilemmas (Schelling 1960, 213–14; G. Hardin 1968). Predictions from the PD game generalize to the CPR and PG dilemmas discussed in this chapter – little or no cooperation if they are one-shot or finitely repeated and a possibility of cooperation only in dilemmas with indefinite repetition.

The behavioral revolution in economics and political science led some to question these predictions (see Ostrom 1998). Experimental game theory has been indispensable in challenging the conventional view of human behavior and improving game-theoretic predictions in collective-action dilemmas. As behaviors inconsistent with predictions from the conventional view are uncovered, analysts change their theories to incorporate the anomalies. For example, some authors have examined preferences for inequality aversion (Fehr and Schmidt 1999), preferences for fairness (Rabin 1993), and emotional states (Cox, Friedman, and Gjerstad 2007) to explain behavior in collective-action dilemmas (see also Camerer 2003, 101–13).

In the next section, we briefly discuss how field evidence suggested that the conventional model did not adequately explain behavior in collective-action dilemmas. The careful control possible in experiments is uniquely suited both to uncovering the precise degree of anomalous behavior and for calibrating new models. As these new models of behavior are developed, they should then be subjected to the same rigorous experimental and nonexperimental testing as the conventional model.

2. Evidence from Observational Studies

Much has been written about collective action in observational studies (NRC 1986, 2002). These studies are particularly useful for experimentalists because they provide examples of behaviors and strategies that people employ in real settings to achieve cooperation. Experiments have then made these behaviors possible in laboratory settings in order to assess exactly how effective they can be. For example, much early work indicated that people are willing to sanction noncooperators at a personal cost to themselves (Ostrom 1990). This literature influenced experimentalists to create the possibility of allowing sanctioning in laboratory environments (Ostrom, Gardner, and Walker 1994; Fehr and Gächter 2000), which led to the development of

the inequality aversion model (Fehr and Schmidt 1999). Other variables found in observational research that have influenced experimental work are: group size and heterogeneity (Olson 1965), rewarding cooperators (Dawes et al. 1986), communication and conditionally cooperative strategies (Ostrom 1990), and interpersonal trust (Rothstein 2005). By and large, these same factors appear to be important in experimental studies.

We view field and experimental work as complementary; that is, we have more confidence in experimental results because they are confirmed by fieldwork and vice versa. While observational studies make important contributions to our understanding of what determines if a group acts collectively, they are limited in that there are a host of confounding factors that might account for the effects. Experimental work is uniquely suited to isolate and identify these effects and then to calibrate new models of human behavior.

Let us now turn to experimental contributions to collective-action theory. Many types of experiments involve collective-action dilemmas for the subjects involved, including Trust Game experiments (see Wilson and Eckel's chapter in this volume), Ultimatum Game experiments, and a host of others (see Camerer 2003 for a review). In this chapter, we will focus on two games that have received much attention in the experimental literature. In Section 2, we review research and contributions from PG experiments, and in Section 3, we review CPR experiments. In Section 4, we discuss the emerging role of laboratory experiments in the field, and in Section 5, we conclude.

3. Public-Goods Experiments

The PD game discussed in the previous section is a special case of a PG game. Suppose that instead of Cooperate and Defect, we labeled these columns Contribute and Withhold. The game is structured such that a public good is provided to all players in proportion to the number of players that contribute. Suppose that the public good can be monetized as 4 dollars per player

contribution. If both players contribute, each receives 4 dollars. If only one player contributes, each receives 2 dollars. Suppose further that it costs each player 3 dollars to contribute. This is the same game structure as depicted in the PD game in Figure 24-1.ⁱⁱ We can also write the payouts from this PG game in equation form. Let $C_i \in \{0,1\}$ represent the decision to contribute, so that C_i takes the value of 1 if player i decides to contribute and C_i takes the value of 0 if player i decides to withhold. The payment to player i in this one-shot PD game is

$$\pi_i = \frac{4(C_i + C_j)}{2} - 3C_i. \quad (1)$$

If player i maximizes their own income, then they will select the level of C_i that maximizes Equation 1. In this case, the prediction is that $C_i = 0$, because the net effect of one person's contribution is -1 . Let us relax some of the assumptions from Equation 1 to develop a general PG game. First, we add n players to the game and relax the assumption that the decision to contribute is binary. That is, instead of contributing or not contributing, suppose that the subjects can determine a specific amount to contribute, C_i . In general, C_i can be allowed to vary up to some initial endowment, and can take any value between 0 and the endowment, E_i . Let the marginal benefits to unilateral contribution be any value A_i . Let us call the costs of contributing B_i . The general PG game, then, is

$$\pi_i = \frac{A_i}{n} \sum_{j=1}^n C_j - B_i C_i,$$

where

$$C_i \in [0, E_i]. \quad (2)$$

The parameter B_i is often set to 1, and A_i is set to be the same for every player. In this case, the primary characteristic used to describe the game is the ratio of marginal benefits to the number of subjects, $\frac{A}{n}$. This ratio is known as the Marginal Per Capita Return (MPCR). If $B = 1$, then A must

be less than n for this to be a PG game and for this to remain a collective-action dilemma. The closer the MPCR is to 1, the higher the benefits from cooperation.ⁱⁱⁱ

The experimental protocols for PG games are typically abstract, instructing subjects to allocate their endowment to a group or an individual fund. The individual fund has a rate of return equal to B . The rate of return on the group fund is equal to the MPCR. The PG game is also often referred to as a Voluntary Contribution Mechanism (VCM) because subjects make voluntary contributions to this group fund.

Baseline Public-Goods Experiments

In the baseline PG experiment, subjects are each endowed with the same number of tokens, and receive the same MPCR and the same rate of return to their private accounts. If one assumes that subjects behave according to narrow self-interest, then one would expect no contributions in any round of the game.

In one of the first PG experiments, Isaac and Walker (1988a) endowed each subject, in groups of four, with 62 tokens and repeated play for twenty rounds. The MPCR was \$0.003 per token, while the return to the private account was \$0.01. The dotted line (NC–NC) at the bottom of Figure 24-2 shows that in the first round, subjects contributed about 50 percent of their endowment to the public good. Over time, these contributions steadily fall toward zero. This result is fairly robust, having been replicated in a number of studies (Isaac and Walker 1988b, 184). This result appears to confirm the traditional model of narrow self-interest, especially if the anomalies at the beginning rounds can be attributed to learning (see Muller et al. 2008).

Communication

When participants in the CPR experiment cannot communicate, their behavior approaches zero contributions over time. Participants in most field settings, however, are able to communicate

with one another at least from time to time, either in formally constituted meetings or at social gatherings. In an effort to take one step at a time toward the fuller situations faced by groups providing public goods, researchers have tried to assess the effects of communication.

Figure 24-2 shows the results from two additional treatments in Isaac and Walker (1988b). In the treatment C–NC, subjects were allowed to communicate at the beginning of each of the first ten rounds but were not allowed to communicate thereafter. In the treatment NC–C, subjects were not allowed to communicate for the first ten rounds, but starting in round 11 were allowed to communicate in every round thereafter. In the baseline treatment (NC–NC), described in the previous section, no communication was allowed in any round. Nonbinding communication is referred to as *cheap talk* and is predicted to have no effect on outcomes in the PG game (Harsanyi and Selten 1988, 3).

[Figure 24-2 about here]

Figure 24-2 clearly shows, however, that communication has a profound effect. Take the case where communication was allowed in the last ten rounds of play, the dashed line in the left panel of Figure 24-2. It appears for the first ten rounds that the subjects are on a similar trajectory as those who are never allowed to communicate; that is, mean contributions to the public good are steadily falling. After communication in round 10, however, mean contributions increase substantially. In the second half of the game, contributions are near 100 percent of the total endowments. The right panel of Figure 24- 2 shows that the mean contributions are significantly higher in the second ten rounds. Perhaps more astonishing is that when communication is allowed in the first ten rounds, contributions continue to remain high in the second ten rounds (the solid line in the left panel of Figure 24-2), although this tends to taper off in the last three rounds. Still, mean contributions remain essentially the same across ten-round

increments, as indicated in the right panel of Figure 24-2. Such strong effects of communication have been found in many studies (Sally 1995). In fact, Miettinen and Suetens (2008, 945) have argued that “Researchers have reached a rather undisputed consensus about the prime driving force of the beneficial effect of communication on cooperation.”

Subject communication tends to focus around strategies for the game. Often, subjects will agree on some predetermined behavior. While they frequently do what they promise, some defections do occur. If promises were not kept, subjects use the aggregated information on the outcomes from the previous round to castigate the unknown participant(s) who does not keep to their agreement. Subjects can be indignant about evidence of defection and express their anger openly. Not only does the content of communication matter, but the medium of communication is also important. Frohlich and Oppenheimer (1998, 394) found that those allowed to communicate face-to-face reach nearly 100 percent contribution to the public good, while communication via e-mail improves contributions to about 75 percent.^{iv}

While the findings show that communication makes a major difference in outcomes, some debate exists as to *why* communication alone leads to better results (Buchan, Johnson, and Croson 2006). A review by Shankar and Pavitt (2002) suggests that voicing of commitments and development of group identity and norms seem to be the best explanations for why communication makes a difference. Another reason may be the revelation of a participant’s type, which is one source of incomplete information in experimental games. For example, face-to-face communication and verbal commitments may change participants’ expectations of other participants’ responses. In particular, if a participant believes that other participants are of a cooperative type (i.e., will cooperate in response to cooperative play), that participant may play

cooperatively to induce cooperation from others. In this case, cooperating can be sustained as rational play in the framework of incomplete information regarding participant types.

Leadership

Political leadership has long been linked to the provision of public goods (Frohlich, Oppenheimer, and Young 1971). Economists have traditionally modeled leadership in PG settings as “leading by example” (Levati, Sutter, and van der Heijden 2007). That is, one subject is randomly selected from the experiment to be a first mover in the PG game. Other subjects are then hypothesized to take cues from the first mover. Evidence suggests that this type of leadership increases cooperation in PG games (Levati, Sutter, and van der Heijden 2007).

Leading by example, however, seems to be a coarse operationalization of leadership. Experimental research would benefit from more thoughtful insights from political leadership theory, and these theories could be carefully examined by endowing leaders with different capabilities. For example, characteristics of Machiavelli’s prince might be manipulated in the laboratory. Is it truly better to be feared than loved? Do leaders who devise punishments for those who do not contribute to the public good fare better than those who offer rewards?^v

In addition, while much research has been conducted on the election process of political leadership (see Morton and Williams’s chapter in this volume), the effects of such institutions on PG provision have not been thoroughly explored. One notable exception is a recent paper by Hamman, Woon, and Weber (2008). The authors investigate the effects of political leadership by forcing groups to delegate authority to one elected (majority-rule) leader who then determines the contributions of each member to the public good. The delegate is then reelected in subsequent periods, ensuring some accountability to other group members. The authors find that

under delegated PG provision, groups elect delegates who ensure that the group optimum is almost always met.

4. Common-Pool Resource Experiments

Common-pool resources such as lakes, forests, fishing grounds, and irrigation systems are resources from which one person's use subtracts units that are then not available to others, and it is difficult to exclude or limit users once the resource is provided by nature or produced by humans (Ostrom et al. 1994). When access to a common-pool resource is open to all, anyone who would like to use the resource has an incentive to appropriate more resource units when acting independently than if they could find some way of coordinating their appropriation activities with others.

CPR games are different from PG games in two ways: (1) the decision task of a CPR is removing resources from a joint fund instead of contributing and (2) appropriation is rivalrous. This rivalry can be thought of as an externality that occurs because the payout rate from the common-pool resource depends nonlinearly on total group appropriation. Initially, it pays to withdraw resources from the common-pool resource, but subjects maximize group earnings when they invest some, but not all, of their effort to appropriate from the CPR.

The first series of CPR experiments was initiated at Indiana University to complement ongoing fieldwork. The series started with a static, baseline situation that was as simple as possible while keeping crucial aspects of the problems that real harvesters face. The payoff function used in these experiments was a quadratic function similar to the theoretical function specified by Gordon (1954). The initial resource endowment of each participant consisted of a set of tokens that the participants could allocate between two situations: Market 1, which had a fixed return and Market 2, which functioned like a common-pool resource so that the return was determined in part by the actions of all participants in the experiment.

Each participant i could choose to invest a portion x_i of his/her endowment of ω in the common resource Market 2, and the remaining portion $\omega - x_i$ is then invested in Market 1. The payoff function used in Ostrom et al. (1994) is

$$u_i(\mathbf{x}) = \begin{cases} 0.05 \cdot \omega & \text{if } x_i = 0 \\ 0.05 \cdot (\omega - x_i) + (x_i / \sum x_i) \cdot F(\sum x_i) & \text{if } x_i > 0 \end{cases} \quad (3)$$

where

$$F(\sum x_i) = (23 \cdot \sum_{i=1}^8 x_i - 0.25 \cdot (\sum_{i=1}^8 x_i)^2) / 100. \quad (4)$$

According to this formula, the payoff of someone investing all ω tokens in Market 1 ($x_i = 0$) is $0.05 \cdot \omega$. The payoff from Market 1 is like a fixed wage paid according to the hours devoted to working. Investing part or all of the tokens in Market 2 ($x_i > 0$) yields an outcome that depends on the investments of the other participants.

Basically, if appropriators put all of their assets into the outside option (working for a wage rather than fishing), they are certain to receive a fixed return equal to the amount of their endowment times an unchanging rate. If appropriators put some of their endowed assets into the CPR, they received part of their payoff from the outside option and the rest from their proportional investment in the CPR. The participants received aggregated information after each round, so they did not know individual actions. Each participant was endowed with a new set of tokens in every round of play. Their outside opportunity was valued at \$.05 per token. They earned \$.01 on each outcome unit they received from investing tokens in the CPR. The number of rounds in each experiment varied between twenty and thirty rounds, but participants were informed that they were in an experiment that would last no more than two hours.

The solid line in Figure 24-3 shows the relationship between total group investments in Market 1, the fixed wage rate, and group earnings from that market. The dashed line shows the

relationship between group investments in Market 2, the CPR, and its group earnings. Wage earnings are interpreted as the opportunity costs of investing in the CPR. Total earnings, represented by Equation 3, are maximized when the CPR earnings minus wage earnings is maximized. Given the parameterization of Equation 4, this occurs at total investment in the CPR of 36 tokens.

[Figure 24-3 about here]

The symmetric Nash equilibrium for this finitely repeated game (if subjects are not discounting the future, and each participant is assumed to be maximizing own-monetary returns) is for each participant to invest 8 tokens in the common-pool resource for a total of 64 tokens (see Ostrom et al. 1994, 111–12). They could, however, earn considerably more if the total number of tokens invested were 36 tokens (rather than 64 tokens). The baseline experiment is an example of a commons dilemma in which the Nash equilibrium outcome involves substantial overuse of a common-pool resource, while a much better outcome could be reached if participants were to lower their joint use relative to the Nash equilibrium.

Baseline CPR Experiments

Participants interacting in baseline experiments substantially overinvested as predicted. At the individual level, participants rarely invested eight tokens – the predicted level of investment at the symmetric Nash equilibrium. Instead, all experiments provided evidence of an unpredicted and strong pulsing pattern in which individuals appear to increase their investments in the common-pool resource until there is a strong reduction in yield, at which time they tend to reduce their investments leading to an increase in yields. At an aggregate level, behavior begins to approach the symmetric Nash equilibrium level in later rounds.

Voting

If subjects are allowed to make binding agreements about their behavior in the CPR game, they might overcome the free-rider problem. People may be willing to voluntarily precommit to limit the choices available to the group in the future in order to achieve a more preferred group outcome (Elster 1977).

Ostrom et al. (1994) investigated if subjects were willing to precommit to binding contracts in the CPR game and if those contracts would produce efficient results. The authors gave groups of seven subjects an opportunity to use simple majority rule to develop an appropriation system for themselves. In the lab, they found people moving toward a minimum winning coalition. Subjects knew the computer numbers and began to make proposals like “Let’s give all the optimal resources to computer number one, two, three, and four.” Of course, this came from somebody who was using computer number one, two, three, and four; and they zeroed out five, six, and seven. When the voting rule was changed to require unanimity, the subjects also went to the optimum, but they allocated it across the entire group.

Sanctioning

In the field, many users of CPRs do monitor and sanction one another (Coleman and Steed 2009). Engaging in costly monitoring and sanctioning behavior is not consistent with the theory of norm-free, full rationality (Elster 1989, 40–41).

To test if participants would use their own resources to sanction others, Ostrom, Walker, and Gardner (1992) conducted a modified CPR game. Individual investments in each round were reported as well as the total outcomes.^{vi} Participants were then told that in the subsequent rounds, they would have an opportunity to pay a fee in order to impose a fine on the payoffs received by another participant. In brief, the finding from this series of experiments was that much sanctioning occurs. Most of the sanctions were directed at subjects who had ordered high levels of tokens. Participants react both to the cost of sanctioning and to the fee/fine relationships. They

sanction more when the cost of sanctioning is less and when the ratio of the fine to the fee is higher (Ostrom et al. 1992). Participants did increase benefits through their sanctioning but reduced their net returns substantially due to the high use of costly sanctions.

5. Bringing the Lab to the Field

We think laboratory experiments in field settings hold a challenging, yet potentially fruitful avenue for political scientists to investigate collective-action theory. Many collective-action dilemma experiments have been conducted in developed countries with undergraduate students from university settings. The initial reasons for this selected sample of participants were their accessibility, control for the experimenters, and lower overall costs.^{vii} Experiments have now been conducted with nonstudent populations and with more salient frames of the decision tasks, and there are often striking differences in behavior across these populations (Henrich et al. 2004).

Because of the increased costs and logistical problems associated with these types of experiments, researchers should think carefully about the reasons for extending their research to field settings. Harrison and List (2004) argue that key characteristics of subjects from the experimental sample need to match the population for which inferences will be generalized. That is, if age, education, or some political or cultural phenomenon unique to students are not key characteristics of the theory being tested, then a student sample may be appropriate to test the theory (see Druckman and Kam's chapter in this volume). On the other hand, if one wishes to investigate the effects of communism, for example, then a sample of U.S. students would not be appropriate; an older age sample from a post-communist country would be needed for the experiment (Bahry et al. 2005).

Ethnic Diversity and the Mechanisms of Public-Goods Provision

Habyarimana et al. (2007) were interested in why ethnic heterogeneity leads to decreased investments in public goods. In order to test a number of possible mechanisms, the authors conducted a set of surveys and experiments using 300 randomly selected subjects recruited from a slum in Kampala, Uganda. It was necessary to use such subjects because “ethnicity is highly salient in everyday social interactions” and the subjects had almost exclusive responsibility for supplying local public goods (Habyarimana et al. 2007, 712).

The authors identified three different potential mechanisms. First, they tested the effects of ethnic heterogeneity of tastes – the extent to which different ethnic groups care about different types of public goods and their preferences that public goods are provided to their own ethnic group and not others. Using a survey instrument, the authors found that there was little difference in tastes both as to which types of public goods subjects preferred (drainage, garbage collection, or security) or to the means of their provision (government versus local). The authors then had subjects play an anonymous dictator game to test if noncoethnic pairs have different tastes for income distribution than coethnic pairs and found that this was not the case. In the dictator game, a subject is given some sum of money and is simply asked to divide the money with a partner. The subject can give all, none, or anything in between.

Second, they tested the effects of technological advantages of homogeneous groups. Such groups can draw on common language and culture to produce public goods and are better able to identify noncooperative members. To test the first proposition, that coethnics work well together, the authors had subject pairs solve puzzles. They found that while coethnic pairs were more likely to solve the puzzle than noncoethnic pairs, the difference was not significant. The second submechanism is that members of homogeneous groups can find and identify noncooperative members through social networks. The authors had subjects locate randomly selected

nonexperimental subjects as “targets” in Kampala. Those of the same ethnicity as their target found the target 43 percent of the time, while those of different ethnicities found the target only 28 percent of the time.

Third, ethnicity might serve to coordinate strategies through social sanctioning. To test this mechanism, the authors had subjects play a nonanonymous dictator game. The authors reason that in such a game, if subjects give nothing to their partner, they might still be subject to social shame for acting noncooperatively. The authors found that certain types of subjects discriminate their giving based on ethnicity when play is not anonymous. That is, they give less to noncoethnics than they do to coethnics.

In this study, experimental methods allowed the researchers to carefully parse out and test different causal mechanisms that explain why ethnically heterogeneous groups provide fewer public goods than homogeneous groups. Experimental methodology was needed to explore these mechanisms, as all three seem equally plausible when analyzing observation data. In addition, the field setting allowed the authors to test these theories with samples from a population where ethnic diversity was a major factor in public-goods provision.

Social Norms and Cultural Variability in Common-Pool Resources

A very interesting series of replications and extensions of CPR experiments have been conducted by Cardenas (2000) and colleagues using field laboratories set up in villages in rural Colombia. The villagers whom Cardenas invited were actual users of local forests. Cardenas wanted to assess whether experienced villagers, who were heavily dependent on local forests for wood products, would behave in a manner broadly consistent with that of undergraduate students in a developed country.

The answer to this first question turned out to be positive.^{viii} Cardenas asked villagers to decide on how many months a year they would spend in the forest gathering wood products as contrasted to using their time otherwise. Each villager had a copy of an identical payoff table. In the baseline, no-communication experiments, Cardenas found a pattern similar to earlier findings from the baseline CPR experiments. Villagers substantially overinvested in appropriation from the resource.

Face-to-face communication enabled the villagers to increase total earnings on average from 57.7 percent to 76.1 percent of optimal. Subjects filled in surveys after completing the experiments; Cardenas used these to explain the considerable variation among groups. He found, for example, that when most members of the group were already familiar with resources, they used the communication rounds more effectively than when most members of the group were dependent primarily on individual assets. Cardenas also found that “social distance and group inequality based on the economic wealth of the people in the group seemed to constrain the effectiveness of communication for this same sample of groups” (Cardenas 2000, 317; see also Cardenas 2003). In five other experiments, subjects were told that a new regulation would go into force mandating that they should spend no more than the optimal level of time in the forest each round (Cardenas, Stranlund, and Willis 2000). Subjects were also told that there would be a 50 percent chance that someone would be monitored each round. The experimenter rolled a dice in front of the participants each round to determine whether the contributions of any participant would be monitored. If an even number appeared, someone would be inspected. The experimenter then drew a number from chits numbered between one and eight placed in a hat to determine who would be inspected. Thus, the probability that anyone would be inspected was 1/16 per round – a low but realistic probability for monitoring forest harvesting in rural areas.

The monitor checked the investment of the person whose number was drawn. A penalty was subtracted from the payoff of anyone over the limit, and no statement was made to others as to whether the appropriator was complying with regulations or not.

The participants in this experiment with a rule to withdraw the “optimal” amount imposed on them actually *increased* their withdrawal levels in contrast to behavior when no rule was imposed and face-to-face communication was allowed. Thus, participants who were simply allowed to communicate with one another on a face-to-face basis were able to achieve a higher joint return than those who had an optimal but imperfectly enforced external rule imposed on them.

Some Considerations

While investigating the differences in experimental behavior across societies holds the potential for important new insights into collective-action theory, one would be remiss without mentioning some of the ethical concerns attendant to such research. Generally, payments for participation in these experiments are large compared to local wage rates. Average payments in these games generally range from one-half days wage to a week’s wage, although in some instances the stakes are even much greater. Researchers should consider both benefits that subjects receive from participating, as well as the potential for conflict if some subjects are dissatisfied with the results. Every effort should be taken to ensure that earnings remain anonymous.

6. Conclusion

While much important work has already been done in collective-action experiments, interesting questions remain. It is perhaps not surprising that considerable variation in behavior is recorded in these experiments across different societies. However, what is unclear is explaining the cultural and political dimensions driving these differences. Political scientists can make

important contributions to understanding such behavior by reference to variation in political phenomenon at the local and national level. Political corruption, for example, may be very important for determining why subjects in some societies are more cooperative than subjects in others (Rothstein 2005).

Important advances can also be made in understanding the role of different political structures and the incentives they provide in CPR and PG games. We do not understand, for example, what effect different voting rules have on the propensity to delegate punishment authority or allocative authority and what effects this may have on cooperation. Research has yet to be done that examines the effects of oversight, a third-order collective-action dilemma, on the propensity to sanction and the subsequent collective-action outcome. There is still much work to be done examining the role of different institutional arrangements on collective action.

References

- Alchian, Armen A. 1950. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy* 58: 211-21.
- Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.
- Axelrod, Robert, and William D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211(4489): 1390-6.
- Bahry, Donna, Michail Kosolapov, Polina Kozyreva, and Rick K. Wilson. 2005. "Ethnicity and Trust: Evidence from Russia." *American Political Science Review* 99: 521-32.
- Bentley, Arthur F. 1949. *The Process of Government*. Evanston, IL: Principia Press.
- Buchan, Nancy R., Eric J. Johnson, and Rachel T. A. Croson. 2006. "Let's Get Personal: An International Examination of the Influence of Communication, Culture and Social Distance on Other Regarding Preferences." *Journal of Economic Behavior and Organization* 60: 373-98.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

- Cardenas, Juan-Camilo. 2000. "How Do Groups Solve Local Commons Dilemmas? Lessons from Experimental Economics in the Field." *Environment, Development and Sustainability* 2: 305-22.
- Cardenas, Juan-Camilo. 2003. "Real Wealth and Experimental Cooperation: Evidence from Field Experiments." *Journal of Development Economics* 70: 263-89.
- Cardenas, Juan-Camilo, John K. Stranlund, and Cleve E. Willis. 2000. "Local Environmental Control and Institutional Crowding-Out." *World Development* 28: 1719-33.
- Coleman, Eric A., and Brian Steed. 2009. "Monitoring and Sanctioning on the Commons: An Application to Forestry." *Ecological Economics* 68: 2106-13.
- Cox, James C., Daniel Friedman, and Steven Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior* 59: 17-45.
- Dawes, Robyn M., John M. Orbell, Randy T. Simmons, and Alphons J.C. van de Kragt. 1986. "Organizing Groups for Collective Action." *American Political Science Review* 80: 1171-85.
- Elster, Jon. 1977. "Ulysses and the Sirens: A Theory of Imperfect Rationality." *Social Science Information* 16: 469-526.
- Elster, Jon. 1989. *Solomonic Judgments: Studies in the Limitations of Rationality*. New York: Cambridge University Press.
- Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90: 980-94.
- Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114: 817-68.
- Frohlich, Norman, and Joe A. Oppenheimer. 1998. "Some Consequences of E-Mail vs. Face-To-Face Communication in Experiment." *Journal of Economic Behavior and Organization* 35: 389-403.
- Frohlich, Norman, Joe A. Oppenheimer, and Oran R. Young. 1971. *Political Leadership and Collective Goods*. Princeton, NJ: Princeton University Press.
- Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory*. Cambridge, MA: MIT Press.
- Gordon, H. Scott. 1954. "The Economic Theory of a Common Property Resource: The Fishery." *Journal of Political Economy* 62: 124-42.
- Habyarimana, James, Macartan Humphreys, Daniel N. Posner, and Jeremy M. Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101: 709-25.

- Hamman, John, Jonathan Woon, and Roberto A. Weber. 2008. "An Experimental Investigation of Delegation, Voting, and the Provision of Public Goods." Presented at the Graduate Student Conference on Experiments in Interactive Decision Making, Princeton University, Princeton, NJ. Retrieved from <http://www.princeton.edu/~deconf/FinalPapers/HammanPaper.pdf> .
- Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162: 1243-8.
- Hardin, Russell. 1982. *Collective Action*. Baltimore, MD: Johns Hopkins University Press.
- Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42: 1009-55.
- Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.
- Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford, UK: Oxford University Press.
- Henrich, Joseph, and Natalie Smith. 2004. "Comparative Experimental Evidence from the Machiguenga, Mapuche, Huinca, and American Populations." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, eds. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford, UK: Oxford University Press.
- Isaac, R. Mark, and James M. Walker. 1988a. "Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism." *Economic Inquiry* 26: 585-608.
- Isaac, R. Mark, and James M. Walker. 1988b. "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism." *Quarterly Journal of Economics* 103: 179-99.
- Levati, M. Vittoria, Matthias Sutter, and Eline van der Heijden. 2007. "Leading by Example in a Public Goods Experiment with Heterogeneity and Incomplete Information." *Journal of Conflict Resolution* 51: 793-818.
- Miettinen, Topi, and Sigrid Suetens. 2008. "Communication and Guilt in a Prisoner's Dilemma." *Journal of Conflict Resolution* 52: 945-60.
- Muller, Laurent, Martin Sefton, Richard Steinberg, and Lise Vesterlung. 2008. "Strategic Behavior and Learning in Repeated Voluntary Contribution Experiments." *Journal of Economic Behavior and Organization* 67: 782-93.
- National Research Council (NRC). 1986. *Proceedings of the Conference on Common Property Resource Management*. Washington, DC: National Academies Press.

- National Research Council (NRC). 2002. *The Drama of the Commons*, eds. Elinor Ostrom, Thomas Dietz, Nives Dolšak, Paul Stern, Susan Stonich, and Elke Weber. Committee on the Human Dimensions of Global Change. Washington, DC: National Academies Press.
- Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.
- Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.
- Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review* 92: 1-22.
- Ostrom, Elinor. In press. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review*.
- Ostrom, Elinor, Roy Gardner, and James Walker. 1994. *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.
- Ostrom Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance is Possible." *American Political Science Review* 86: 404-17.
- Poteete, Amy R., Marco A. Janssen, and Elinor Ostrom. 2010. *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*. Princeton, NJ: Princeton University Press.
- Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281-302.
- Rothstein, Bo. 2005. *Social Traps and the Problem of Trust*. New York: Cambridge University Press.
- Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society* 7: 58-92.
- Satz, Debra, and John Ferejohn. 1994. "Rational Choice and Social Theory." *Journal of Philosophy* 91: 71-87.
- Schelling, Thomas. 1960. *The Evolution of Cooperation*. Cambridge, MA: Harvard University Press.
- Sefton, Martin, Robert Shupp, and James Walker. 2007. "The Effect of Rewards and Sanctions in Provision of Public Goods." *Economic Inquiry* 45: 671-90.
- Shankar, Anisha, and Charles Pavitt. 2002. "Resource and Public Goods Dilemmas: A New Issue for Communication Research." *The Review of Communication* 2: 251-72.

- Smith, Vernon L. 1991. *Papers in Experimental Economics*. New York: Cambridge University Press.
- Taylor, Michael. 1987. *The Possibility of Cooperation*. New York: Cambridge University Press.
- Truman, David B. 1958. *The Governmental Process*. New York: Knopf.
- van Soest, Daan, and Jana Vyrastekova. 2006. "Peer Enforcement in CPR Experiments: The Relative Effectiveness of Sanctions and Transfer Rewards and the Role of Behavioral Types." In *Using Experimental Economics in Environmental and Resource Economics*, ed. John List. Cheltenham, UK: Edward Elgar.

ⁱ The authors wish to acknowledge financial support from the National Science Foundation and from the Workshop in Political Theory and Policy Analysis at Indiana University. Suggestions and comments from Roy Duch, Arthur Lupia, and Adam Seth Levine are appreciated, as is the editing assistance of David Price and Patty Lezotte.

ⁱⁱ If both players contribute, they each receive 1 dollar (4 dollars from the public good minus 3 dollars of contribution costs). If only one player contributes, that player receives -1 dollar (2 dollars from the public good minus 3 dollars of contribution costs), but the noncontributing player receives 2 dollars (2 dollars from the public good and no expense incurred from contributing). If neither player contributes, both receive zero dollars.

ⁱⁱⁱ See Isaac and Walker (1988b) for results related to changing the relative size of the MPCR.

^{iv} See Sally (1995) for meta-analysis relating communication treatments to cooperation in experimental games.

^v For the PG game, see Sefton, Shupp, and Walker (2007); for the CPR game, see van Soest and Vyrastekova (2006).

^{vi} See Fehr and Gächter (2000) for an application in PG games.

^{vii} While laboratory experiments conducted in a university setting usually pay participants more than they would earn in a local hourly position, the costs of the experiment itself are substantially less than experiments conducted in field settings.

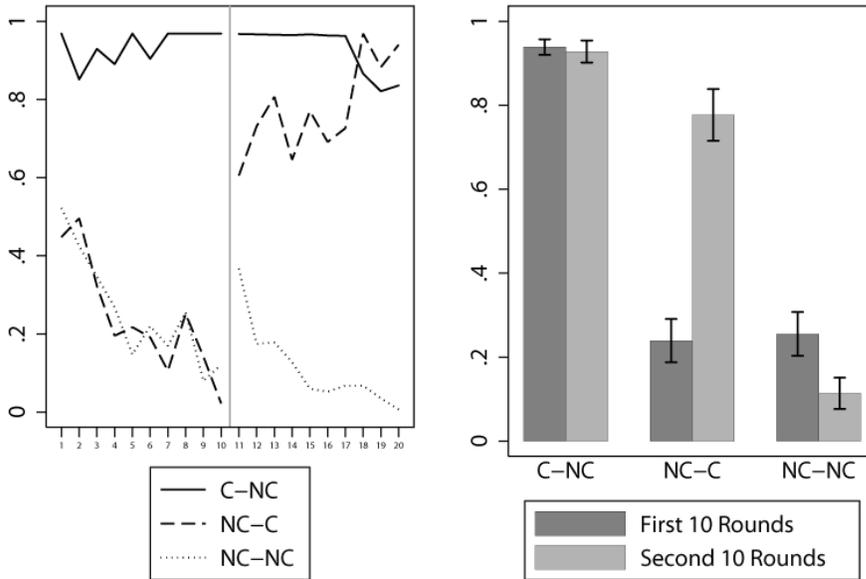
^{viii} Although, see Henrich and Smith (2004) for a counterexample.

Figure 24-1. A Prisoner's Dilemma Game

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	(1,1)	(-1,2)
	Defect	(2, -1)	(0,0)

Figure 24-2. Contributions in a Public-Goods Game

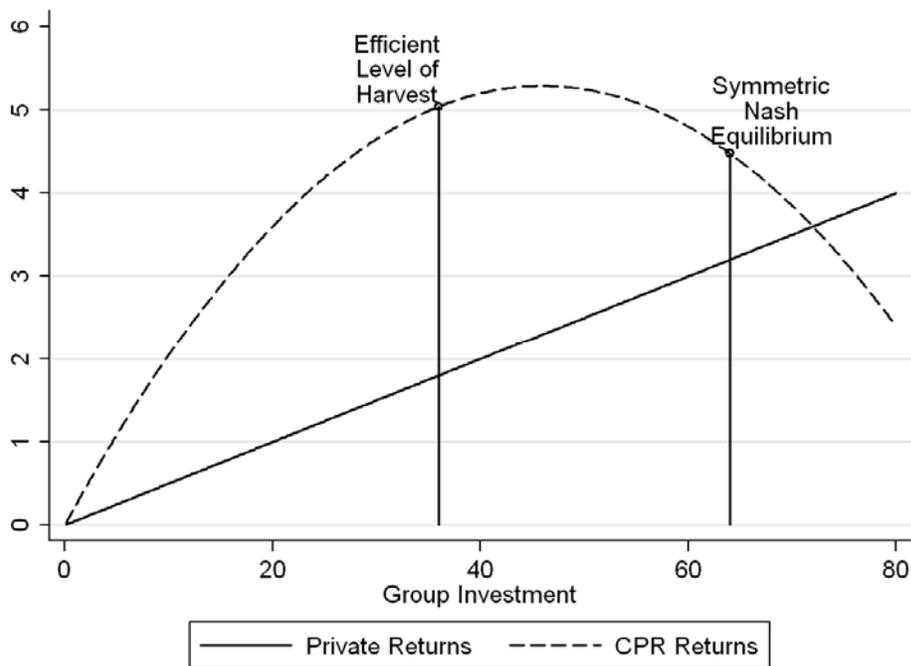
Mean Contributions by Treatments and Rounds



Source: Data from Issac and Walker (1988a).

The left panel shows the proportion of contributions to the group fund in a PG experiment, by round, for three treatments. Rounds were broken into two halves of ten rounds each. In the C-NC treatment, communication was allowed in each of the first ten rounds and no communication in the last ten rounds. In the NC-C treatment, no communication was allowed in any of the first ten rounds, but communication was allowed in each of the last ten rounds. In the NC-NC treatment, no communication was allowed in any round. The right panel shows the proportion of contributions to the group fund by halves of the experiment for each treatment, as well as 95 percent confidence intervals for those means.

Figure 24-3. A Common-Pool Resource Game



The dashed black line represents total group earnings from the CPR for all levels of group investment, and the solid black line represents earnings from the private fund. Earnings from the private fund are the opportunity costs of earnings from the CPR. The efficient level of investment in the CPR, that is, that maximizes group earnings from the CPR (CPR earnings minus the opportunity costs from the private fund), is a group investment of 36 tokens (see Ostrom et al. 1994). Note that net earnings from the CPR are negative when group investment equals 72 tokens. The symmetric Nash equilibrium is to invest 8 tokens each, or 64 total tokens, in the CPR.

25. Legislative Voting and Cycling

Gary Millerⁱ

Discoveries regarding the scope and meaning of majority rule instability have informed debate about the most fundamental questions concerning the viability of democracy. Are popular majorities the means of serving the public interest, or a manifestation of the absence of equilibrium (Riker 1982)? Should majority rule legislatures be suspect, or even avoided in favor of court decisions and bureaucratic delegation? Are the machinations of agenda setters the true source of what we take to be the legislative expression of majority rule? Are rules themselves subject to the vagaries of shifting majority coalitions (McKelvey and Ordeshook 1984)?

These and other questions were raised as a result of explorations in Arrowian social choice theory, which visualized group decisions as being the product of individual preferences and group decision rules, such as majority rule. The biggest challenge to the research agenda was majority rule instability. In general, majority rule may not be able to produce a majority rule winner (an outcome that beats every other in a two-way vote). Rather, every possible outcome could lose to something preferred by some majority coalition. McKelvey (1976) showed that the potential for instability was profound, not epiphenomenal. A population of voters with known preferences might easily choose any outcome, or different outcomes at different times. If this were true, how could scholars predict the outcome of a seemingly arbitrary and unconstrained majority rule institution, even with perfect knowledge of preferences? Was literally anything possible?

Limitations of Field Work on Legislative Instability

Political scientists tried to use legislative data to answer fundamental questions about the scope and meaning of majority rule instability. Riker (1986) used historical examples to illustrate the potential for majority rule instability. A favorite case was the 1956 House debate on federal aid to education. Despite the fact the majority party was prepared to pass the original bill over the status quo, Republicans and northern Democrats preferred the bill with the Powell Amendment, which would send no aid to segregated schools. Once amended, a third majority (southern Democrats and Republicans) preferred no bill at all. This seemed to be a graphic example of cyclic instability, because *some* majority was ready to vote down any of the three alternatives (original bill, Powell amendment, status quo) in favor of a different outcome. The outcomes “cycled”, and none of the three outcomes seemed to have a privileged position either in terms of legitimacy, manipulability, or likelihood. The determinant of the final outcome seemed to have more to do with manipulation of the agenda than anything else.

Nevertheless, this case and others like it were open to debate. Riker had made assumptions about the preferences of the legislators that were open to different interpretations. Wilkerson (1999) believed the possibility of instability as manifested by “killer amendments” was minimal. In general, political scientists could only guess about the connection between voting behavior and the underlying preferences of legislators. Without a way to measure the independent variables (preferences and rules) or the dependent variable (legislative outcomes), rational actor models seemed singularly handicapped.

Of course, the effect of a shift in preferences or rules change might offer a “natural experiment” on the effect of such a change on policy outcomes, but usually such historic changes were hopelessly confounded with other historical trends that might impact the outcome. Were the

1961 rule changes governing the House Rules Committee responsible for the liberal legislation of the next decade, or the manifestation of a change in preferences that would have brought that legislation into being in any case? Was the Republican ascendancy of 1994 the cause of welfare reform, or the vehicle for public pressure that would have brought welfare reform about in any case? Research on readily observable features of legislatures – partisanship, committee composition, constituency, etc. – led to “ambiguous and debated correlations” (Druckman et al. 2006, 629). Experimental research offered the prospect of nailing down the causation that was inevitably obscured by field data.

Early Spatial Experiments: Fiorina and Plott

It did not take long after the emergence of early rational choice models of legislative decision making for the advantages of experiments to become apparent. Fiorina and Plott (1978) set out to assess McKelvey’s demonstration that voting in two-dimensions could cycle to virtually any point in the space. “McKelvey’s result induces an interesting either-or hypothesis: ‘if equilibrium exists, then equilibrium occurs; if not, then chaos’” (Fiorina and Plott 1978, 590). In setting out to examine this hypothesis, Fiorina and Plott were the precursors of an experimental research agenda assessing the predictability of majority rule.

In addressing this question, Fiorina and Plott created what became the canonical design for majority rule experiments. They used students as subjects, presenting them with two-dimensional sets of possible decision “outcomes” – the crucial dependent variable. The dimensions were presented in an abstract way intended to render them neutral of policy or personal preferences.

The two dimensions were salient only for their financial compensation; the students saw payoff charts showing concentric circles around their highest-paying “ideal point”. One student might receive a higher payoff in the upper-right hand corner of the space, while others would prefer other areas in the space. The students were quite motivated by the payoffs, a fact which gave the experimenters control over the key independent variable – preferences.

The experimenter deliberately presented a passive, neutral face while reading instructions that incorporated a carefully specified regime of rules, determining exactly how a majority of voters could proceed to enact a policy change or to adjourn. They were recognized one at a time to make proposals, and each proposal was voted on against the most recent winning proposal. They could discuss alternatives and seek supporters for particular proposals. Subjects had few constraints beyond a prohibition against side-payments and (famously) “no physical threats.” This procedure provided rigorous control over both preferences and rules.

However, did the students behave in a way that could generalize to real legislature – and was it important if they did not? As Fiorina and Plott (1978) put it, “What makes us believe [...] that we can use college students to simulate the behavior of Congress members? Nothing” (576). They made no claims about the generalizability of the results, but did make claims about the implications of the outcomes for theory; “if a given model does not predict well relative to others under a specified set of conditions [designed to satisfy the specifications of the theory], why should it receive preferential treatment as an explanation of non-laboratory behavior . . . (376)?”

Fiorina and Plott designed two experimental settings, one with an equilibrium, and one without. The equilibrium concept of interest was the majority rule equilibrium or *core* – an alternative that could defeat, in a two-way vote, any other alternative in the space. The core

existed as a result of specially balanced preferences by the voters – the core was the ideal point of one voter, and the other pairs of voters had delicately opposing preferences on either side of the core. The Fiorina-Plott results provided significant support for the predictive power of the core, when it existed. Outcomes chosen by majority rule clustered close to the core. This conclusion was supported by subsequent experiments (McKelvey and Ordeshook 1984). Wilson (2008a, 875) analyzed experimental decision trajectories demonstrating the attractive power of the core. Majorities consistently proposed new alternatives that moved the group choice toward the core.

[Figure 25-1 here]

The other treatment did not satisfy the fragile requirements for a core (see Figure 25-1). For example, Player 1's ideal point, although a centrist outcome, could be defeated by a coalition of Players 3, 4, and 5 preferring a move to the northeast. The other treatment thus provided the crucial test of what would happen when anything could happen.ⁱⁱ

However, Fiorina and Plott's noncore experiments showed a great deal more "clustering" than could have been expected, given McKelvey's result (Figure 25-1). The variance in the outcomes was greater without a core than it was with a core – but the differences were not as striking as they had expected. They concluded, "The pattern of experimental findings does not explode, a fact which makes us wonder whether some unidentified theory is waiting to be discovered and tested." (1978, 590)

Fiorina-Plott's invitation to theorize on the apparent constraint of noncore majority rule was taken up promptly by at least two schools of thought. One school held that subjects acting on their preferences in reasonable ways produced constrained, centrist results – core-like, even

without a core. The other school emphasized that institutional structure and modifications of majority rule generated the predictable constraint on majority rule. The first school examined hypotheses about the effects of preferences changes (holding rules constant) and the second, the effects of rule changes (holding preferences constant). Experimenters could randomly assign subjects to legislative settings that varied by a tweak of the rules or a tweak of the preferences, allowing conclusions about causation that were impossible with natural legislative data.

1. Institutional Constraints on Majority Rule Instability

The behavioral revolution of the 1950s and 1960s consciously minimized the importance of formal rules in social interaction. In light of that, probably the most innovative and far-reaching idea that came out of Arrowian social choice was neo-institutionalism – the claim that rules can have an independent and sometimes counterintuitive effect on legislative outcomes.

Once again, natural legislative settings did not supply much definitive evidence one way or the other. Even if scholars could point to a significant rule change – for example, the change in the Senate cloture rule in 1975 – and even if that change coincided with a change in the pattern of legislation, it was impossible to sort out whether the rule change was causal, spurious, or incidental to the policy change. One research agenda that followed from Arrowian social choice was to examine the effect of rules themselves, while holding preferences constant.

Procedural Rules: Structure-Induced Equilibrium

The institutional approach was kicked off by Shepsle (1979), who initiated a florescence of theory about institutions as constraints on majority rule instability. For example, Shepsle argued that germaneness rules, which limited voting to one dimension at a time, would induce a structure-induced equilibrium located at the issue-by-issue median.

McKelvey and Ordeshook (1984) ran experiments showing that issue-by-issue voting does not seem to constrain outcomes to the proposed structure-induced equilibrium, as long as subjects can communicate openly. In Figure 25-2, Player 5 is the median voter in the X dimension, as is Player 4 in the Y dimension. The results indicate a good deal of logrolling, for instance by the 1, 2, and 5 coalition, that pulls outcomes away from the structure-induced equilibrium. They conclude that theorists who “seek to uncover the effects of procedural rules and institutional constraints must take cognizance of incentives and opportunities for people to disregard those rules and constraints” (201). The germaneness rule does not seem a sturdy source of majority rule stability.

[Figure 25-2 here]

Forward and Backward Agendas.

Wilson (1986; 2008b) ran experiments on a different procedural variation – forward- vs. backward-moving agendas. A forward-moving agenda considers the first proposal against the status quo, and then the second alternative against the winner of the first vote, and so on. Each new proposal is voted on against the most recent winner. Presumably, the first successful proposal will be in the *winset* of the status quo, where the winset of X is the set of alternatives that defeat X by majority rule. A core has an empty winset, but when there is no core, every alternative has a nonempty winset. The winset of the status quo is the propeller-shaped figure shown in Figure 25-3.

An alternative is a backward-moving agenda, in which alternatives are voted on in backwards order from the order in which they were proposed. If alternatives 1, 2, and 3 are proposed in that order, then the first vote is between 2 and 3, with the winner against 1, and the

winner of that against the status quo. With this agenda, the final outcome should be either the status quo, or an alternative in the winset of the status quo. Theoretically, a backward-moving agenda is more constrained – more predictable – than a forward-moving agenda.

Figure 25-3 shows one typical voting trajectory for each treatment. The soft gray line shows a typical forward-moving agenda. The first proposal was in the win set of the status quo, backed by voters 2, 3, and 4. Subsequent moves were supported by coalitions 3, 4, and 5, then 1, 2, and 5 and then 2, 3, and 4 to restore the first successful proposal, and complete a cycle. A forward-moving agenda did nothing to constrain majority rule instability.

[Figure 25-3 here]

The dark line shows that the first alternative introduced was not in the win set of the status quo, so the final vote resulted in the imposition of the status quo. This could have been avoided with strategic voting by Player 5 on the penultimate step, leaving the committee with an outcome closer to five's ideal point than the status quo.

Overall, Wilson reports that eight of twelve experiments run with the backward-moving agenda treatment were at the initial status quo, and the other four trials were in the winset of the status quo. This contrasted sharply with the forward-moving agenda, which never ended at the original status quo, and which frequently cycled through the policy space.

The conclusion is that forward-moving agendas do not constrain majority rule instability, or provide the leverage necessary for accurate prediction. On the other hand, the backward-moving agenda is an institution that does effectively constrain majority rule.

Monopoly Agenda Control.

In simple majority rule, every majority coalition has the power and motivation to move an outcome from outside its Pareto-preferred set to some point inside. No point outside the Pareto set of every majority coalition can be in equilibrium. When the Plott symmetry conditions hold, a single internal voter's ideal point is included in every majority coalition's Pareto set. Since, in general, there is no point that is internal to the Pareto sets of all decisive coalitions, there is in general no core. Instability is the result of too many decisive majority coalitions.

The rules can create stability by mandating that some majority coalitions are not decisive. For example, the rules may specify that every proposal to be considered must be approved by a single actor – the agenda monopolist. In other words, every majority coalition that does not include the agenda monopolist is not decisive.

This greatly reduces the number of decisive majority coalitions. In particular, the intersection of the Pareto sets for all decisive coalitions is guaranteed to include only one point – the agenda-setter's ideal point. As a result, the core of a game with an agenda monopolist necessarily includes the agenda-setter's ideal point.

To test the effect of this institutional feature on majority rule instability, Wilson (2008b) ran experiments with constant preferences and no simple majority rule core. In one treatment, there was an open agenda and in the other, a monopoly agenda setter. In this latter case, the agenda-setter's ideal point was the unique core. Wilson showed that the outcomes in the open agenda had high variance; the outcomes with an agenda setter had lower variance and were significantly biased toward the agenda setter's ideal point.

Figure 25-2 shows the trajectory for a typical agenda-setter experiment. The agenda-setter, Player 5, consistently plays off the coalition with 1 and 2 against the coalition with 3 and

4. The power to do so means, of course, that majority rule instability can be replaced by coherence – at the cost of making one Player a dictator.

[Figure 25-4 here]

2. Preference-Based Constraints on Majority Rule Instability

Shepsle’s original hypothesis – that institutional variations of majority rule can sharply constrain majority rule instability and allow prediction of experimental outcomes – has proven both true and of the utmost significance for studying democracy. Rules defining control over the agenda, the size of the majority, or bicameralism have all been shown to lead to an improvement in prediction accuracy.

However, the patterning of outcomes in simple majority rule experiments, as illustrated in Figure 25-1, reveals that institutional rules are a sufficient, but not necessary, condition for constraint. Experimental outcomes cluster even with simple majority rule – even without monopoly agenda control, germaneness rules, or a backward-moving agenda.

Despite the fact that McKelvey was the author of what came to be known as the “chaos” theorem, he himself was an early advocate of finding a preference-based solution concept. That is, he believed that the actions of rational voters, negotiating alternative majority coalitions to advance their own preferences, would somehow constrain majority rule to a reasonable subset of the entire policy space – without requiring the constraint of rules other than simple majority rule. With Ordeshook, McKelvey advanced the solution concept known as the “competitive solution” for simple majority rule games. By understanding coalition formation as a kind of market that established the appropriate “price” for coalitional pivots, McKelvey, Ordeshook and Winer (1978) generated predictions that worked rather well for five-person spatial games. However, the

authors gave up on the competitive solution when other experimental results, using discrete alternatives, proved to be sensitive to cardinal payoffs (McKelvey and Ordeshook 1983).

The Uncovered Set

An alternative preference-based solution concept was the uncovered set, developed in the context of discrete alternatives by Miller (1980). It is a solution concept that identifies a set of moderate outcomes in the “center” of the space of ideal points as the likely outcome of strategic voting and the coalition formation process.

Outcomes that are far from the “center” of the ideal points are certain to be *covered*, where a covered alternative B is one such that there is some alternative A that beats B, and every alternative X that beats A also beats B. If A covers B, it implies that B is a relatively unattractive alternative with a large enough winset to encompass the winset of A.ⁱⁱⁱ

An alternative is in the *uncovered set* if it is not covered by any other alternative. If D is uncovered then for every C that beats D, there is some alternative X such that D beats X and X beats C. This means that an uncovered alternative can either defeat every other alternative directly or via an intermediate alternative. The uncovered set is the set of centrist outcomes that constitute the (unstable) center of the policy space.

Early theoretical results showed that the uncovered set had several striking characteristics. For one thing, the uncovered set was shown to be a subset of the Pareto set. For another, it shrank in size as preference profiles approximated those producing a core, and collapsed to the core when the core existed (Cox 1987).

The uncovered set has proven to be of interest to noncooperative game theory as well as cooperative game theory. The reason is that, as McKelvey argued, the uncovered set contains the noncooperative equilibria arising under a variety of institutional rules.

Shepsle and Weingast (1984) proposed that “The main conclusion is that institutional arrangements, specifically mechanisms of agenda construction, impose constraints on majority outcomes” (49). McKelvey took away a quite different interpretation. In an article provocatively titled “The Institution-Free Properties of Social Choice,” McKelvey (1986) argued that if a single solution concept encompasses the equilibrium results of a variety of institutions, then the choice process is “institution free.” That is, “the actual social choice may be rather insensitive to the choice of institutional rules” (283).

In the paper, McKelvey demonstrated that a variety of distinct institutions theoretically lead to equilibrium outcomes inside the uncovered set. He confirmed the result that legislative voting under a known, fixed agenda should lead inside the uncovered set. Cooperative coalition formation should lead to outcomes in the uncovered set, as should two-candidate elections. Hence, McKelvey could argue, constraint on simple majority rule instability seemed to be “institution-free” – the ideal points of voters provide enough information to predict where outcomes should end up, even without knowing exactly which of the three institutions would be used to select the outcome.

The problem was that neither McKelvey nor anyone else knew exactly how much the uncovered set constrained majority rule decision making, because no one had a way to characterize the uncovered set for a given set of preferences.

Looking Backward with the Uncovered Set

The recent invention of an algorithm for precise estimation of the uncovered set (Bianco, Jeliaskov, and Sened 2004) has allowed the testing of that solution concept against previously reported experimental results (Bianco et al. 2006), and with new data (Bianco et al. 2008). Figure 25-1 is a case in point. Figure 25-1 shows the Fiorina-Plott noncore experiments. The uncovered set for their experimental configuration of preferences is shown as the small shaded region. In Figure 25-1, the uncovered set is a relatively precise and promising predictor of the noncore experiments. The same is true for the uncovered set shown (as a gray shaded region) for the McKelvey-Ordeshook experiments on germaneness and communication – nearly all of the outcomes were in the uncovered set (see Figure 25-2). For the McKelvey-Ordeshook experiments, with different proposal rules and different degrees of constraint on communication, the uncovered set performs equally well.

We can do the same with other majority rule experiments run in two-dimensional policy space with simple majority rule. The results for a series of simple majority rule experiments are shown Table 25-1. Out of 272 total majority rule experiments, administered by 8 different teams of experimentalists, ninety-three percent were in the uncovered set.

[Table 25-1 here]

Testing the Uncovered Set

While the results in Table 25-1 are noteworthy, the experiments reported there were not designed to test the uncovered set. In particular, several of these experiments typically imposed maximal dispersion of ideal points, resulting in quite large uncovered sets – perhaps an “easy test” of the uncovered set. Consequently, Bianco et al. (2008) designed computer-mediated five

person majority rule experiments with two treatments creating relatively small and nonoverlapping uncovered sets – designed to be a difficult case for the uncovered case.

The two treatments were based on two configurations of preferences, shown in Figures 25-5a and 25-5b. In each case, the preferences were “clustered” rather than maximally dispersed; this had the effect of producing smaller uncovered sets. Configurations 1 and 2 are identical except for the location of Player 1’s ideal point. In Configuration 1, Player 1 was clustered with Player 4 and 5; in Configuration 2, Player 1 was in an even tighter majority cluster with Players 2 and 3. The change in Player 1’s ideal point shifted the uncovered set dramatically.

The alternative hypothesis is what may be called the partisan hypothesis, based on the obvious clustering of ideal points. Poole and Rosenthal (1997), Bianco and Sened (2005), and others have estimated the preferences of real-world legislatures – finding that they are organized in two partisan clusters. So the differences between the two configurations could be thought of as a shift of majority party control with a change in representation of district 1. The members of the majority cluster in either configuration could easily and quickly pick an alternative within the convex hull of their three ideal points and, resisting the attempts by the members of the minority cluster, vote to adjourn.

It is worth noting that the uncovered set in this setting is primarily located between the Pareto sets for the majority and minority parties, and thus will only occur if there is a significant amount of cross – partisan coalition formation and no party solidarity. In other words, in Configuration 1, if Players 4 and 5 can offer Player 3 an outcome that is more attractive than that offered by Players 1 and 2, then the uncovered set has a chance of being realized. But if Player 3 (for example) refuses offers especially made to move her away from her “natural” allies, then the

outcome should be well within the Pareto set of the partisan coalition, rather than in the uncovered set.

Figure 25-5a shows a sample committee trajectory for Configuration 1. As can be seen, there was a great deal of majority rule instability. A variety of coalitions formed, including coalitions across clusters. However, the instability was constrained by the borders of the uncovered set. Despite frequent successful moves to outcomes close to the contract curve between Players 1 and 3, Players 4 and 5 were repeatedly able to pull the outcome modestly in their direction by offering Player 3 more than Player 1 had offered.

[Figure 25-5a here]

Configuration 2 is more difficult; any outcome in the Pareto set of the tight cluster of 1, 4, and 5 is very attractive to these three voters – making it hard for 2 and 3 to offer proposals that will break up the 1-4-5 coalition. Yet even here, Players 2 and 3 occasionally make proposals that attract support from members of the majority cluster. This tends to pull outcomes out of the 1-2-3 Pareto triangle toward the minority cluster. The result is cycling within the smaller uncovered set.

[Figure 25-5b here]

Twenty-eight experiments were done with each treatment. Figure 25-6a shows the final outcome in the twenty-eight Configuration 1 experiments. The percentage of final outcomes in the uncovered was 100 percent.

[Figure 25-6a here]

Figure 25-6b shows the final outcomes in the twenty-eight Configuration 2 experiments. In four committees, the outcome seemed to be influenced by fairness considerations.

In seven of the committees, the opposite occurred – the partisan 1-4-5 coalition formed and imposed an outcome in their Pareto triangle but outside of the uncovered set. In either case, the presence of an extremely tight cluster of three ideal points seemed to decrease the likelihood of the kind of multilateral coalition formation that could pull outcomes into the uncovered set. Overall, the proportion of Configuration 2 outcomes in the uncovered set was still 60.7 percent.

[Figure 25-6b here]

While either fairness considerations or partisan solidarity can result in outcomes outside of the uncovered set, it seems fair to say that, as long as the coalition formation process is cross-partisan and vigorous, the outcome will likely be within the uncovered set. Overall, the uncovered set experiments suggest that the majority rule coalition formation process does constrain outcomes as argued by McKelvey (1986). Even more importantly, outcomes tend to converge to centrist, compromise outcomes.

3. The Challenges and Opportunities for Further Research

While the past generation of majority rule experiments has largely tested either an institutional or preference-based effect on majority rule outcomes, the McKelvey (1986) hypothesis offers a research agenda that involves both institutions and preferences, both noncooperative and cooperative game theory.

The McKelvey challenge: Endogenous Agendas in Legislatures

In the two decades since McKelvey wrote his paper on “Institution-Free Properties of Social Choice,” scholarly research on legislative institutions has flowered, especially with the aid of noncooperative game theory (e.g. Baron and Ferejohn 1989). However, little of that research has served to respond to McKelvey’s challenge to examine whether the equilibria of

noncooperative games associated with particular institutional rules in fact are located in the uncovered set.

One institution McKelvey was interested in was that in which amendments are generated by an open amendment process from the floor, in the absence of complete information about how the amendments might be ordered or what additional motions might arise. The proposal stage would be followed by a voting stage in which all the voters would know the agenda. Viewing this institution as an n -person, noncooperative game, the equilibrium should be contained in the uncovered set as long as voters vote sophisticatedly.

We know that voters sometimes make mistakes, i.e. fail to vote in a sophisticated manner (Wilson 2008). So the outcome of such endogenous agenda institutions is an open question for experimental research. Given McKelvey's result, there are three logical possibilities: 1) outcomes will be at the noncooperative equilibrium (and therefore in the uncovered set), 2) outcomes will be in the uncovered set but not at the noncooperative equilibrium, or 3) outcomes will be outside the uncovered set (and therefore not at the noncooperative equilibrium).

Connection to the Psychological Literature on Negotiation. There is a large and established psychological literature on negotiation, touching on the effect of such factors as risk preferences, cognitive biases, trust, egalitarian norms, cultural considerations, and ethical considerations. Since implementation of majority rule ultimately boils down to negotiating majority coalitions, it is important to begin to incorporate insights from that literature into the design of majority rule experiments. For example, the core is a cooperative solution concept that assumes a contract enforcement mechanism, which is uniformly lacking in majority rule experiments. Why does the core work so well in experiments that uniformly lack any contract enforcement mechanism? One

answer is suggested by Bottom et al. (1996). In this experiment, examining an institution of decentralized agenda control, getting to the core from some status quos required forming and then renegeing on a coalition – actions that many subjects were unwilling to undertake. Groups were “constrained by a complicated set of social norms that prevents the frictionless coalition formation and dissolution assumed by cooperative game theory” (Bottom 1996, 318). The net result is that informal social processes may substitute for formal contract enforcement, resulting in experimental support for cooperative solution concepts like the core and the uncovered set.

Fairness and Other Nonordinal Considerations

The Fiorina-Plott experiments were designed in such a way that subject payoffs fell off very quickly from ideal points. As a result, there was no single outcome that would give three voters a significant payoff; at least one majority coalition voter had to vote for an outcome that yielded only pennies. And there was certainly no outcome that could provide a lucrative payoff for all five voters.

In one sense, this was a difficult test for the core. It proved a good predictor even though it did not create a gleeful majority coalition. However, it also raised the question of whether the choice of the core was sensitive to changes in cardinal payoffs that left the ordinal payoffs unchanged. Eavey (1991) ran simple majority rule experiments with the same ordinal preferences as in Fiorina and Plott, but Eavey constructed less steep payoff gradients for the voters to the west, creating a benign Rawlsian alternative to the east; that is, the point that maximized the payoff of the worst-paid voter lay east of the core, and gave all five members of the committee a moderate payoff. While the attraction of the core was still apparent (Grelak and Koford 1997), the new cardinal payoffs tended to pull outcomes in the direction of the fair point,

as participants in these face-to-face committees seemed to value outcomes supported by supermajorities, rather than a minimal winning coalition. Further research is needed to explore the sensitivity of computer-mediated experiments to cardinal payoffs. Understanding the degree of sensitivity to cardinal values is potentially important for evaluating our ability to control subjects induced valuation of alternatives.

Parallelism. One challenge facing students of majority rule has been persistently voiced since Fiorina and Plott (1978). Their defense of experiments was grounded in an acknowledged need for *parallel* experimental and field research: “we reject the suggestion that the laboratory can replace creative field researchers” (576). Since that time, parallelism in research has been advocated a good many more times than it has been attempted.

The recent development of techniques for estimating the spatial preferences of real world legislators, using roll call data offers the prospect of parallel research using laboratory and real world data. An ideal point estimation method called “agenda-constrained” estimation (Clinton and Meirowitz 2004; Jeong 2008) relies on the knowledge of the agenda and legislative records on roll call votes on amendments; with this information, they obtain estimates of both legislative preferences and the outcomes they are. This information is just what is needed to test the uncovered set with real legislative data.

For example, Figure 25-7 shows estimates of senators’ ideal points, and a trajectory of winning outcomes, using Senate voting on 109 roll call amendments for the Civil Rights Act of 1964. The figure also shows the estimated uncovered set given the locations of the senators over the two key dimensions of the bill – scope and enforcement. The uncovered set lies between the cluster of ideal points of the strongest civil rights supporters and the strongest opponents. As in

the laboratory experiments, the senators created a variety of cross-cluster coalitions; the civil rights opponents repeatedly tried to weaken the enforcement or limit the scope by picking off the weakest supporters in one dimension or the other (Jeong et al 2009).

The coalitional negotiations in the Senate was much like that in experiments: new coalitions were formed to propose and vote on new amendments, and as these succeeded or failed, coalitional negotiations continued to generate yet more amendments. The administration bill, as modified by the House, was located in the uncovered set. An amendment to guarantee a trial by jury for those state and local officials found in contempt for their opposition to civil rights was popular enough to generate a majority coalition that moved the bill to location B to the left of the civil rights bill. A leadership substitute form of the bill was much stricter in enforcement at point C, but a weakening amendment protecting southern officials from double jeopardy brought the location of the bill back inside the uncovered set, where it remained despite a slight weakening of scope. The second to last vote pitted the administration bill as amended against the leadership substitute as amended; the final vote ran the leadership substitute against the perceived status quo. The final bill, located at E, was well within the uncovered set.

What does the data in Figure 25-7 suggest for an integrated research agenda involving both Senate data and experiments? One possibility is that preferences estimated from real world legislators on actual legislation may be replicated in the laboratory; a unique legislative history can potentially be repeated many times over. The debate on the civil rights bill can be replayed by inducing preferences similar to those of the senators to see if a similar outcome occurs. We can find out whether, given the preferences of legislators, the outcome was in some sense inevitable or if a dispersion of final outcomes could have been the basis for alternative histories.

Modifications in real world preferences can be examined to examine counterfactuals such as: What would have happened to this bill if Midwestern Republicans had been less supportive of the civil rights act? Or, could the bill have been passed if Tennessee's senators had been more opposed?

Or, the same preferences can be examined under different institutional rules to examine what might have occurred if the legislature had operated under a different set of rules. What if the Senate had used a different agenda procedure, or had enacted the 1975 cloture reform before 1964?

[Figure 25-7 here]

4. Conclusion

Experimental research has to some extent substantiated the concern with majority rule instability. As Wilson (2008b) noted, given appropriate institutions, “voting cycles, rather than being rare events, are common” (887). Given an open, forward agenda, and minimally diverse preferences, cycles can be readily observed.

Nevertheless, experiments have also shown that cycles are contained within the uncovered set, and can be tamed by institutional rules and procedures. There is a place for more theoretical endeavors and further experimental research on ideological (spatial) decision making – both in the lab and in parallel fieldwork on questions generated by experimental research. Indeed, the results of majority rule experiments have both informed the political science debate about the meaning and limits of majority rule (McKelvey and Ordeshook 1990, 99-144, and guided theorists as they seek explanations for both the observed instabilities of majority rule and the observed constraints on that instability. And if, as McKelvey hypothesized, a variety of

institutional rules can only manipulate outcomes within the uncovered set, then the degree to which behind-the-scenes agenda setters can manipulate the outcome of majority rule processes is itself limited.

References

- Bianco, William T., and Itai Sened. 2005. "Uncovering Conditional Party Government: Reassessing the Evidence for Party Influence in Congress and State Legislatures." *American Political Science Review* 99: 361-72.
- Bianco, William T., Ivan Jeliaskov, and Itai Sened. 2004. "The Uncovered Set and the Limits of Legislative Action." *Political Analysis* 12: 256-76.
- Bianco, William T., Michael S. Lynch, Gary J. Miller, and Itai Sened. 2006. "'Theory Waiting to Be Discovered and Used': A Reanalysis of Canonical Experiments on Majority Rule Decision-Making." *Journal of Politics* 68: 837-50.
- Bianco, W., Michael S. Lynch, Gary Miller, and Itai Sened. 2008. "The Constrained Instability of Majority Rule: Experiments on the Robustness of the Uncovered Set." *Political Analysis* 16: 115-37.
- Bottom, William, Cheryl Eavey, and Gary Miller. 1996. "Getting to the Core: Coalitional Integrity as a Constraint on the Power of Agenda Setters." *Journal of Conflict Resolution* 40: 298-319.
- Clinton, Joshua D., and Adam Meirowitz. 2001. "Agenda Constrained Ideal Points and the Spatial Voting Models." *Political Analysis* 9: 242-59.
- Cox, Gary W. 1987. "The Uncovered Set and the Core." *American Journal of Political Science* 31: 408-22.
- Druckman, James N. Donald Green, James Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100: 627-36.
- Eavey, Cheryl L. 1991. "Patterns of Distribution in Spatial Games." *Rationality and Society* 3: 450-74.
- Endersby, James W. 1993. "Rules of Method and Rules of Conduct: An Experimental Study on Two Types of Procedure and Committee Behavior." *Journal of Politics* 55: 218-36.

- Fiorina, Morris P., and Charles R. Plott. 1978. "Committee Decisions under Majority Rule: An Experimental Study." *American Political Science Review* 72: 575-98.
- Grelak, Eric, and Kenneth Koford. 1997. "A Re-examination of the Fiorina-Plott and Eavey Voting Experiments." *Journal of Economic Behavior and Organization* 32: 571-89.
- Jeong, Gyung-Ho. 2008. "Testing the Predictions of the Multidimensional Spatial Voting Model with Roll Call Data." *Political Analysis* 16: 179-96.
- Jeong, Gyung-Ho, Gary Miller, and Itai Sened. 2009. "Closing the Deal: Negotiating Civil Rights Legislation." *American Political Science Review* 103: 588-606.
- McKelvey, Richard D. 1976. "Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control." *Journal of Economic Theory* 12: 472-82.
- McKelvey, Richard D. 1986. "Covering, Dominance and Institution Free Properties of Social Choice." *American Journal of Political Science* 30: 283-314.
- McKelvey, Richard D., and Peter C. Ordeshook. 1983. "Some Experimental Results That Fail to Support the Competitive Solution." *Public Choice* 40: 281-91.
- McKelvey, Richard D., and Peter C. Ordeshook. 1984. "An Experimental Study of the Effects of Procedural Rules on Committee Behavior." *Journal of Politics* 46: 182-205.
- McKelvey, Richard D., and Peter C. Ordeshook. 1990. "A Decade of Experimental Research in Spatial Models of Elections and Committees." In *Advances in the Spatial Theory of Voting*, eds. James M. Enelow, and Melvin J. Hinich. Cambridge: Cambridge University Press.
- McKelvey, Richard D., Peter C. Ordeshook, and Mark D. Winer. 1978. "The Competitive Solution for N-Person Games Without Transferable Utility, With an Application to Committee Games." *American Political Science Review* 72: 599-615.
- Miller, Nicholas. 1980. "A New Solution Set for Tournament and Majority Voting." *American Journal of Political Science* 24: 68-96.
- Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll-Call Voting*. New York: Oxford University Press.
- Riker, William H. 1982. *Liberalism Against Populism*. San Francisco: Freeman.
- Riker, William H. 1986. *The Art of Political Manipulation*. New Haven: Yale University Press.
- Shepsle, Kenneth A. 1979. "Institutional Arrangements and Equilibria in Multidimensional Voting Models." *American Journal of Political Science* 23: 27-59.

- Shepsle, Kenneth A., and Barry R. Weingast. 1984. "Uncovered Sets and Sophisticated Voting Outcomes with Implications for Agenda Institutions." *American Journal of Political Science* 28: 49-74.
- Wilkerson, John. 1999. "'Killer' Amendments in Congress." *American Political Science Review* 93: 535-52.
- Wilson, Richard. 1986. "Forward and Backward Agenda Procedures: Committee Experiments on Structurally Induced Equilibrium." *Journal of Politics* 48: 390-409.
- Wilson, Richard. 2008a. "Endogenous Properties of Equilibrium and Disequilibrium in Spatial Committee Games." In *Handbook of Experimental Economics Results*, eds. Charles R. Plott, and Vernon Smith. Amsterdam: North Holland.
- Wilson, Richard. 2008b. "Structure-Induced Equilibrium in Spatial Committee Games." In *Handbook of Experimental Economics Results*, eds. Charles R. Plott, and Vernon Smith. Amsterdam: North Holland.

Table 25-1. Testing the Uncovered Set with Previous Majority Rule Experiments - Bianco et al. (2006)

Table 1. Summary of the Uncovered Set's Predictive Power			
ARTICLE	EXPERIMENT	TOTAL OUTCOMES	% IN UNCOVERED SET
Fiorina and Plott	Series 3	15	80.00%
McKelvey, Ordeshook, and Winer	Competitive Solution	8	100.00%
Laing and Olmsted	A ₂ -The Bear	19	89.47%
	B-Two Insiders	19	94.74%
	C ₁ -House	19	73.68%
	C ₂ -Skewed Star	18	83.33%
McKelvey and Ordeshook	PH-Closed Communication	17	100.00%
	PH-Open Communication	16	93.75%
	PHR-Closed Communication	15	100.00%
	PHR-Open Communication	18	100.00%
Endersby	PH-Closed Rule	10	100.00%
	Closed Communication		
	PH-Open Rule	10	100.00%
	Closed Communication		
	PH-Open Rule	10	100.00%
	Open Communication		
	PHR-Closed Rule	10	100.00%
	Closed Communication		
	PHR-Open Rule	10	100.00%
	Closed Communication		
PHR-Open Rule	10	100.00%	
Open Communication			
Wilson	Forward Agenda	12	91.67%
	Backward Agenda	12	100.00%
Wilson and Herzberg	Simple Majority Rule	18	94.44%
King	Non-Voting Chair	6	100.00%
TOTAL		272	93.75%

Figure 25-1. Outcomes of Majority Rule Experiments without a Core - Fiorina and Plott (1978)

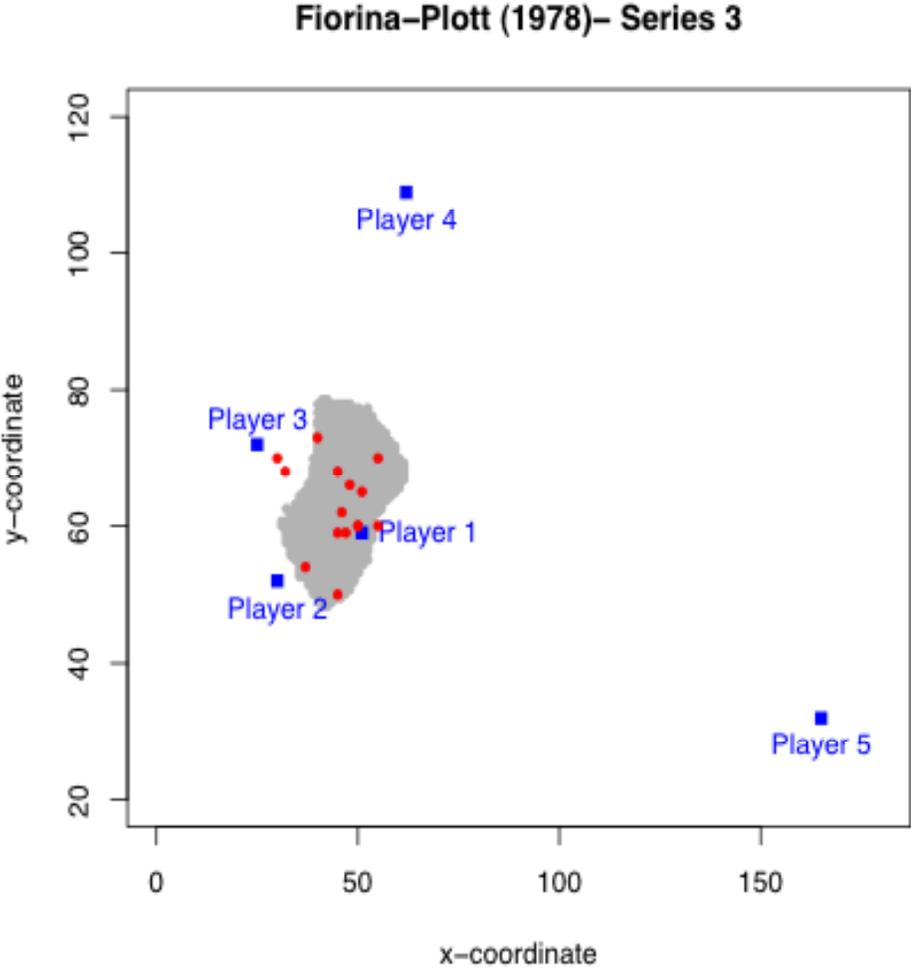


Figure 25-2. Majority Rule with Issue-by-Issue Voting - McKelvey and Ordeshook (1984)

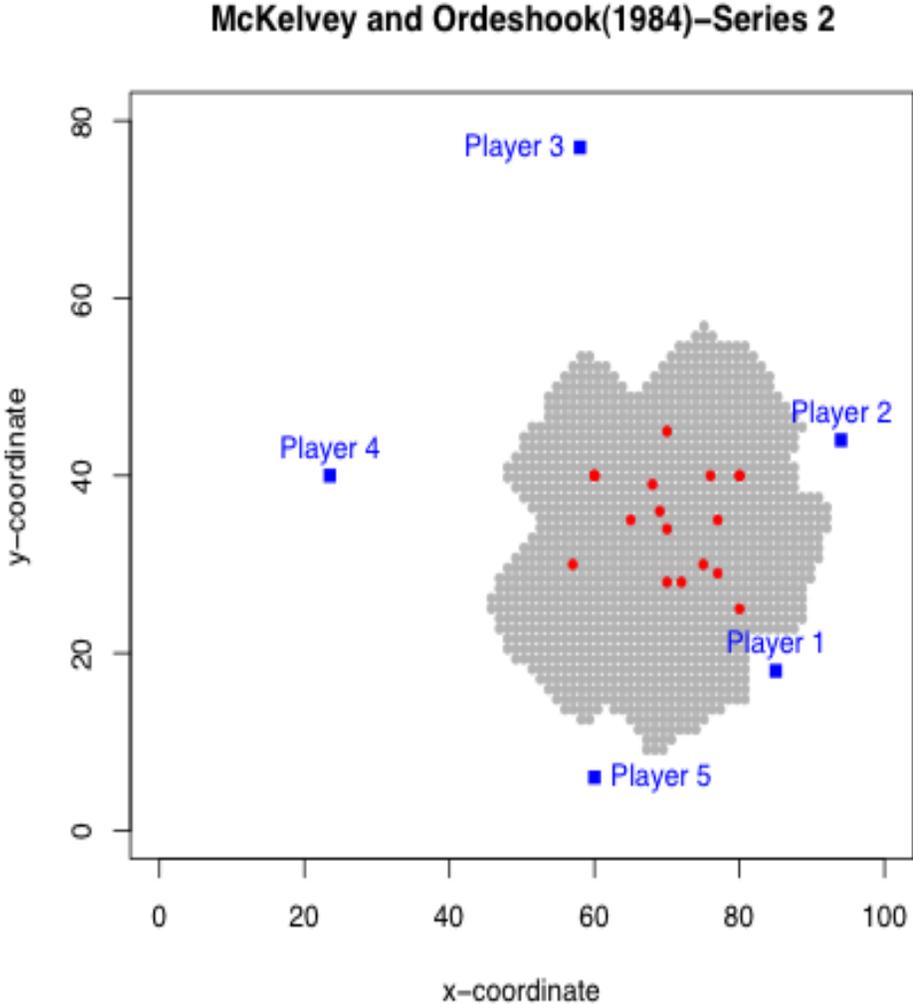


Figure 25-3. The Effect of Backward and Forward Agendas - Wilson (2008b)

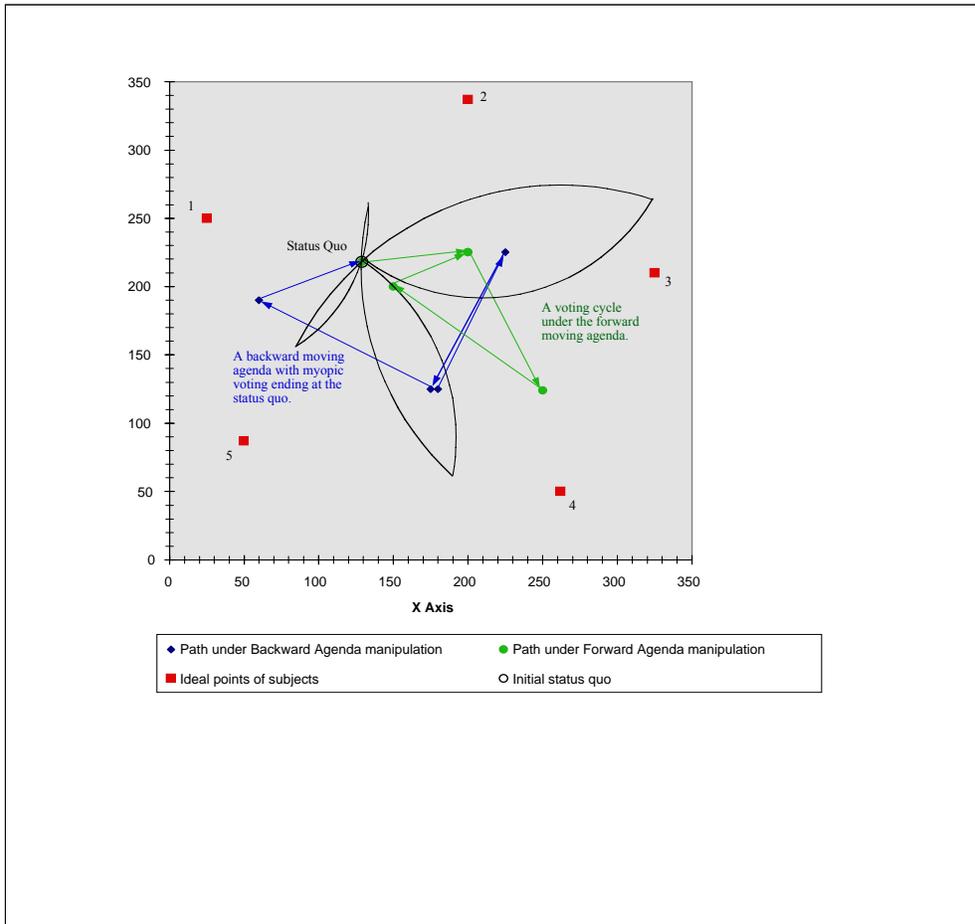


Figure 25-4. The Effect of Monopoly Agenda Setting - Wilson (2008b)

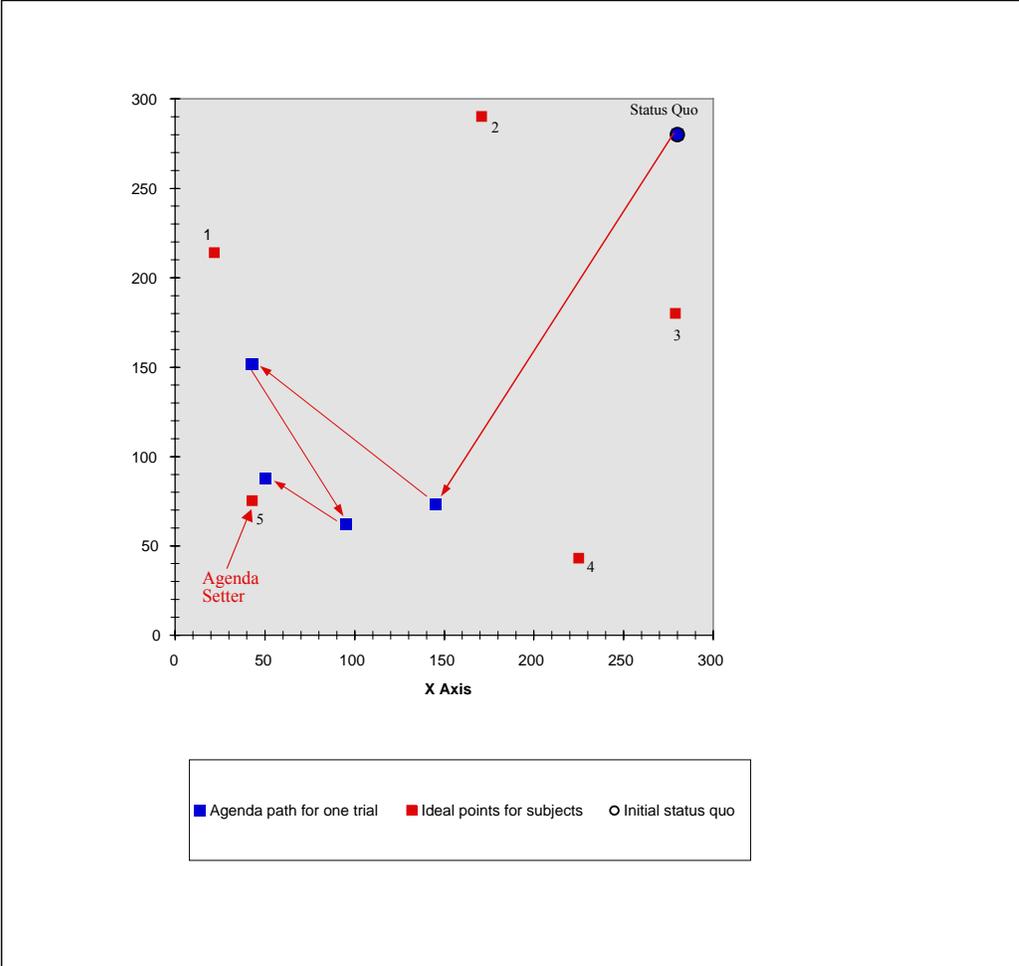


Figure 25-5a. A Sample Majority-Rule Trajectory for Configuration 1 - Bianco et al. (2008)

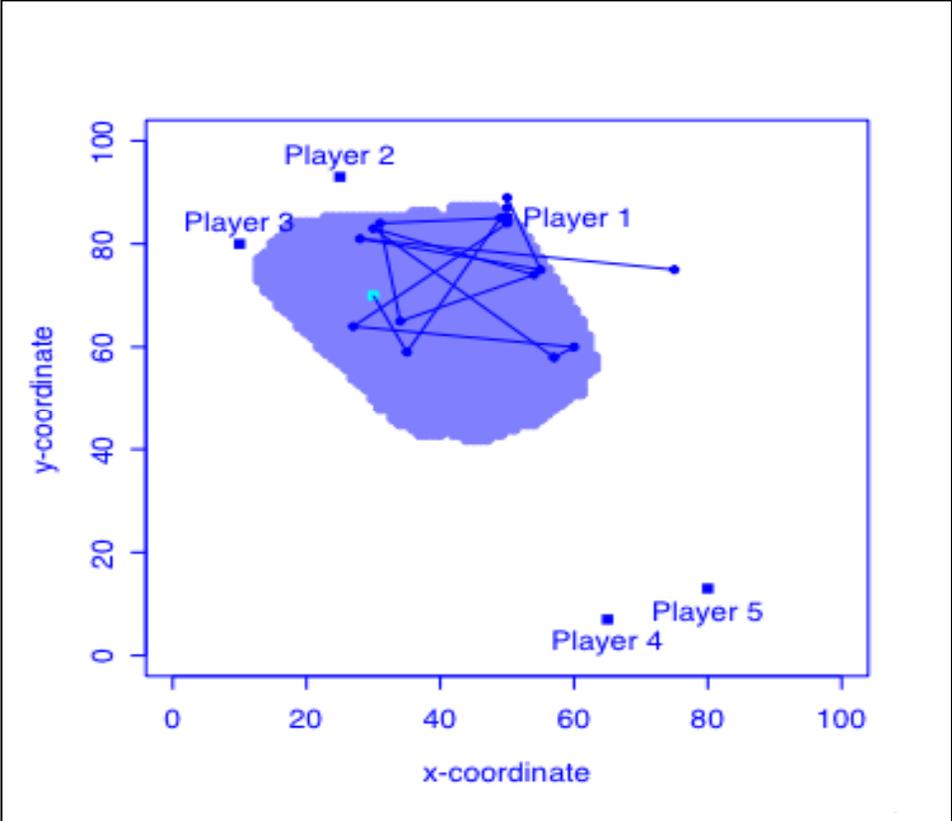


Figure 25-5b. A Sample Majority-Rule Trajectory for Configuration 1 - Bianco et al. (2008)

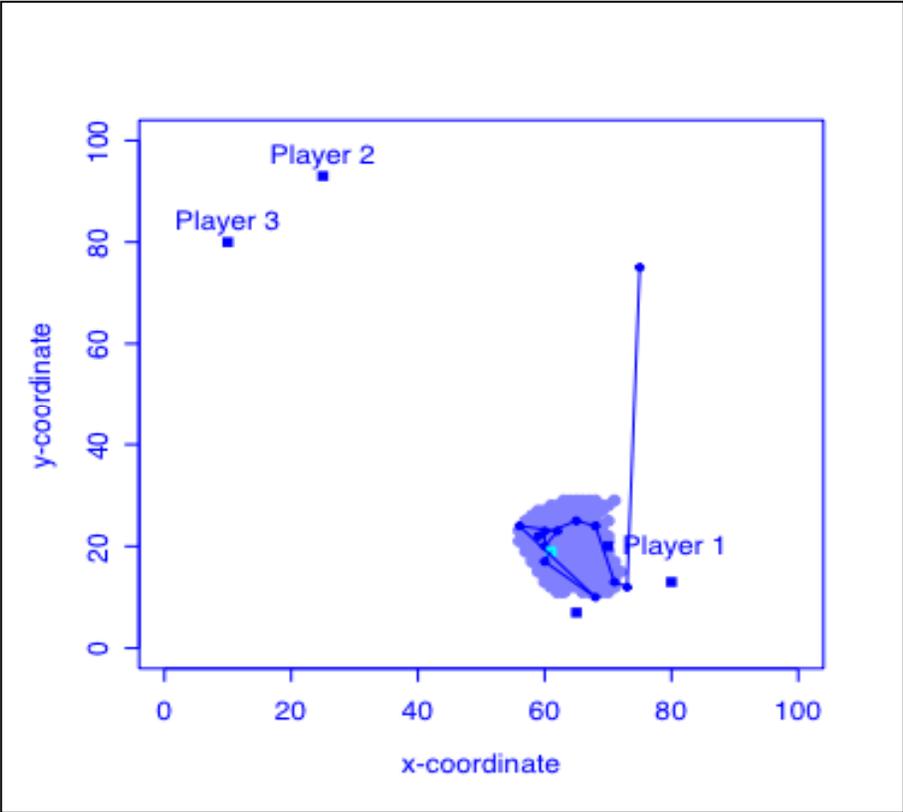


Figure 25-6a. The Uncovered Set and Outcomes for Configuration 1 - Bianco et al. (2008)

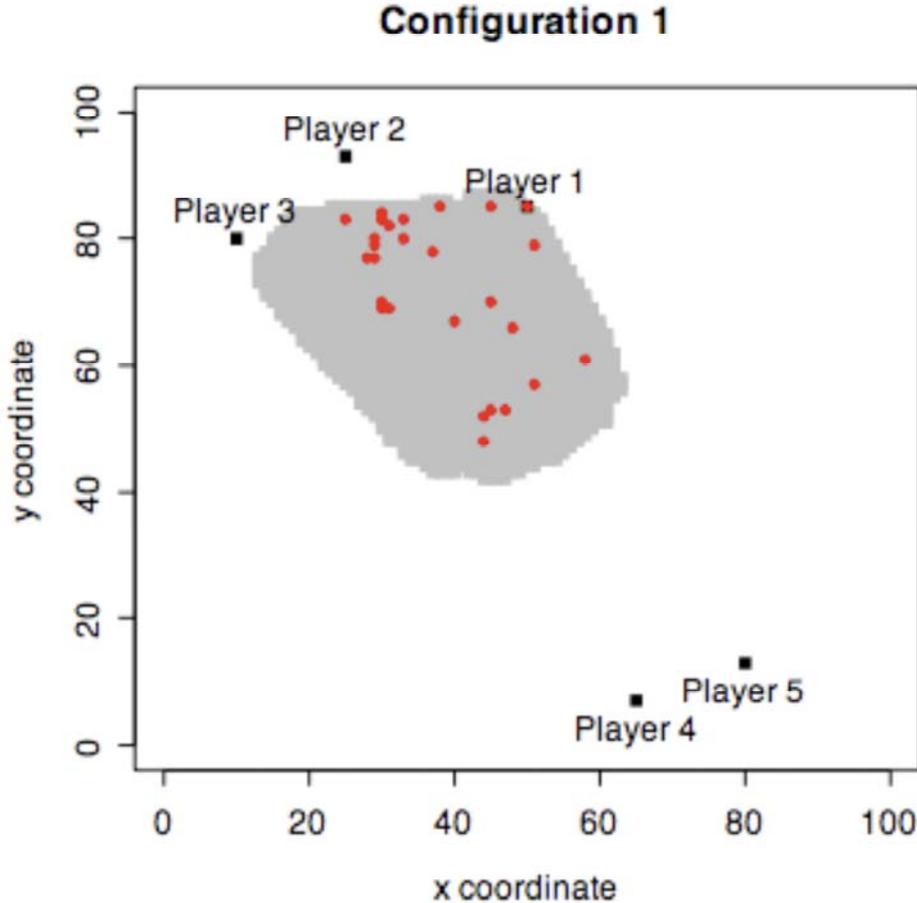


Figure 25-6b. The Uncovered Set and Outcomes for Configuration 2 - Bianco et al. (2008)

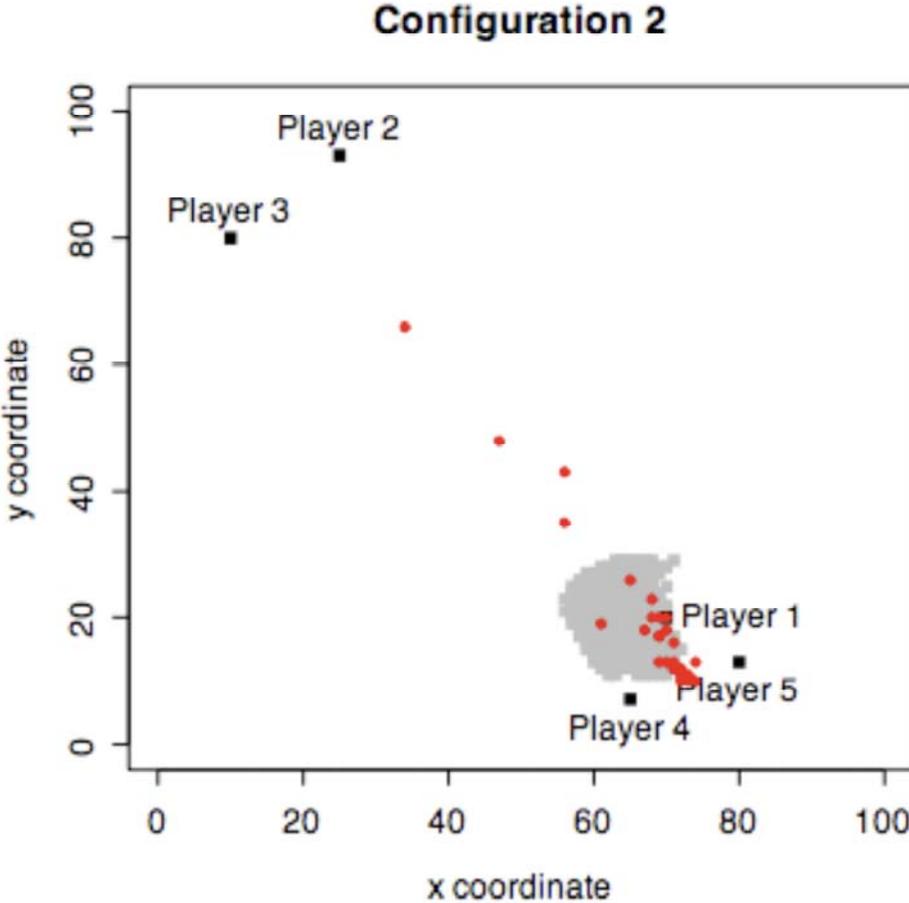
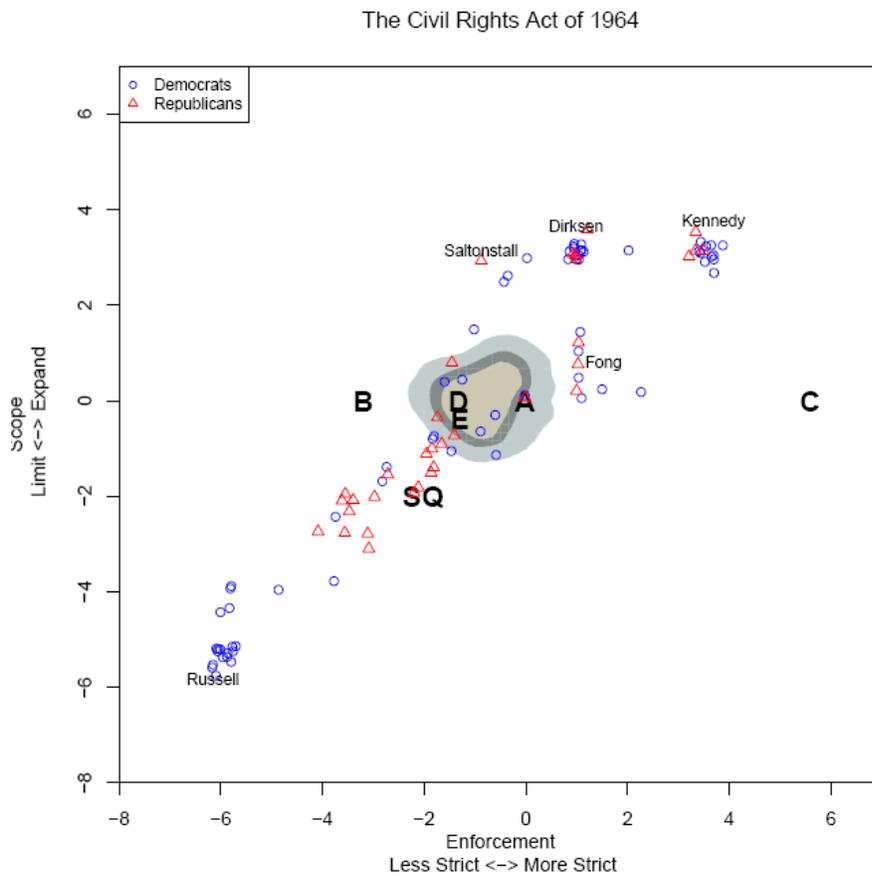


Figure 25-7. Senatorial Ideal Points and Proposed Amendments for the Civil Rights Act of 1964 - Jeong et al. (2009)



ⁱ Thanks to William Bottom, Jamie Druckman, Raymond Duch, Dima Galkin, Gyung-Ho Jeong, Yanna Krupnikov, Arthur Lupia, Michael Lynch, Itai Sened, Jennifer Nicoll Victor, Robert Victor, and Rick Wilson.

ⁱⁱ The gray shaded area will be explained later in the paper.

ⁱⁱⁱ More centrist outcomes have smaller win sets.

26. Electoral Systems and Strategic Voting (Laboratory Election Experiments)

Rebecca B. Morton and Kenneth C. Williams

It can be complicated to attempt to understand how election mechanisms and other variables surrounding an election determine outcomes. This is because the variables of interest are often intertwined so it is difficult to disentangle them to determine the cause and effect that variables have on each other. Formal models of elections are used to disentangle variables so that cause and effect can be isolated. Laboratory election experiments are conducted so that the causes and effects of these isolated variables from these formal models can be empirically measured. These types of experiments are conducted within a single location where it is possible for a researcher to control many of the variables of the election environment and thus observe the behavior of subjects under different electoral situations. The elections are often carried out in computer laboratories via computer terminals and the communication between the researcher and subjects occurs primarily through a computer interface. In these experiments, subjects are assigned as either voters or candidates, and in some cases both roles. Voters are rewarded based upon a utility function that assigns a preference for a particular candidate or party. Candidates are typically rewarded based on whether they win the election but sometimes their rewards depend on the actions they take after the election.

By randomizing subjects to different treatments and controlling many exogenous variables, a laboratory election experiment is able to establish causality between the variables of interest. For example, if a researcher is interested in how different types of information (such as reading a newspaper editorial versus reading a blog on an internet site) impact voting decisions,

then in the laboratory it is possible to control all the parameters in the experiment, such as voter preferences for different candidates, parties or issues, while only varying the types of information voters receive. Randomizing subjects to two different types of information treatments (for example, editorial versus blog) allows the researcher to establish if one type of information (editorial) causes voters to behave differently than other voters who were given a different type of information (blog) under the same electoral conditions and choices. Hence, laboratory election experiments are concerned with controlling aspects of an election environment and randomizing subjects to treatments in order to determine causality of the election variable(s) a researcher has selected. Field experiments also allow researchers to vary variables, such as types of information, but they do not allow for the control of other factors, such as voter preferences over candidates, parties or issues (see Gerber's chapter in this volume). As opposed to field experiments, the laboratory allows a researcher to control all aspects of the election environment.

Laboratory election experiments started in the early 1980s and were directly derived from committee experiments, since the researchers working on some of these committee experiments, McKelvey and Ordeshook, began conducting the first laboratory election experiments (for a review of committee experiments, see Miller's chapter in this volume). Similar procedures used in the committee experiments were carried over to test competitive elections where the substantive questions that led to laboratory election experiments were questions relating to the median voter theorem (Downs 1957). To illustrate this theory, consider Figure 26-1. In this case, there is a single policy dimension where three voters have single peaked utility functions or preferences over the dimension (the y axis). The property of single peakedness of utility

functions ensures that each voter has a unique ideal point over the policy dimension. For each voter, a dashed line represents their ideal point or their best policy (where they receive the highest payoff on the x axis); each voter prefers policies closer to their ideal point than policies further away.

[Figure 26-1 about here]

In this simple theory, it is assumed that two candidates compete in an election by adopting policies on the single dimension and voters vote for candidates who adopt policies closer to their own ideal points. To see this, consider that if Candidate 1 adopts position 10 and Candidate 2 adopts a position of 60, then voter 1 will vote for Candidate 1 but voters 2 and 3 will vote for Candidate 2 since her position is closer to their ideal points than is Candidate 1's position. Since both candidates know the distribution of voter ideal points, then both candidates realize that the optimal location for placement of positions is at voter's 2 ideal point, or the median voter that guarantees the candidates of attaining a tie.

The theorem shows the importance of a central tendency in a single dimension election. However, the theorem relies on restrictive assumptions such that voters have full information about candidate positions, candidates know the distribution of voter preferences, and there is no abstention. When the theorem is extended to two dimensions, then no equilibrium exists unless other rigid restrictions are made.

Experiments first considered the assumption that voters possess full information about candidate positions, which was challenged by empirical findings that voters were relatively uninformed about the policies or even the names of the candidates (Berelson, Lazarsfeld, and McPhee 1954; Almond and Verba 1963; Converse 1975). The purpose of these early

experiments was to determine whether the general results would still hold if the information assumption were relaxed. Laboratory experiments were an ideal way to study this question, since it was possible to replicate the assumptions of the model in an experimental setting allowing for variables, such as the information levels that voters have about candidates, to be relaxed. Hence, causality could be directly measured by determining if candidate positions converged when voters possessed only incomplete information about candidate positions. We will discuss the results of these types of experiments later in this chapter.

1. Methodology

In terms of incentives, laboratory election experiments pay subjects based on their performance during the experiment. For a fuller discussion of the use of financial incentives, see Dickson's chapter in this volume. Financial incentives allow experimenters to operationalize a monetary utility function that establishes performance-based incentives for subjects within the experimental environment. This procedure follows the principals of Induced Value Theory (see Smith 1976), which essentially means that payments awarded during the experiment must be salient in order to motivate subjects to choose as if the situation or election were natural. That is, within the election environment, voter and candidate subjects must feel that there are real consequences to their actions.

Another aspect of laboratory elections is that a number of repeated trials or elections are conducted within a single experimental session. Usually in a typical election, voters are assigned a type (or a preference for a particular candidate), but in the next election period the voter is randomly reassigned another type and exposed to a different treatment and so on until the completion of the experimental session. This is referred to as a "within-subject treatment" design

that allows researchers to vary treatments (election parameters) while holding subject identities constant. One reason that voters are randomly assigned different types for each election period is to avoid repeated game effects. This means that the researcher does not want subjects to view each election as a function of the last election, but rather they want subjects to think they are participating in a single or new election each period. Randomly assigning subjects different types each period and varying treatments does in fact create a new electoral environment for subjects at each period since they have different decisions to make. Also, by having a number of elections, it allows the researcher to observe learning effects. In some experiments subjects need experience with the election environment to figure out equilibrium behavior, or what is the optimal decision to make given the environment and this can often take a number of rounds. Hence, in most analyses, researchers look at the behavior of subjects in the beginning, middle and end of the experiment to determine learning effects. Finally, conducting multiple elections during a single experimental session simply gives the researcher more data to analyze.

Again, one of the advantages of using laboratory experiments is that it is easy to measure causality about subject behavior within the election environment when different treatments are compared and subjects are randomly assigned to treatments. Consequently, experimental controls and randomization of treatments allow for easy establishment of causality of the variables of interest without the use of sophisticated statistical methods. Laboratory election experiments allow the researcher to examine electoral phenomena that observational studies cannot, since there is simply no data or very little data such as the electoral properties of election rules that have not been instituted or used very little in real-world elections. Also researchers who rely on observational data are constrained by the number of election occurrences, whereas in the

laboratory, election researchers can conduct hundreds of elections under various manipulations. Finally, and more importantly, laboratory elections allow the researcher to play the role of God over the election environment since manipulations are generally easy to induce in a laboratory.

In this chapter, we present laboratory election experiments from a wide range of approaches in which we hope to show that results from experimental elections have, like results from observational data, provided findings deemed to be fruitful for the body of literature on election behavior and electoral mechanisms. First, we will discuss the early election experiments that were concerned with testing the robustness of the median voter theorem. We continue with an examination of experiments on theories that explain candidate policy divergence in elections. We then examine experiments on multi-candidate elections and the coordination problem that is involved with strategic voting, and we discuss experiments on sequential elections in which different sources of information are relayed to early and late voters in the voting queue. Finally, we present experiments on voter turnout. Although we have attempted to provide a comprehensive review of laboratory experimental work on elections, due to space constraints we are not able to discuss all of the experimental work in detail. We encourage readers to explore this literature further.

2. The Literature on Laboratory Election Experiments

Tests of the Median Voter Theorem

The first significant laboratory election experiments were conducted by McKelvey and Ordeshook (1985a, 1985b), in which they relaxed the full information condition of the median voter theorem.¹ These experiments are what is referred to as a “stress test” that examines how a model’s results hold when some of its assumptions are relaxed. McKelvey and Ordeshook tested

a rational expectations theory of markets, which dictates that information is aggregated so that informed traders transmit information about the price of commodities to uninformed traders and, as a result, the market behaves as if it were fully informed. In applying this theory to elections, they designed an experiment in which there were informed and uninformed voters who were divided into three groups and had ideal points on three locations on a single dimensional space. Uninformed voters only knew the location of their ideal points, which voter groups were to the left and right of their ideal points, and an interest group endorsement which specified which candidate was furthest left. Informed voters knew the precise location of candidates' positions. Candidate subjects did not know the location of the median voter's ideal point.

Prior to the election, a poll was conducted that revealed the percentage of subjects within a group who reported favoring either candidate. With these two pieces of information (the interest group endorsements and the poll results), uninformed voters, in the theoretical equilibrium, behaved or voted like informed voters. In the experiment, there was significant evidence in support of the rational expectations equilibrium, in which uninformed voters behaved as if they were informed eighty percent of the time. The experimental results also found that, although candidate subjects did not know the median voter's ideal point, their positions converged near the median after a few rounds of the experiment. As discussed earlier, these experimental results were possible to observe primarily because of the control that the researchers were able to exert over voter information. Such control would not have been possible in observable elections or in a field experiment on an election.

A retrospective voting experiment conducted by Collier et al. (1987) severely limited the information of voters and candidates as well. In this experiment, voter subjects were assigned an

ideal point over a single dimension but, unlike the previous experiment, they did not know the location of their or other voters' ideal points. Candidates also did not know the distribution of the voters' ideal points. During the experiment, two subjects posed as candidates and one of the candidate subjects was randomly chosen to be an initial incumbent. This subject then selected a position on the issue space, which translated into a monetary payoff for the voter subjects based on the location of their ideal point relative to the incumbent's position. The amount of their payoffs was revealed to voters and they voted either to retain the incumbent or to vote for the challenger. Between twenty-three and forty-five elections were conducted.

Some sessions used a within-subjects design, in that they shifted the voter ideal points by thirty-five units in round twenty-one in order to determine whether candidates could track the shift in the median. The results showed that candidate positions did converge to the median voter in both the nonshift and shift experiments. Hence, these experiments show that candidate subjects largely located the median position even when they had no information on its location. And, although voters did not know candidate positions, based on retrospective evaluations, they learned within the first ten periods how to identify the candidate who was closest to their ideal points.

In a variant of this experiment, Collier, Ordeshook, and Williams (1989) allowed voter subjects the option of purchasing information about the challenger's proposed position. One hypothesis tested was that when the electoral environment is stable, such that policies enacted are invariant from election to election, voters will rely on retrospective cues, but when the political environment is unstable – i.e., when policies vary from election to election -- voters will invest in information to discover the policy of the challenger. Again, the researchers used a

within-subject design. In order to create systems of electoral stability and instability, “dummy” candidates were used where candidate positions remained constant for a fixed number of periods and fluctuated for a fixed number of periods.² Again the research question was whether voters can track shifting candidates’ positions and under what conditions will voters purchase information (stable versus unstable environments). The experiments showed that when the electoral system was stable, voters tended to purchase less information and they relied more on retrospective evaluations, and when the electoral system was unstable they tended to purchase more information about the incumbent’s positions. In real elections it is difficult to measure the costs of being informed for individuals, since these costs vary by individual, but in the laboratory this cost can be explicitly measured.

The experiments discussed in this section revealed, as the pundits have observed in observational elections, that the tendency toward the median distribution of voter ideal points is a powerful pull in electoral politics. Although the median voter theorem espoused by Downs was criticized for its simplicity, the experiments showed that this theorem holds up when its restrictive conditions have been relaxed, surviving a number of tests that the laboratory allowed researchers the ability to implement. Even when voters are uninformed about the exact location of candidate positions and candidates do not know the exact distribution of voter ideal points, the median is still a magnet for electoral outcomes. Although observational data can point to a central tendency of candidate positions in elections, they cannot determine how this pull to the median is affected by voter information, since this variable cannot be controlled, measured, or randomized in real-world elections.

Models of Candidate Divergence

The experiments we have discussed emphasize the robustness of the pull of the median for candidates in elections. Although these tendencies exist, it is also well documented that there are policy differences in candidates and parties; a number of formal models and experimental tests have been proposed to explain these differences. One of the first formal models to explore why candidates might diverge in policy positions is Wittman's (1977) model of candidates with divergent policy preferences. Calvert (1985) demonstrates that for such policy preferences to lead to divergence of candidates' positions in elections, candidates must also be uncertain about the ideal point of the median voter. Morton (1991) presents an experimental test of Calvert's proposition. In her experiment, subjects were assigned as candidates and their payments depended solely on the policy position chosen by the winning candidate. In the treatment with incomplete information, the median voter's ideal point was a random draw each period. This treatment was compared to one in which the median voter's ideal point was constant across periods. Morton finds that indeed, as Calvert predicts, candidates choose more divergent positions when the median voter ideal point is randomly determined in each period, but converge in positions when the median voter ideal point is fixed.

Morton's experiment illustrates how laboratory elections can provide tests of theories that are nearly impossible using observable data or field experiments. The comparative static prediction would be difficult, if not impossible, to evaluate in observational elections since a researcher would only be able to estimate the knowledge of candidates and parties about voter preferences (as well as only be able to estimate those preferences him or herself) and is unlikely to have exogenous variation in that knowledge. Furthermore, Morton's experiment also illustrates how laboratory experiments can provide researchers with new, unexpected results that

are not observable if we only rely on observable studies and field experiments. That is, Morton finds that subjects converge more than is theoretically predicted in the uncertainty condition, suggesting that subjects value winning independent of their payoffs. Given the ability in the laboratory to control both the information subjects had and their monetary or financial payoffs from the election, it was possible to discern that the subjects received intrinsic, nonmonetary payoffs from winning, something that would be extremely tough to observe outside of the lab.

An alternative explanation for candidate divergence is the existence of valence quality differences between candidates (differences that are independent of policy positions), posited in Londregan and Romer (1993), Groseclose (2001), and Aragonés and Palfrey (2004). Aragonés and Palfrey present experimental tests of a theory that suggests that candidates who are perceived by voters to have lower valence quality advantages will adopt more extreme positions and that, similar to Calvert's results, the more uncertain the position of the median voter, the greater the divergence of the candidates in policy positions. Aragonés and Palfrey find that these theoretical predictions are supported in the laboratory elections.³ Furthermore, they demonstrate that the results are robust to variations in subject pools (the experiments were conducted both at the California Institute of Technology in Pasadena, California and Universitat Pompeu Fabra in Barcelona, Spain) and in framing (some treatments at Caltech used a game framing where subjects made choices in a payoff matrix and others used a political context). Thus, they not only establish that the theory is supported in the laboratory but that their results are robust to a number of external validity tests.

Related to the experimental research on theories of why candidates might diverge in candidate policy positions is the work of Houser and Stratmann (2008), who consider how

candidates with fixed, divergent policy positions but uncertain quality differences might use campaign advertisements to convey information on their qualities to voters, testing a theory of campaign contributions posited by Coate (2004). In their experiment, subjects are both voters and candidates. Candidates are rewarded solely based on whether they win an election and different schemes for paying for campaign advertisements are considered (that is, sometimes campaign advertisements reduce voter payoffs if the advertiser wins, as if the campaign spending was funding by special interests, and other times campaign advertisements are not costly to voters, although always costly to candidates). Note that they use a within-subject design so they are able to measure the effects of the different finance schemes, while controlling for specific subject effects.

Houser and Stratmann find that indeed candidates use advertisements to convey policy information and that voters use that information to choose candidates who provide them with greater payoffs. Furthermore, they find that less information is provided to voters when campaign contributions are costly in terms of voter payoffs. Houser and Stratmann, because they can control and randomize how campaign contributions directly affect voter payoffs, provide useful information on how campaign advertisements financed by interest groups might affect voter information, voter choices, and electoral outcomes. Such control and randomization is largely impossible outside of the laboratory. Furthermore, a within-subject design, which allows for the ability to control for subject specific effects, cannot be done outside of the laboratory.

Strategic Voting and Coordination in Three-Candidate Elections.

Strategic voting in three-candidate elections means that a voter realizes that her most preferred candidate will lose the election so she votes for her second preferred candidate to

prevent her least preferred candidate from winning the election. The problem is that a substantial number of other voters in the same situation must also decide to vote strategically so that their least preferred candidate does not win. Hence, it is a coordination problem among voters whereby they figure out that supporting their second most preferred candidate leads to a better outcome, i.e., their least preferred candidate is not elected.

Forsythe et al. (1993) consider experimentally the voter coordination problem in strategic voting situations using an example from the theoretical work of Myerson and Weber (1993). They present a simple example that illustrates the problem voters confront, as detailed in Table 26-1. In the table there are three types of voter preferences where Type 1 prefers A to B to C, since the monetary amount associated with each candidate is higher. Type 2 voters prefer B to A to C, and Type 3 voters prefer C and are indifferent between B and A (since the monetary amounts are the same). Also assume that there are four Type 1 voters, four Type 2 voters but six Type 3 voters.

[Table 26-1 about here]

In this configuration, candidate C is a Condorcet loser since he would lose in a two-way contest against either of the other candidates. However, if each type voted their sincere preference in a plurality election, then candidate C, the Condorcet loser, would be the winning candidate. The problem is how can Type 1 and Type 2 voters coordinate their votes so as to prevent candidate C, their least preferred candidate, from winning? Should Type 1 and Type 2 voters vote strategically for candidates B or A? Without a coordination device, subjects in the Type 1 and 2 groups behave poorly and fail to coordinate and the Condorcet loser wins about 87.5 percent of the time.

Three types of coordination devices were instituted in the experiments: polls, history of past elections, and ballot location. Polls in the experiment were implemented by allowing subjects to vote in a nonbinding election where the results were revealed to all subjects, and then a binding election took place. The theory of Myerson and Weber provided no predictions as to whether any of these coordination devices would work or whether one would be more successful than the others. Thus, the experiments provided new information for our understanding of voter coordination. The researchers found that with the use of polls the Condorcet loser only won 33 percent of the time. Also when either A or B was leading in the poll results the Condorcet loser only won 16 percent of the time. The experiment also found a small bandwagon effect, in which the candidate who won in a past election garnered more support and there was also a small ballot location effect. Thus, the researchers found that polls were only weak coordination devices, albeit stronger than the other alternatives.

Using a similar design, Rietz, Myerson, and Weber (1998) consider another coordination device, campaign contributions where subjects can purchase ads for candidates. They find that campaign contributions are more successful than polls as a coordination device. Note that, unlike the Houser and Stratmann experiments, the ads provided no information to other voters about voter payoffs, but were merely ways in which voters could attempt to coordinate before an election. Morton and Rietz (2008) analyze majority requirements as coordination devices. In contrast to the other devices, majority requirements theoretically should result in full coordination since the requirements theoretically eliminate the equilibria where the Condorcet loser wins. They find that, indeed, majority requirements are far more effective as coordination devices than the others studied.

Forsythe et al. (1996) conducted a related experiment using the same preference profile that replicates the “Condorcet loser” paradox but they varied the voting rules. They altered three alternative voting rules: plurality, approval voting, and the Borda count. Under the plurality rule, voters voted for just one candidate; under approval voting, subjects voted for as many of the candidates as they wanted to; and under the Borda rule, voters were required to give two votes to their two most preferred candidates and one vote to their second preferred candidate. While there were multiple equilibria, in general: under the plurality rule, they predicted the Condorcet loser would win more elections; under approval voting, the Condorcet loser would win less often; and under the Borda rule, the Condorcet loser would lose even more elections. The results generally support the equilibrium predictions. Hence, the Borda and approval rules were more efficient in defeating the Condorcet loser. The ability to hold preferences of voters over candidates constant and to vary the electoral rule provided researchers with the opportunity to gain causal information on the effects of these rules, which would not be possible using observational data or field experiments.

Gerber, Morton and Rietz (1998) used a similar candidate profile to examine cumulative versus straight voting in multi-member districts (i.e., three candidates compete for two seats). In their setup there are three voter types with the following preferences: Type 1 voters (4) prefer A to B to C, Type 2 voters (4) prefer B to A to C, while Type 3 voters (6 are the minority type) only get utility from C. Hence, like the experiment previously described, candidate C is the Condorcet loser. In straight voting, voters can cast one vote for up to two candidates, while in cumulative voting they can cast two votes for one candidate. Voters can also abstain. They find that minority representation increases with the use of cumulative voting, since minority voters

can cumulate their votes on the minority candidate. Again, the use of the laboratory allows for the researchers to make comparisons, holding voter preferences constant; again, this is not possible in observational data or in field experiments.

These experiments illustrate how, by using laboratory elections, it is possible to vary different aspects of an election -- for example, whether there is a poll or not, whether campaign contributions are allowed, whether majority requirements are instituted, and whether alternative voting systems are instituted -- and thus measure the causal effects of these different aspects, again something not possible using observational data or field experiments. Real-world elections have not provided sufficient observational data on these electoral mechanisms, yet laboratory elections are able to test the efficiencies of these mechanisms.

Sequential Voting

Morton and Williams (1999, 2001) examine multi-candidate elections to test the difference between simultaneous and sequential voting in terms of their impact on voter behavior and electoral outcomes. Simultaneous voting represents the American general presidential elections where voters vote once and the results are revealed and a winning candidate is announced. Sequential voting represents American presidential primaries where voting takes place over time in stages and the winning candidate is the one who accumulates the most votes over the stages. The two main hypotheses Morton and Williams test are whether sequential elections give candidates who are well known to voters an electoral advantage, and whether sequential elections lead to more informed voter decisions. The voters in the experiment have preferences over three candidates, where Group x has ten members and prefers x to y to z, Group z has ten members and prefers z to y to x, and Group y has four members and prefers y and is

indifferent between x and z . In this case, y is the Condorcet winner and will beat the other candidates in a pairwise competition, but will lose if all voters vote sincerely, in which x and z tie.

In the simultaneous treatment, the identity of one candidate was revealed to voters prior to casting of votes. This was an incomplete information treatment, in which one candidate is better known to voters. In the sequential treatment, voters were divided into two Groups, A and B, where Group A voted first, followed by Group B, and then the results were tallied across the two groups. There were treatments that varied the level of information that voters had but the primary treatment involved letting Group A voters know the identity of the candidate and Group B would have poll results or they would know how Group A voted. The results show that later voters, or Group B voters, were able to use poll results to make more informed decisions that reflected their preferences. The researchers also find that under certain conditions when the Condorcet winner, y , is unknown, this candidate does better under sequential voting than under simultaneous voting. The laboratory provides Morton and Williams a unique opportunity to compare the effects of these two voting systems, simultaneous and sequential, on the election outcomes, and to compare voter information about their choices. Given that presidential elections occur only every four years and there is considerable variation over time in the candidates, voters, and countless other factors that confound such a comparison in observational data, the laboratory experiments provide information that cannot be learned otherwise.

In a related setup that is a blend of Morton and Williams and the Forsythe et al. experiments, Dasgupta et al. (2008) examine coordinated voting in a five-person sequential voting game. In this experiment, five subjects vote sequentially, and each subject knows how

each subject prior to them has voted (the first voter does not have any vote information and the last voter knows how everyone else has voted). In this experiment, voters maximize their payoff when they vote for the alternative that garners a majority of the votes. This experiment examines situations where a voter is concerned with voting for the majority preferred alternative, or what is referred to as conformity voting (Coleman 2004). Subjects were randomly assigned one of two types (they either preferred Green or Red). Subjects in the experiment were shown the following payoff matrix.

[Table 26-2 about here]

This is a payoff matrix for subjects who were randomly assigned to be a Green type. Notice that if a subject is randomly assigned to be a Green type, then if she thinks the Green candidate will win the election she should vote for the Green candidate since this will yield her \$1.50 for this election period. However, if she thinks Green will lose the election, then she should vote for the Red candidate and receive \$1.00 instead of receiving nothing when she votes for the Green candidate and the Green candidate loses.

Two information conditions were varied; one where subjects only knew their type, and the other where they knew the type of all other subjects. For full information, the equilibrium prediction is that all voters who are in the majority should vote for the most preferred candidate. The equilibrium prediction for the incomplete information treatment depends on the order in which types are revealed, but generally voters later in the voting queue should switch their votes more often (i.e., we should see more strategic voting.) The results generally confirm the equilibrium predictions. As in the Morton and Williams experiment (1991, 2001), these results show that, in elections where sequential choice is the chosen voting rule, such as in European

Union elections, then voters at different positions in the voting queue will be voting with different available information and this can have an effect on electoral outcomes.

These aforementioned experiments on sequential voting elections do not allow for abstention. Battaglini, Morton, and Palfrey (2007) examine the difference between sequential and simultaneous voting with two choices under uncertainty when there is a cost to voting and abstention is allowed. In the experiment they conduct, there are two voting cost treatments: a low and high cost environment. In this experiment, the researchers used a nonpolitical frame, unlike those previously discussed. That is, in this case, subjects were presented with two jars (via a computer image) where one jar contained six red balls and two blue balls, and the other jar contained six blue balls and two red balls. The subjects then selected one ball from one of the jars, which was only revealed to that subject. Subjects then guessed which jar was the correct one; they could guess Jar 1, Jar 2, or abstain. In the simultaneous treatment, subjects all guessed or abstained together, and in the sequential treatment, subjects were assigned a position (first, second, or third) and they guessed in that order. If a majority of the subjects guessed for the correct jar, then all subjects received a payoff minus the cost of voting.

The theory predicted that, in simultaneous voting, the probability of abstention decreases with the cost of voting. In sequential voting under low cost of voting, early voters should bear the cost and later voters should vote as if their vote is pivotal, whereas with high costs, later voters will be forced to bear the cost. In terms of equity of the electoral mechanism, they show that simultaneous voting is more equitable, as opposed to sequential elections, since all voters derive the same expected utility, although sequential voting may lead to higher aggregate payoffs and be more economically efficient. As predicted, the researchers find that abstention increases

with voting costs in simultaneous voting. In the case of sequential voting, the researchers find that the later voters are advantaged by having more information to determine when to abstain. However, they also find that subjects often make choices at variance with the theoretical predictions, voting more often than theoretically predicted in the sequential voting elections. Again, these experiments illustrate the effect of sequential elections in terms of voter turnout. Since sequential elections are not that common, laboratory election experiments are able to illustrate that these types of elections do impact voting and turnout behavior.

Turnout Experiments

One of the riddles in the voting literature is the basic question of why people vote. The voting paradox posits that in large elections the probability of one vote affecting the outcome is relatively small, almost zero, so that if there is a cost to voting, then actually going to the polls to vote will outweigh the benefits (i.e., being a pivotal vote) and, therefore, it is not rational to vote; yet, we observe large number of voters participating in elections. In an early attempt to address this paradox, Ledyard (1984) and Palfrey and Rosenthal (1985) noted that the probability that a vote is pivotal is endogenously determined and that, when candidates have fixed policy positions, equilibria exist with positive turnout and purely rational voter decision making. This is because, if everyone assumed that the probability of being pivotal was smaller than the cost of voting and, consequently, no one voted, then the probability of being pivotal for any one voter would be 100 percent, since that one voter would determine the outcome. Thus, when electorates are finite and candidate positions are fixed, endogeneity of pivotality means that equilibria with positive voting are possible.

Levine and Palfrey (2007) conducted the first direct experimental tests of the Palfrey Rosenthal (1985) model. They find that the comparative statics of the theory are supported in the laboratory – specifically, that turnout decreases when the electorate increases (size effect), turnout is higher when the election is close (closeness effect), and that supporters of minority candidates turn out in greater numbers than do those of majority candidates (an underdog effect). Levine and Palfrey also find that turnout is higher than predicted in the large elections (and smaller than predicted in the small elections), which they explain to be a consequence of voter errors. Levine and Palfrey’s experiments demonstrate the advantage of the laboratory, in that they can both control and manipulate a number of important theoretical independent variables that are difficult to control or measure in the field, such as voter costs, electorate size, and size of the majority, while holding voter preferences constant.

An alternative model of voter turnout is that voters respond to group influences and are not purely individually motivated as formulated in Morton (1987, 1991), Uhlaner (1989), and Schram and Van Winden (1991). The Morton and Uhlaner models view turnout decisions as responses to group leader manipulations, while Schram and Van Winden posit that groups have a psychological impact on how individuals vote. Feddersen and Sandroni (2006) provide a micromodel of how groups might influence turnout, in which some voters turn out even when their votes are not likely to be pivotal because of ethical concerns.

A number of experimentalists have considered the group turnout models. For example, Schram and Sonnemans (1996a, b) and Grosser and Schram (2006) present experimental evidence on the impact that psychological influence and contact with other voters has on voters’ turnout decisions. Gailmard, Feddersen, and Sandroni (2009) provide an interesting experimental

test of the ethical voter model in that they vary both the probability that a subject's vote is decisive in the outcome and the benefit to other voters from voting. They find support for the argument that ethnical motives might explain turnout decisions. These results may also partially explain the tendency of excessive voting in the large elections found by Levine and Palfrey.

The models we have thus far discussed examine turnout decisions when voting is costly. However, voters often abstain even when voting is costless, for example when voters choose to not vote in some races while in the voting booth. Feddersen and Pesendorfer (1996) create a model of elections between two alternatives, one in which all the voters have the same preferences (common preferences) but are asymmetrically informed and one in which some voters are more informed than other voters. They show that when the poorly informed voters are indifferent between the alternatives, in equilibrium it is optimal for these voters to abstain and let the more informed voters vote, since their misguided choices may sway the election in the wrong direction. Consequently, even if there is no cost to voting, it is rational for the poorly informed to forgo voting and delegate the decision to the more informed voters. This phenomenon is referred to as the "Swing Voter Curse" (SVC). One corollary of this theory is that if these uninformed voters know that some voters are partisans and will vote for their favored candidates regardless of the information, then the uninformed voters will abstain less often and choose to vote in order to cancel out the votes of the partisans, even if doing so is contrary to their prior information about which choice is best for them. To understand the intuition of this result, suppose that there are two options before voters, a and b , and two states of the world, A and B . A set of voters (swing voters) prefers option a in state of the world A and option b in state of the world B . Some of the swing voters are informed and know for sure the true state of the world, while others are

uninformed and only have probabilistic information about the true state of the world. Their prior information is that state of the world A is more likely than state of the world B . However, there are also a group of partisan voters who will always vote for option a regardless of the state of the world. In the event that an uninformed voter is pivotal, then it is likely that all informed voters are voting for b since partisans vote for a . Thus, uninformed voters have an incentive to offset the partisan votes and vote for b as well, even though their prior information suggests that the state of the world is A .

Battaglini, Morton, and Palfrey (2009) test this theory using a similar procedure as in their experiment on sequential voting (previously discussed). In this experiment, subjects were shown one of two colored jars on their computer screens and subjects voted for the jar that was randomly deemed to be the correct jar. Some voters were notified which jar was the correct jar and other voters were not informed. The results show that uninformed voters delegated their vote or abstained about 91 percent of the time. They also find evidence that when uninformed voters know partisans are voting in the election, they tend to abstain less often and to vote in order to cancel out the votes of the partisans. The results further show that turnout and margin of victory tend to increase with the number of informed voters in the election environment. It is noteworthy that uninformed voters voted to cancel out the votes of the partisans, even though doing so involved voting against their own prior information in some of the treatments. As Lassen (2005) explains, this type of nuanced voting behavior is unlikely to be measurable in observational elections even in the best of circumstances (when information is provided to voters arguably exogenously). Thus the laboratory provides an important environment for testing these sorts of precise predictions that are largely unobservable outside of the laboratory.

Morton and Tyran (2008), using an experimental design similar to that of Battaglini, Morton, and Palfrey, examine voting cases where voters vary in information quality but no voter is completely uninformed and no voter is completely informed. In the games explored, there are multiple equilibria, equilibria where only highly informed voters participate (SVC Equilibria), and equilibria where all voters participate (All Vote Equilibria). In some cases, the Pareto Optimal equilibrium (that is, the equilibrium that provides subjects with the highest expected payoffs) is SVC and in others it is All Vote, in contrast to Battaglini, Morton, and Palfrey, where SVC is always Pareto Optimal. Morton and Tyran find that the tendency of less informed voters to abstain is so strong that, even in the cases where it is Pareto Optimal for all to vote and share information, less informed voters delegate their votes to the more informed voters, letting the experts decide. They find that the tendency to delegate is so strong that even in a voting game where such delegation by all less informed voters is not an equilibrium, subjects were drawn to such behavior. These results suggest that the tendency observed by Battaglini, Morton, and Palfrey, may reflect a norm of behavior to delegate to experts that may not always be optimal. The ability to discern the importance of such a norm that the laboratory provides is yet another example of what one can learn from a laboratory election that is not generally possible using observational or field experimental data alone.

Other experimental research demonstrates that the tendency of uninformed voters to abstain when voting is costless appears to be related to experimental context and other aspects of the voting situation. Houser, Morton, and Stratmann (2008) extend the aforementioned Houser and Stratmann experiments to allow voters to abstain. Thus, their experiments consider the extent that uninformed voters abstain in a situation in which campaign information is

endogenous and may be ambiguous (that is, when the advertising is funded at a cost to voters). They find that abstention rates of uninformed voters are higher, as is theoretically predicted, but are much lower than those found by Battaglini, Morton, and Palfrey. They also find that all voters, both informed and uninformed, abstain significantly more when campaign information is costly to voters. These results provide unique information on the effects of campaign financing of information on voter turnout that is difficult to observe in observational elections or to manipulate in field experiments, since such an experiment would require manipulating how candidates finance their campaigns -- something to which we would expect only a rare viable candidate in an election to consent.

3. Conclusion

As we have illustrated, laboratory election experiments allow researchers to conduct tests of predictions from formal models, both relationship and dynamic predictions, and to conduct stress tests where our theory does not give us a guide. The laboratory provides researchers with the ability to control many aspects of an election environment enabling researchers to better measure the causal effects of different election variables, such as information, timing of voting, variations in voting rules, coordination devices such as polls, and ethical motivations on voter and candidate choices. Laboratory experiments have provided significant support for the theoretical predictions of the median voter theorem, even in situations where voter and candidate information is limited. Furthermore, laboratory experiments have allowed for researchers to investigate voting mechanisms that are difficult to study observationally because they are rarely used and their effects on individual voters are often impossible to determine.

One concern about laboratory election experiments is the costs associated with conducting these types of experiments. Because these experiments assume Induced Value Theory, subjects must be financially rewarded based on their performance in the experiment and these costs can be prohibitive for some researchers. However, there are many methods to satisfy Induced Value Theory and keep the costs low, such as paying subjects only for randomly selected rounds (see Morton and Williams 2010 for other techniques.) Programming costs must also be taken into account, since most of these types of experiments are conducted on a computer network and some software must be developed to operationalize the experiment in the laboratory. However, there is free software that allows researchers to design network-based experiments, even with little programming knowledge.⁴

Laboratory election experiments have been criticized for failing to provide externally valid claims because of the artificiality of the laboratory. However, external validity does not refer to whether the laboratory election resembles some election in the observable world, but rather to whether the results can be generalized to variations in treatments and target populations; this can only be shown empirically through replication, not through logic or supposition. In behavioral economics, experimental economists have engaged in an increasing amount of such replication, taking many of the basic experiments in economics to new target populations across the world, allowing for a better understanding of the validity of the results from these experiments.⁵ We believe that the next challenge facing experimental political scientists is to similarly consider the external validity of laboratory election experiments by conducting such experiments with new and different target populations, as in the Aragoes and Palfrey experiment (2004) discussed in this chapter. In other experimental disciplines, it is common to

engage in replication and to use such replications to conduct meta-analyses and systematic reviews to determine which results are valid across the variety of experimental treatments and target populations. It is time for political scientists to investigate whether these important results from laboratory experimental elections are externally valid the only way that such an investigation can be done – by conducting more such experiments but varying the subjects used and the treatments, as is done in other experimental disciplines.

In addition to replication issues, other future developments for laboratory election experiments are the various extensions that are applicable for the experiments discussed in this chapter. For example, varying the preferences and motivations of voters, information that voters and candidates possess about the election environment, the number of voters in the elections, and different types of electoral mechanisms are all valid research agendas. We would also like to see more collaborative projects. Currently, the number of subjects who participate in an election experiment tends to be rather small due to the costs mentioned above. But we can imagine, as a result of Internet collaboration, projects where laboratories across the nation or the world could participate in large-scale experiments and share in the costs.

References

- Almond, Gabriel, and Sidney Verba. 1963. *The Civic Culture*. Princeton, NJ: Princeton University Press.
- Aragones, Enriqueta, and Thomas R. Palfrey. 2004. “The Effect of Candidate Quality on Electoral Equilibrium: An Experimental Study.” *American Political Science Review* 90: 34-45.
- Battaglini, Marco, Rebecca B. Morton, and Thomas R. Palfrey. 2007. “Efficiency, Equity, and Timing in Voting Mechanisms.” *American Political Science Review* 101: 409-24.
- Battaglini, Marco, Rebecca B. Morton, and Thomas R. Palfrey. 2009. “The Swing Voter’s Curse in the Laboratory.” *Review of Economic Studies* 77: 61-89.

- Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting*. Chicago: University of Chicago Press.
- Calvert, Randall L. 1985. "Robustness of the Multidimensional Voting Model, Candidate Motivations, Uncertainty and Convergence" *American Journal of Political Science* 29: 69-95.
- Coate, Stephen 2004. "Pareto-Improving Campaign Finance Policy." *American Economic Review* 94: 628-55.
- Coleman, Stephen. 2004. "The Effect of Social Conformity on Collective Voting Behavior." *Political Analysis* 12: 76-96.
- Collier, Kenneth E., Richard D. McKelvey, Peter Ordeshook, and Kenneth C. Williams. 1987. "Retrospective Voting: An Experimental Study." *Public Choice* 53: 101-30.
- Collier, Kenneth E., Peter Ordeshook, and Kenneth C. Williams. 1989. "The Rationally Uninformed Electorate: Some Experimental Evidence." *Public Choice* 60: 3-39.
- Converse, Phillip E. 1975. "Public Opinion and Voting Behavior." In *Handbook of Political Science*, eds. Fred I. Greenstein, and Nelson W. Polsby. Reading, MA: Addison-Wesley.
- Dasgupta, Sugato, Kirk A. Randazzo, Reginald S. Sheehan, and Kenneth C. Williams. 2008. "Coordinated Voting in Sequential and Simultaneous Elections: Some Experimental Results." *Experimental Economics* 11: 315-35.
- Dasgupta, Sugato, Kenneth C. Williams. 2002. "A Principal-Agent Model of Elections with Novice Incumbents." *Journal of Theoretical Politics* 14: 409-38.
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.
- Feddersen Timothy, and Alvaro Sandroni. 2006. "A Theory of Participation in Elections with Ethical Voters." *American Economic Review* 96: 1271-82.
- Feddersen, Timothy J., and Wolfgang Pesendorfer. 1996. "The Swing Voter's Curse." *American Economic Review* 86: 408-24.
- Forsythe, Robert, Roger B. Myerson, Thomas Rietz, and Robert Weber. 1993. "An Experimental Study on Coordination in Multi-Candidate Elections: The Importance of Polls and Election Histories." *Social Choice and Welfare* 10: 223-47.
- Forsythe, Robert, Roger B. Myerson, Thomas Rietz, and Robert Weber. 1996. "An Experimental Study of Voting Rules and Polls in a Three-Way Election." *International Journal of Game Theory* 25: 355-83.

- Gailmard, Sean, Timothy Feddersen, and Alvaro Sandroni. 2009. "Moral Bias in Large Elections: Theory and experimental evidence." *American Political Science Review* 103: 175-92.
- Gerber, Elisabeth, Rebecca B. Morton, and Thomas Rietz. 1988. "Minority Representation in Multimember Districts." *American Political Science Review* 92: 127-44.
- Groseclose, Tim. 2001. "A Model of Candidate Location when One Candidate has a valence Advantage." *American Journal of Political Science* 45: 862-86.
- Grosser, Jens, and Arthur Schram. 2006. "Neighborhood Information Exchange and Voter Participation: An Experimental Study." *American Political Science Review* 110: 235-48.
- Houser Daniel, Rebecca B. Morton, and Thomas Stratmann. 2008. "Turned Off or Turned Out? Campaign Advertising, Information, and Voting." Working Papers 1004, George Mason University, Interdisciplinary Center for Economic Science.
- Houser Daniel, and Thomas Stratmann. 2008. "Selling Favors in the Lab: Experiments on Campaign Finance Reform." *Public Choice* 136: 215-39.
- Lassen, David 2005. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." *American Journal of Political Science* 49: 103-18.
- Ledyard, John O. 1984. "The Pure Theory of Large Two-Candidate Elections." *Public Choice* 44: 7-41.
- Levine, David K., and Thomas R. Palfrey. 2007. "The Paradox of Voter Participation: A Laboratory Study." *American Political Science Review* 101: 143-58.
- Londregan, John, and Thomas Romer. 1993. "Polarization, Incumbency, and the Personal Vote." In *Political Economy: Institutions, Competition, and Representation*, eds. William A. Barnett, Melvin J. Hinich, and Norman J. Schofield. Cambridge, UK: Cambridge University Press.
- McKelvey, Richard D., and Peter C. Ordeshook. 1982. "Two-Candidate Elections without Majority Rule Equilibria." *Simulation and Games* 3: 311-35.
- McKelvey, Richard D., and Peter C. Ordeshook. 1985a. "Elections with Limited Information: A Fulfilled Expectations Model Using Contemporaneous Poll and endorsement Data as Informational Sources." *Journal of Economic Theory* 36: 55-85.
- McKelvey, Richard D., and Peter C. Ordeshook. 1985b. "Sequential elections with Limited Information." *American Journal of Political Science* 29: 480-512.
- Morton, Rebecca B. 1987. "A Group Majority Voting Model of Public Good Provision." *Social Choice and Welfare* 4: 117-31.

- Morton, Rebecca B. 1991. "Groups in Rational Turnout models." *American Journal of Political Science* 3: 758-76.
- Morton, Rebecca B., and Thomas A. Rietz. 2008. "Majority Requirements and Minority Representation." *New York University Annual Survey of American Law*, 63: 691-726.
- Morton, Rebecca B., and Jean-Robert Tyran. 2008. "Let the Experts Decide: Asymmetric Information, Abstention and Coordination in Standing Committees." Unpublished manuscript, New York University.
- Morton, Rebecca B., and Kenneth C. Williams. 1999. "Information asymmetries and Simultaneous versus Sequential Voting." *American Political Science Review* 93: 51-67.
- Morton, Rebecca B., and Kenneth C. Williams. 2001. *Learning by Voting*. Ann Arbor: University of Michigan Press.
- Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York: Cambridge University Press.
- Myerson, Robert B., and Robert J. Weber. 1993. "A Theory of Voting Equilibria." *American Political Science Review* 87: 102-14.
- Oosterbeek, Hessel, Randolph Sloof, and Gijs Van de Kullen. 2004. "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis." *Experimental Economics* 7: 171-88.
- Palfrey, Thomas R., and Howard Rosenthal. 1985. "Voter Participation and Strategic Uncertainty." *American Political Science Review* 79: 62-78.
- Rietz, Thomas A., Robert B. Myerson, and Robert J. Weber. 1998. "Campaign Finance Levels as Coordinating Signal in Three-Way Experimental Elections." *Economics and Politics* 10: 185-217.
- Schram, Arthur, and Joep Sonnemans. 1996a. "Voter Turnout as a Participation Game: An Experimental Investigation." *International Journal of Game Theory* 25: 385-406.
- Schram, Arthur, and Joep Sonnemans. 1996b. "Why People Vote: Experimental Evidence." *Journal of Economic Psychology* 17: 417-42.
- Schram, Arthur, and Frans Winden. 1991. "Why People Vote: Free Riding and the Production and Consumption of Social Pressure." *Journal of Economic Psychology* 12: 575-620.
- Smith, Vernon L. 1976. "Experimental Economics: Induced Value Theory." *American Economic Review* 66: 274-79.

Uhlaner, Carole Jean. 1989. "Relational Goods and Participation: Incorporating Sociability into a Theory of Rational Action." *Public Choice* 62: 253-85.

Wittman, Donald. 1977. "Candidates with Policy Preferences: A Dynamic Model." *Journal of Economic Theory* 14: 180-9.

Figure 26-1. Median Voter Theorem

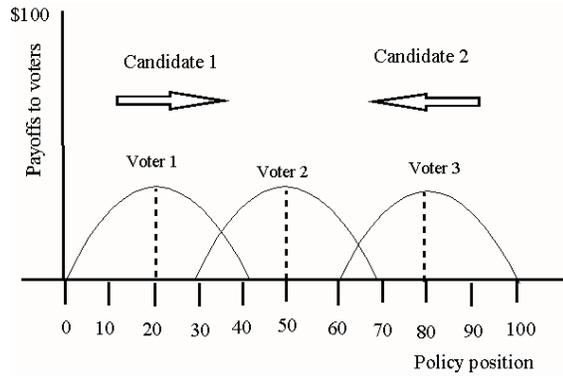


Table 26-1. Forsythe et al (1993) Payoff Schedule

<u>Voter Type</u>	<u>Election Winner</u>			<u>Total Number of Each Type</u>
	<u>A</u>	<u>B</u>	<u>C</u>	
1 (A)	\$1.20	\$0.90	\$0.20	4
2 (B)	\$0.90	\$1.20	\$0.20	4
3 (C)	\$0.40	\$0.40	\$1.40	6

Table 26-2. Dasgupta et al. (2008) Payoff Schedule

Voter Type is Green		
	G wins election	R wins election
Vote G	1.50	0.0
Vote R	0.25	1.0

¹McKelvey and Ordeshook conducted an earlier election experiment in 1982 but this experiment only concerned voters and whether they could calculate mixed strategy candidate equilibria in their minds.

²Unbeknownst to voter subjects, candidate subjects were not used, but rather a researcher manually inputted candidate positions.

³Aragones and Palfrey's experimental design builds on earlier work of Dasgupta and Williams (2002) on candidate quality differences in experiments on the principal-agent problem between voters and candidates.

⁴See <http://www.iew.uzh.ch/ztree/index.php>

⁵See, for example, the recent meta-analysis of ultimatum game experiments conducted using a variety of subject pools in Oosterbeek, et al. (2004).

27. Experimental Research on Democracy and Development

Ana L. De La O and Leonard Wantchekon

Expectations about the role of democracy in development have changed considerably in recent years. In principle, the exercise of political rights sets democracies apart from other political regimes in that voters can pressure their representatives to respond to their needs. It has been argued that such pressure “helps voters constrain the confiscatory temptations of rulers and thereby secure property rights; increases political accountability, thus reduces corruption and waste; and improves the provision of public goods essential to development” (Boix and Stokes 2003, 538). Thus, the argument follows, democracy is development-enhancing. Yet deprivations, such as malnutrition, illiteracy, and inequalities in ethnic and gender relationships have proven to be resilient, even within the nearly two-thirds of the world's countries ranked as electoral democracies. The persistence of deprivations is a reminder that there is still a great deal to be learned about the relationship between democracy and development.

Not surprisingly, scholars have explored numerous ways in which democracy can be related to development, ranging from macropolitical examinations (e.g. are democracies better at producing development than are authoritarian regimes?), to microexplanations (under what circumstances can voters limit bureaucrats' rent-seeking behavior?). Yet the bulk of empirical evidence in this respect is inconclusive (Przeworski and Limongi 1997; Boix and Stokes 2003; Keefer 2007). Is democracy a requirement for development or is it the other way around? Are formal institutions the causes or the symptoms of different levels of development? Which should come first, property rights or political competition? Civil liberties or public service provision?

Why are elections compatible with rampant corruption? As critical as these questions are to the discipline, what we know thus far is plagued by problems of simultaneous causality, spurious correlations, and unobserved selection patterns.

Recently, experimental research on the political economy of development has blossomed. Despite its novelty, progress has been rapid and continues apace. As experiments in this field have evolved, several features distinguish them from earlier empirical contributions. First, scholars have started to address central debates in the field by mapping broad theoretical issues to more specific and tractable questions (Humphreys and Weinstein 2009). For example, instead of asking how different political regimes shape development, recent studies ask whether various methods of preference aggregation produce divergent provisions of public goods. Second, unlike previous macrostudies based on cross-country regressions, recent work has focused on the subnational level. Third, researchers are increasingly making use of field experiments to study how politics affects development and how development shapes politics in developing countries.

Throughout this chapter, as in the rest of this volume, when we speak of experiments we mean research projects where the subjects under study are randomly assigned to different values of potentially causal variables (i.e., different treatment and control groups). For example, a researcher might assign one group of households to receive cash transfers and assign another group of households to receive the same cash transfers but make the latter transfer conditional on parents investing in their children's education. In some, but not all, designs there is also a control group that does not receive any treatment. As Druckman et al. explain in the introduction to this volume, random assignment means that each entity being studied has an equal chance to be in a particular treatment or control condition.

Experimentation in the field of political economy of development has taken several forms: the increasingly popular field experiments take place in a naturally occurring setting; laboratory experiments take place in a setting controlled by the researcher; laboratory experiments in the field resemble field experiments more generally, in that interventions take place in a naturally occurring setting, but researchers have more control over the setting and the treatment; survey experiments involve an intervention in the course of an opinion survey; and finally, there are some instances when interventions of theoretical interest have been randomly assigned not by researchers but by governments. We group studies that take advantage of this type of randomization in the category of natural experiments.

Because experimentation is still a novel research tool in the field, throughout this chapter we review some of the ongoing and published research projects that illustrate how random assignment is being used to tackle questions about the political economy of development. We begin Section 1 by considering examples of pioneering field experiments executed in collaboration with nongovernmental organizations (NGOs). Section 2 describes two unique field experiments done in partnership with political parties. Section 3 presents several studies that took advantage of natural experiments. Section 4 introduces the use of laboratory and laboratory in the field experiments. Section 5 discusses some of the challenges faced by researchers conducting experiments on development and democracy, such as internal and external validity, as well as ethical issues. This section also presents practical solutions to some of these challenges drawing from recent experimental designs.

In section 6, we conclude that, despite the challenges, experiments are a promising research tool that have the potential to make substantial contributions to the study of democracy

and development, not only by disentangling the causal order of different components of democracy and development, but also by providing evidence that other empirical strategies cannot produce. Moving forward, we argue that the best of the experimental work in the field of democracy and development should reflect well-chosen populations, a deep understanding of the interaction of the interventions with their contexts, and should test theoretical mechanisms such that scientific knowledge starts to accumulate.

1. Field experiments in collaboration with NGOs

Olken's (2010) study of two political mechanisms –plebiscites and meetings-- in Indonesia illustrates the use of field experiments to test a particular angle of the relationship between democracy and development. While most of the previous work on the topic takes institutions as a given and studies their effects (Shepsle 2006), Olken's study starts from the recognition that, in countless examples, institutions and the public policies that follow them are endogenous.

Olken, with support from the World Bank and UK's Department for International Development (DfID), conducted a field experiment in forty-eight Indonesian villages, each of which was preparing to petition for infrastructure projects as part of the Indonesian Kecamatan Development Program. All villages in the experiment followed the same agenda-setting process to propose two infrastructure projects -- one general project determined by the village as a whole, and one women's project. The experiment randomly assigned villages to make the final decision regarding the projects either through a meeting or through a plebiscite. Olken examined the impact of meetings and plebiscites on elite capture along two main dimensions. First, he

examined whether the types of projects chosen moved closer to the preferences of village elites. Second, he tested whether the location of projects moved toward wealthier parts of the villages.

The experiment's findings paint a mixed picture. Whether there was a meeting or a plebiscite had little impact on the general project, however the plebiscite did change the location of the women's project to the poorer areas of a village. The type of project chosen by women, however, was closer to the stated preferences of the village elites than to poor villagers' preferences. Olken explains that because the experiment left the agenda-setting process unchanged, the elite's influence over the decision-making process regarding the type of project remained unchallenged. The experiment thus confirms previous arguments on the relevance of political mechanisms to aggregate preferences. At the same time, it shows the resilience of political inequalities.

The persuasiveness of the results comes from the research design, which guaranteed that plebiscites and meetings were allocated to villages regardless of their social and political configuration or any other observed or unobserved characteristic. Therefore, differences in the type and location of projects can be adjudicated with certainty to the political mechanism in place.

Olken's experiment is an example of a growing trend in political science and development economics where researchers collaborate with NGOs in order to implement an intervention and evaluate its effects. This type of partnership has proven fruitful for the study of a vast array of topics central to our understanding of the relationship between democracy and development. For example, Humphreys, Masters, and Sandbu (2006) explore the role of leaders in democratic deliberations in São Tomé and Príncipe; Bertrand et al. (2007) collaborate with the

International Finance Corporation to study corruption in the allocation of driver's licenses in India; Blattman, Fiala, and Martinez (2008) study the reintegration of ex-combatants in Northern Uganda; Collier and Vicente (2008) test the effectiveness of an antiviolenence intervention in Nigeria; Moehler (2008) investigates the role of private media in the strengthening of accountability; Levy Paluck and Green (2009) examine how media broadcasts affect interethnic relations in a post-conflict context; Fearon, Humphreys, and Weinstein (2009) collaborate with the International Rescue Committee to evaluate the impact of a community-driven reconstruction program in Liberia.¹ All of these studies were made possible in large part through collaboration with local and international NGOs.

Interventions led by NGOs can shed much light on social phenomena in contexts where the involvement of independent actors comes naturally, such as in the experiments described previously. There are cases, however, where one must give special consideration to the effect that an NGO's involvement may itself have on the social phenomena at hand. Ravallion (2008) writes:

the very nature of the intervention may change when it is implemented by a government rather than an NGO. This may happen because of unavoidable differences in (inter alia) the quality of supervision, the incentives facing service providers, and administrative capacity (17).

Moreover, there are social contexts where an NGO's involvement is not easily justified. In such cases, researchers have two options. First, they can undertake the enterprise of forging alliances with the relevant actors, such as government officials or politicians, required to randomize an intervention of substantive interest. Second, they can take advantage of the growing number of cases where natural experiments are already in place due to policymakers' decisions to randomize an intervention of interest.

2. Field experiments in collaboration with politicians

Wantchekon's (2003) study of clientelism and its electoral effectiveness in Benin is an example of a unique collaboration between researchers and politicians to implement a treatment. Wantchekon worked directly with presidential candidates to embed a field experiment in the context of the first round of the March 2001 presidential elections. Together with the candidates, Wantchekon randomly selected villages to be exposed to purely clientelist or purely public policy platforms.

Prior to this study, scholars had given little attention to the effectiveness of clientelist and programmatic mobilization strategies. Stokes (2007) notes that “most students and casual observers of clientelism assume that it works as an electoral strategy -- that, all else equal, a party that disburses clientelist benefits will win more votes than it would have had it not pursued this strategy. In general we do not expect parties to pursue strategies that are ineffective. And yet we have some theoretical reasons for believing that conditions are not always ripe for clientelism” (622). The challenge of estimating the effectiveness of clientelism, patronage, and pork-barrel as mobilization strategies rests in the possibility that electoral performance can shape spending decisions (Stokes 2007).

The Benin experiment empirically validates the argument that clientelist appeals are a winning electoral strategy, whereas public policy appeals produce mixed results. Beyond confirming these arguments, the Benin experiment presents a wide range of new results that are counterintuitive and could not likely have been derived from any other form of empirical research because in Benin we almost never observe a candidate campaigning on public policy. The experiment shows for instance that 1) clientelist appeals reinforce ethnic voting (not the

other way around), 2) voters' preference for clientelist or public goods messages depends in large part on political factors, such as incumbency, and on demographic factors, such as gender, and 3) the lack of support for programmatic platforms is not due to opposing preferences among groups, level of education, or poverty, but instead to the fact that public policy platforms lack credibility, presumably because they tend to be vague.

In a follow-up experiment implemented in the context of the 2006 presidential elections, Wantchekon (2009) finds that broad-based platforms can be effective in generating electoral support when they are specific and communicated to voters through town hall meetings. As a result of these experiments, discussions of how to promote broad-based electoral politics in Benin now have empirical basis.

3. Natural Experiments

While experiments like Wantchekon's (2003) are still rare, scattered throughout the literature on development are examples of randomized interventions where assignment of treatment is outside of researchers' control. Chattopadhyay and Duflo's (2004) study of the quota system for women's political participation and the provision of public goods in India is such an example. The natural experiment was facilitated by the 73rd Amendment, which required that one-third of Village Council head positions be randomly reserved for women. Chattopadhyay and Duflo's evidence confirms that correcting unequal access to positions of representation leads to a decrease in unequal access to public goods. To begin with, the quota system was effective. In the two districts studied (West Benagal and Rajasthan), all positions of chief in local village councils (Gram Panchayats, henceforth GPs) reserved for women were, in fact, occupied by females. In turn, having a woman chief increased the involvement of women in GPs' affairs in

West Bengal, but had no effect on women's participation in GPs in Rajasthan. Moreover, the increase in women's nominal representation translated into substantive representation.

The study of the quota system shows that women invest more in goods that are relevant to the needs of local women: water and roads in West Bengal and water in Rajasthan. Conversely, they invest less in goods that are less relevant to the needs of women: nonformal education centers in West Bengal and roads in Rajasthan. The evidence from this study confirms that some classic claims of representative democracy, such as the relevance of rules and the identity of representatives, hold true. Subsequent studies, however, show that despite institutional innovations, political inequalities and prejudice continue to bias the representation system against minority and disadvantaged groups. In particular, once the GPs' chief position was no longer reserved for women, none of the chief women were reelected, even though villages reserved for women leaders have more public goods and the measured quality of these goods is at least as high as in nonreserved villages (Duflo and Topalova 2004).

In Latin America, Ferraz and Finan (2008) make use of a natural experiment to study the effects of the disclosure of local government corruption practices on incumbents' electoral outcomes in Brazil's municipal elections. The research design takes advantage of the fact that Brazil had initiated an anti-corruption program whereby the federal government began to randomly select municipal governments to be audited for their use of federal funds. To promote transparency, the outcomes of these audits were then disseminated publicly to the municipality, federal prosecutors, and the general media. Ferraz and Finan compare the electoral outcomes of mayors eligible for reelection between municipalities audited before and after the 2004 municipal elections.

Ferraz and Finan find that, conditional on the level of corruption exposed by the audit, incumbents audited before the election did worse than incumbents audited after the election. Furthermore, in those municipalities with local radio stations, the effect of disclosing corruption on the incumbent's likelihood of reelection was more severe. This finding is in line with earlier contributions that show how access to information affects the responsiveness of governments. Moreover, it also corroborates findings that the media is important to diffuse information and discipline incumbents for poor performance (Besley and Burgess 2002; Stromberg 2004).

De La O's (2008) study of the electoral effects of the Mexican conditional cash transfer program (Progresa) is a third example of the use of a natural experiment. Finding the electoral effectiveness of programmatic spending presents similar challenges to the ones previously discussed. In order to evaluate the causal effect of spending, one needs to find exogenous variation on it. De La O empirically examines whether Progresa influenced recipients' voting behavior by taking advantage of the fact that the first rounds of the program included a randomized component. Five hundred and five villages were enrolled in the program twenty-one and six months before the 2000 presidential election. De La O finds that the program increased turnout in 2000 by five percentage points and increased the incumbent's vote share by four percentage points.

4. Lab and lab-in-the-field experiments

Research opportunities such as the ones described in previous sections are becoming more common as governments, NGOs and sponsors around the world are giving priority to the systematic evaluation of interventions. There are, however, other questions central to the field of political economy of development that require a deeper understanding of the microfoundations of

social processes. For example, what determines preferences over redistribution? Why do some individuals behave in a self-interested way while others seem to be altruistic? Why do some communities prefer private over public goods? Why is inequality tolerated more in some places than others? What triggers reciprocity?

Political scientists have found experimentation in the laboratory useful to study these and many other questions. The laboratory gives researchers complete control over assignment to treatment, the treatment itself, and -- perhaps most alluring -- control over the setting where subjects are exposed to the treatment. The price that researchers pay for the internal validity of experimental results produced in a laboratory is a well-known critique about external validity. Concerns about generalizability, however, are not a dismissal of laboratory experiments. Rather, they are an opportunity for creative researchers (Camerer 2003). Indeed, recent studies have shown that lab-based experimentation needs not to be confined to universities.

Habyarimana et al. (2007), for example, take the experimental laboratory to Uganda to study the mechanisms that link high levels of ethnic diversity to low levels of public goods provision. In this study, subjects are naturally exposed to high ethnic diversity on a daily basis. Thus the conclusions drawn from the dictator, puzzle, network, and public goods games played by Ugandan subjects speak directly to the social phenomenon of interest.

The games in Uganda show that laboratory experimentation enables researchers to adjudicate among complex mechanisms that in less controlled settings would be confounded. For example, Habyarimana et al. find that ethnic diversity leads to lower provision of public goods, not because co-ethnics have similar tastes or are more altruistic, but because people from

different ethnic groups are less linked in social networks. Therefore, the threat of social sanction for people that do not cooperate is less credible.

5. Challenges for experiments

Internal Validity

The advantage of experiments compared to observational research is that random assignment ensures that, in expectation, the treatment groups have the same observable and unobservable baseline characteristics. As the editors of this volume note in the introduction, however, random assignment alone does not guarantee that the experimental outcome will speak convincingly to the theoretical question at hand. The interpretation of the experimental result is a matter of internal validity -- whether the treatment of interest was, in fact, responsible for changing outcomes. For example, in a pioneering field experiment, Levy Paluck and Green (2009) seek to gauge the causal effect of listening to a radio program aimed at discouraging blind obedience and reliance on direction from authorities, and promoting independent thought and collective action in problem solving in post-genocide Rwanda. Research assistants played the radio program on a portable stereo for the listener groups. The challenge of this experimental design in terms of internal validity is that listener groups often lingered to chat after the radio program finished. Therefore, the effect of the radio program could be conflated with the effect of socialization. Levy Paluck and Green successfully dealt with this challenge by recording on standardized observation sheets the lengths and subjects of discussions during and after the program. With this information, they could test whether groups exposed to a particular radio program socialized more than other groups.

The interpretation of experimental results also depends on what the control group receives as treatment. In the experiment in Rwanda, for example, the control group listened to an educational-entertainment radio soap opera, which aimed to change beliefs and behaviors related to reproductive health and HIV. The average treatment effect is therefore the relative influence of the different content of the radio programs. This comparison is different from a comparison between those who listen to a radio program and those who don't listen to anything at all. A comparison between a group of listeners and a control group, however, would be problematic in terms of internal validity because the treatment group would not only be exposed to radio program content but also to group meetings, interactions with research assistants, and so on.

More generally, researchers in political economy of development face three challenges. First, because of the nature of the subject, researchers in development and democracy need to forge alliances with relevant decision makers to study social phenomena. These alliances make research more realistic, but also more challenging. Policymakers, both in government and NGOs, are interested in maximizing the effect of a specific intervention and it is natural for them to endorse treatments that consist of a bundle of interventions. For example, Green et al. (2010), in partnership with the Sarathi Development Foundation, implemented a field experiment in India during the 2007 election to examine how voters in rural areas would respond to messages urging them not to vote on caste lines but to vote for development. The treatment consisted of puppet shows and posters. This bundle of interventions is attractive from the NGO perspective, but is challenging for researchers who want to estimate the average treatment effect of an educational campaign.

To make the challenge more explicit, assume that in the example of Green et al.'s Indian field experiment compliance with the research protocol was perfect. If the effects of posters and puppet shows are independent from each other, then the effect of the bundled intervention is equal to the sum of the effects of the individual components of the intervention. By contrast, if the effects of posters and puppet shows are not independent, then there are four possibilities: posters might magnify the effect of puppet shows and vice versa or, alternatively, posters might cancel out the effect of puppet shows (and vice versa). In this particular application, it might not be theoretically relevant to isolate the effects of the two components of the treatment. In other applications, however, the degree to which an experiment can shed light onto a theoretical question will depend on how the individual components of bundled treatments map onto theoretically relevant variables.

The second challenge faced by experimental researchers is that logistical difficulties of working in the field oftentimes compromise compliance with research protocols. One form of noncompliance occurs when those assigned to the treatment group do not receive the treatment. In this case, the randomly assigned groups remain comparable, but the difference in average outcomes does not measure the average treatment effect. For example, De La O et al. (2010) design an informational campaign in Mexico where households in randomly selected polling precincts receive a flyer with information about their municipal government's use of a federal transfer scheme aimed at improving the provision of public services. Complying with the research protocol was more challenging in some of the experimental sites than in others because some of the polling precincts were more isolated. Naturally, easy-to-access precincts are different than harder-to-access precincts – easy-to-access precincts are more urban and wealthier

than the other precincts. These sociodemographic differences are directly correlated to partisanship. Thus, in this example, noncompliance in the form of failure-to-treat could greatly compromise the experimental design. De La O et al. circumvent the problem of noncompliance by including several mechanisms of supervision in the distribution of flyers, including the use of GPS receivers and unannounced audits.

An alternative form of noncompliance occurs when a treatment intended for one unit inadvertently treats a unit in another group. The risk of spillover effects is prevalent in the study of politics of development. In the Rwanda experiment, for example, the radio program was also being nationally broadcasted, so listeners in both treatment groups could listen to the program independent of the study. To minimize spillover effects, Levy Paluck and Green use strategies, such as offering to give participants in both groups the cassettes containing the radio program they were not suppose to listen to at the end of the study. An alternative strategy to deal with problems generated by spillovers is for researchers to choose a unit of analysis that enables them to estimate overall treatment effects. For example, Miguel and Kramer (2004) design a field experiment in Kenya where de-worming drugs are randomly phased into schools, rather than provided to individuals. With this design, Miguel and Kramer can take into account the fact that medical treatment at the individual level has positive externalities for nontreated individuals in the form of reduced disease transmission.ⁱⁱ

External Validity

Field experiments are valuable tools for the study of development and democracy, but designing and executing an experiment that speaks convincingly to theoretical questions of interest to the field presents some challenges, in addition to the ones discussed in the previous

section. Just like field researchers, experimental researchers face a tradeoff between the depth of knowledge that comes from studying a particular population and the generalizability of their findings (Wood 2007).

In order to address challenges to external validity, researchers must design their experiments with four things in mind. First, it is often the case that researchers need to exert great effort to include in a study the subset of the population worth studying, rather than the subset of the population that is most readily available to participate in a randomized trial. For example, Habyarimana et al. (2007) recruit their subjects from an area in Uganda characterized by high levels of ethnic diversity and low levels of public goods provision. In the Rwandan experiment, Levy Paluck and Green (2009) include two genocide survivor communities and two prisons in their fourteen experimental sites. Fearon, Humphrey, and Weinstein's (2009) study includes communities in post-conflict Liberia where the majority of the population had been affected by war either because they experienced violence or were displaced.

Second, the context of an experiment must resemble the context of the social phenomenon of interest. For example, in the experiment in Mexico, De La O et al. (2010) distribute to households the information about municipal spending of the infrastructure fund close to the election day. An alternative design would be to recruit individuals for a study where similar information would be distributed in informational meetings directed by the researchers. This design, however, comes less naturally than that of flyer distribution – a widely used communication technique in developing countries.

Third, researchers must find creative ways to design treatments that resemble the variables of interest in the real world. In this sense, not only the treatment but the scale of a field

experiment must be taken into account when thinking about external validity. Consider the recent trend in the field, where researchers collaborate with policymakers to evaluate an intervention in its pilot phase. Within these partnerships, policymakers welcome researchers' interventions in small-scale versions of larger policy projects. Yet, as Deaton (2009) explains:

small scale projects may operate substantially different than their large scale version. A project that involves a few villagers or a few villages may not attract the attention of corrupt public officials because it is not worth their while to undermine or exploit them, yet they would do so as soon as any attempt were made to scale up. So that there is no guarantee that the policy tested by the randomized controlled trial will have the same effects as in the trial, even on the subjects included in the trial (42).

Finally, researchers must find ways to measure outcomes that resemble the actual outcomes of theoretical interest. Indeed, experiments have in some cases started to revolutionize the field by presenting alternative measures of key concepts, such as corruption and vote buying. Consider Olken's (2007) field experiment in 608 Indonesian villages where treatments were designed to test the effectiveness of top-down and bottom-up monitoring mechanisms to reduce corruption. Unlike much of the empirical work that measures corruption based on perceptions, Olken measured corruption more directly, by comparing two measures of the same quantity, one before and one after corruption. With this innovative measure, Olken found that bottom-up interventions were successful in raising participation levels. However, when compared to the top-down intervention, the bottom-up interventions proved to be less successful at reducing corruption.

Nickerson et al. (2010) present another example where a field experiment innovatively measures a critical concept on the field. Numerous qualitative studies of vote buying have concluded that the exchange of votes for gifts or cash is a prevalent practice around the world. Yet studies based on survey research have consistently found surprisingly little evidence of vote

buying. Nickerson et al. measured the frequency of vote buying in the 2008 Nicaraguan municipal elections using a survey-based list experiment. All respondents were asked how many activities from a list were carried out by candidates and party operatives during the elections. The control group was given a list of four activities, including typical campaign activities like hanging posters, visiting homes, placing advertisements in the media, as well as not-so-typical activities like making threats. The treatment group was given the same list of activities, with the addition of vote buying. Since respondents were not asked which of the activities they witnessed but rather how many, a certain degree of anonymity when reporting vote buying was guaranteed. The proportion of respondents receiving a gift or favor in exchange for their vote was then measured as the difference in responses between the treatment and the control group. Based on the list experiment, the authors estimated that nearly a quarter of respondents received a gift or favor in exchange for their vote. In contrast, less than three percent of respondents reported that they had received a gift or favor when asked directly.ⁱⁱⁱ

Moving forward, researchers will be confronted with the challenge of designing field experiments in a way that enables the accumulation of knowledge. According to Martel Garcia and Wantchekon (2010), there are two ways to achieve this goal. One option is to replicate as much as possible the relationship between two variables under different conditions (the robustness approach). The ongoing research on the role of information in community development projects illustrates this approach. Banerjee et al. (2010) find that in India a randomly assigned information campaign was not effective at fostering community involvement in Village Education Committees and, ultimately, had no impact on teacher effort or student learning outcomes. By contrast, a similar study in Uganda reveals that, as a result of an

informational campaign, people became more engaged in Community-Based Organizations and began to monitor the health units more extensively. This community-based monitoring increased the quality and quantity of primary health care provision (Bjorkman and Svensson 2007).

The examples provided in this section show that, even in cases where similar experiments are executed across two different populations, contextual differences could cause the same intervention to have different effects. An alternative to replicating similar treatments in different contexts is to use an analytical approach that makes the theoretical foundations of an experiment more explicit (Martel Garcia and Wantchekon 2010). This analytical approach brings front and center the mechanisms that link a causal variable to an outcome. By being explicit about mechanisms, researchers can develop trajectories of experiments that are suitable to test theoretically informed hypotheses.

Consider, for example, the Benin electoral experiments (see Wantchekon 2003, 2009). One of the findings of the 2001 experiment is that voters are more likely to react positively to a public goods message when it comes from a coethnic candidate. A possible explanation for this finding is that voters trust a candidate from their ethnic group more than they trust a candidate from another group. This means that the mediating variable between ethnic ties and votes is trust, or the credibility of the candidate. By testing the relationship between credibility of candidates and voting behavior in a follow-up experiment in 2006, Wantchekon (2009) improves the external validity of the results of the 2001 experiment. As the Benin electoral experiments illustrate, to make scientific progress in this field, new experimental designs should not only take into consideration the context of current experiments, but should also focus on testing various aspects of a theory in a coherent way.

On the Ethics of getting involved in elections

One of the most striking features of experiments on democracy is that they require researchers to work directly with policymakers, politicians or government officials and to get involved in, in many cases, with running elections, government programs, or education campaigns. Embedding experiments in the context of real elections and programs brings a great degree of realism to the treatments. However, what is gained in terms of the external validity of the experimental results may not sufficiently offset ethical concerns.

We are far from having a consensus on where to draw the line between interventions that are ethical and interventions that are not. Nevertheless, there are several guidelines that researchers can follow when designing an experiment. First of all, an intervention will raise fewer ethical concerns if the units under study are exposed to a treatment they would ordinarily seek. In the Benin experiments, for example, the clientelist treatment could at first glance be a source of concern. Candidates in Benin, however, typically run campaigns based on clientelist appeals, regardless of researchers' presence. In such experiments, the researcher was merely acting as an unpaid campaign advisor to the candidate or civic educator. The researcher's main contribution was to suggest random assignment of campaign messages to districts. If anything, random assignment of messages is more ethical than the standard opportunistic tailoring of messages to what voters want to hear.

A similar concern is raised by experimental designs where subjects in one group are denied a treatment that they would ordinarily seek. For example, a study undertaken to examine the effect of international aid, where some villages are randomly selected to receive aid and some equally needy villages are randomly selected to be denied aid, is bound to raise ethical questions.

Practical considerations, however, can help researchers mitigate these concerns. For example, in most cases, NGOs and governments have limited budgets that force them to make decisions regarding where to start an educational campaign, a social policy, or any other intervention of interest. Random assignment in these cases provides policymakers with a transparent and fair way to decide the order in which subjects are, for example, enrolled in a program.

An ongoing field experiment in Uganda illustrates this empirical strategy. Annan et al. (2010), in collaboration with Innovations for Poverty Action and the Association of Volunteers in International Service (AVIS), are evaluating the Women's Income Generating Support (WINGS) program that provides women with grants and business training. To find whether small grants empower women and shape their political participation, Annan et al. will enroll women to the program in different phases over the course of three years. The order of enrollment is randomly assigned. This design enables causal inferences, but no vulnerable household contacted by researchers will be left out of the program.

A second way to think about ethical issues is to ask: what are the costs to subjects of participating in an experiment? In the Benin examples, if there were a cost to voters for being exposed to clientelist messages, this cost is already routinely incurred in all elections. In fact, the whole purpose of the experiment was to lower the cost of this campaign strategy for voters in future elections. More generally, experimental designs must take into account the costs of exposing subjects to treatments including, but not limited to, material costs (e.g., the opportunity costs of spending time in the study), psychological costs, and even physical costs.

A third set of ethical issues that researchers must take into account is the degree to which interventions alter the outcomes and the costs associated with such departures. For example, in

the experimental study of elections, one common concern is that researchers change the result of an election. A 2002 *New York Times* article commenting on the 2001 Benin experiment stated: “There are some major ethical concerns with field experiments in that they can affect election results and bring up important considerations of informed consent.”^{iv} Wantchekon, however, suppressed this possibility by including in the experiment only safe districts, where candidates collaborating in the study had a stronghold.

In this particular example, the subset of districts where ethical concerns are manageable coincided with the subset of districts that were theoretically relevant to study, because clientelism is more resilient in districts where one political machine has a monopoly than in districts where there is more political competition. In other applications, restricting the experiment to certain subpopulations where ethical concerns are manageable may compromise the external validity of the experiment’s results.

Finally, many research questions in the political economy of development, like the effect of violence on development, involve interventions that are difficult to study through experimentation without raising ethical concerns. Creative experimental designs, however, can enable researchers to study social phenomena that at first glance seem out of reach. For example, Vicente (2007) conducted a field experiment in São Tomé and Príncipe to study vote buying. As in many other countries, buying votes is illegal in São Tomé. Thus, Vicente randomly assigned subjects to be exposed to an anti-vote buying campaign, which was sponsored by the National Electoral Commission.

6. Concluding Remarks

The rise of experiments as one of the most prominent empirical strategies has led to new advances in the study of democracy and development. So far, some experimental results have confirmed previous arguments, such as the effectiveness of clientelism as a mobilization strategy and the prevalence of political and social inequalities despite institutional innovations. Other experiments have revealed relationships that only a randomized control trial could uncover, like the fact that clientelist appeals reinforce ethnic voting and not the other way around. Finally, some experiments are revolutionizing the measurement of core concepts in the field. For example, we now know that vote buying measured experimentally is more prevalent than what observational studies suggested.

Going forward, field experiments in collaboration with policymakers, governments, and NGOs are a promising line of research. The next round of experiments, however, faces considerable challenges, including those we have highlighted throughout this chapter. First, researchers must find creative ways to design interventions that are attractive to potential partners but that still speak convincingly to theoretically relevant questions. In doing so, researchers must pay special attention to internal validity issues. Second, a more analytical approach would help guide researchers to design experiments that enable significant accumulation of knowledge to take place. Finally, as the scope of experimentation expands, the tradeoff between external validity and ethical concerns will become more salient.

Despite these challenges, experimental research on development and democracy is a productive and exciting endeavor. As insightful as the experimental research has been up until now, numerous substantive questions remain unanswered. Hopefully the selection of studies

covered in this chapter illustrate how experiments can be used as a research tool to study broader and more central questions about the relationship between democracy and development.

References

- Annan, Jeannie, Christopher Blattman, Eric Green, and Julian Jamison. 2010. "Uganda: Enterprises for Ultra-poor Women after War." Unpublished manuscript, Yale University. Retrieved from <http://chrisblattman.com/projects/wings/>.
- Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence From a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy* 2: 1-30.
- Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122: 1639-76.
- Besley, Timothy, and Robin Burgess. 2002. "The Political Economy of Government Responsiveness: Theory and Evidence from India." *Quarterly Journal of Economics* 117: 1415-52.
- Blattman Christopher, Nathan Fiala, and S. Martinez. 2008. "Post-Conflict Youth Livelihoods: An Experimental Impact Evaluation of the Northern Uganda Social Action Fund (NUSAF)." Monograph. Washington, DC: World Bank.
- Boix, Carles, and Susan Stokes. 2003. "Endogenous Democratization." *World Politics* 55: 517-49.
- Bjorkman, Martina, and Jakob Svensson. 2007. "Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda." World Bank Policy Research Working Paper No. 4268.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.
- Chattopadhyay, Raghavendra, and Esther Duflo. 2004. "Women as Policymakers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72: 1409-43.
- Collier P, Vicente P. 2008. "Votes and Violence: Experimental Evidence from a Nigerian Election." Households in Conflict Network (HiCN) Working Paper No. 50.
- De La O, Ana L. 2008. "Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico." Unpublished manuscript, Yale University.

- De La O, Ana L., Alberto Chong, Dean Karlan, and Leonard Wantchekon. 2010. "Information Dissemination and Local Governments' Electoral Returns, Evidence from a Field Experiment in Mexico." Paper presented at the conference on Redistribution, Public Goods and Political Market Failures, Yale University.
- Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." Unpublished manuscript, Princeton University.
- Duflo, Esther, and Petia Topalova. 2004. "Unappreciated Service: Performance, Perceptions, and Women Leaders in India." Working paper, Massachusetts Institute of Technology.
- Fearon James D., Macarthan Humphreys, and Jeremy Weinstein. 2009. "Can Development Aid Contribute to Social Cohesion After Civil War? Evidence from a Field Experiment in Post-conflict Liberia." *American Economic Review: Papers and Proceedings* 99: 287-91.
- Ferraz, Cláudio, and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effect of Brazil's Publicly Released Audits on Electoral Outcomes" *Quarterly Journal of Economics* 123: 703-45.
- Green, Jennifer, Abhijit Banerjee, Donald Green, and Rohini Pande. 2010. "Political Mobilization in Rural India: Three Randomized Field Experiments." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Habyarimana, James, Macarthan Humphreys, Daniel N. Posner, and Jeremy Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101: 709-25.
- Humphreys, Macartan, William A. Masters, and Martin E. Sandbu. 2006. "The Role of Leaders in Democratic Deliberations: Results from a Field Experiment in Sao Tome and Principe." *World Politics* 58: 583-622.
- Humphreys, Macartan, and Jeremy Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12: 367-78.
- Keefer, Philip. 2007. "The Poor Performance of Poor Democracies." In *The Oxford Handbook of Comparative Politics*, eds. Carles Boix, and Susan Stokes. New York: Oxford University Press.
- Levy Paluck, Elizabeth and Donald P. Green. 2009. "Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda." *American Political Science Review* 103: 622-644.
- Martel Garcia, Fernando, and Leonard Wantchekon. 2010. "Theory, External Validity, and Experimental Inference: Some Conjectures." *The Annals of the American Academy of Political and Social Science* 628: 132-47.

- Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72: 159-217.
- Moehler, Devra C. 2008. "Tune in to Governance: An Experimental Investigation of Radio Campaigns in Africa." Paper presented at the conference on Field Experiments in Comparative Politics and Policy, University of Manchester, UK.
- Moehler, Devra C. 2010. "Democracy, Governance, and Randomized Development Assistance." *The Annals of the American Academy of Political and Social Science* 628: 30-46.
- Nickerson, David, Ezequiel Gonzalez-Ocantos, Chad Kiewiet de Jonge, Carlos Meléndez, and Javier Osorio. 2010. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." Unpublished manuscript, University of Notre Dame.
- Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115: 200-49.
- Olken, Benjamin. 2010. "Direct Democracy and Local Public Goods, Evidence from a Field Experiment in Indonesia." *American Political Science Review* 104: 243-67.
- Przeworski, Adam, and Limongi Neto, Fernando. 1997. "Modernization: Theories and Facts." *World Politics* 49: 155-83.
- Ravallion, Martin. 2008. "Evaluation in the Practice of Development." Policy Research Working Paper No. 4547. Washington, DC: World Bank.
- Shepsle, K. 2006. "Old Questions and New Answers about Institutions: The Riker Objection Revisited." In *The Oxford Handbook of Political Economy*, eds. Barry R. Weingast, and Donald A. Wittman. New York: Oxford University Press.
- Stokes, Susan. 2007. "Political Clientelism." In *The Oxford Handbook of Comparative Politics*, eds. Carles Boix, and Susan Stokes. New York: Oxford University Press.
- Stromberg, David. 2004. "Radio's Impact on Public Spending." *Quarterly Journal of Economics* 119: 189-221.
- Vicente, Pedro C. 2007. "Is Vote Buying Effective? Evidence from a Randomized Experiment in West Africa." Economics Series Working Papers No. 318, University of Oxford.
- Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55: 399-422.
- Wantchekon, Leonard. 2009. "Can Informed Public Deliberation Overcome Clientelism? Experimental Evidence from Benin." Working Paper, New York University.

Wood, Elizabeth. 2007. "Field Research." In *The Oxford Handbook of Comparative Politics*, eds. Carle Boix, and Susan Stokes. New York: Oxford University Press.

ⁱ For two excellent summaries of these studies, see Humphreys and Weinstein (2009) and Moehler (2010).

ⁱⁱ For more details on Miguel and Kramer (2004) experiment please see Nickerson's chapter in this volume.

ⁱⁱⁱ For more details on the origins of the list experiment see Sniderman's chapter in this volume.

^{iv} Lynnley Browning. 2002. "Professor Offer A Reality Check for Politicians." *The New York Times*, August 31. Retrieved from <http://www.nytimes.com/2002/08/31>.

VIII. Elite Bargaining

28. Coalition Experiments

Daniel Diermeierⁱ

The literature on coalition experiments is blessed by a close connection between theory and empirical work, using both experimental and field data.ⁱⁱ In political science the main research goal has been to understand the formation of coalition governments in multi-party democracies. While other aspects of coalitions could be considered, for example coalition stability or the creation and maintenance of military alliances, they have not been the focus of much existing research. We will therefore focus on experiments in coalition formation.

The goal of this chapter is not to provide an accurate description of the complete history of coalition experiments, but to discuss some of the key questions that continue to occupy researchers today. For example, much of the recent experimental literature has tested non-cooperative models of coalition government. While there has been a substantive amount of research on solution concepts from cooperative game theory,ⁱⁱⁱ their performance to explain experimental data has long been considered unsatisfactory, leading some researchers such as Gamson (1964) to propose an “utter confusion” theory.

In addition to testing of theoretical models, experiments can also be used to supplement field studies. For example, field studies (Martin and Stevenson 2001, Diermeier, Eraslan, and Merlo 2003) have indicated the importance of constitutional features on government formation, specifically the presence of an investiture vote. Yet, the equilibria predicted by non-cooperative models in general depend on institutional detail, for example the protocol in which offers can be made and so forth. However, there is no straightforward match between (existing) theoretical models and field data. For example, the predictions of a model may depend on the protocol by

which proposers of possible cabinets are selected (so-called formateurs), but such protocols may be based on implicit conventions that are not easily inferred from publicly available data. To address these issues some researchers (Diermeier et al. 2003, 2007) have proposed structural estimation techniques, but these models are difficult to construct and still limited by existing field data. By using experiments we can fill these gaps and directly study the consequences of institutional differences. Until very recently the literature on coalition experiments has largely followed this research agenda; its goal has been to carefully test theories of coalition formation, usually formal theories formulated in the language of non-cooperative game theory. The goal then has been to implement these theories as faithfully as possible in a laboratory setting, following the methodology developed in experimental economics and game theory (Roth 1995): experimental subjects interact anonymously via computer terminals, payments are based on performance with payment schedules carefully constructed to avoid any contamination from factors such as risk preferences, desire to please experimenters, and so forth. While this approach has been very fruitful and led to important insights, recent research has also pointed to its limitations. After reviewing the research that has followed the experimental economics, we will introduce a different research tradition which introduced some methods from social psychology into the study of coalition formation. This approach is emphatically context-rich, allowing face-to-face interaction with little restrictions on communication or negotiation protocols. Of particular promise is the ability to study the role of language in coalition negotiations, which opens to the possibility to systematically study communication and framing strategies.

1. The Baron-Ferejohn Model

During the late 1980s non-cooperative game theory became the dominant paradigm in the study of coalitions in the form of sequential bargaining models under majority rule, specifically reflected in the Baron-Ferejohn (1989) model. In all variants of the Baron-Ferejohn model a proposer is selected according to a known rule and then proposes an alternative to a group of voters. According to a known voting rule the proposal is either accepted or rejected. If the proposal is accepted, the game ends and all actors receive pay-offs as specified by the accepted proposal. Otherwise, another proposer is selected etc.^{iv} This process continues until a proposal is accepted.

Consider a simple version of the model where there are three political parties that need to decide on how to split one dollar. Suppose that no party has a majority of seats. The Baron-Ferejohn model predicts that the party with proposal power will propose a minimal winning coalition consisting of itself and one other member, leaving the third party with zero. The proposing party will offer its coalition partner just the amount necessary to secure acceptance. This amount (or continuation value) equals the coalition partner's expected pay-off if the proposal was rejected and the bargaining continued. Proposals are thus always accepted in the first round. Note that the proposing party will always choose as its coalition partner the party with the lowest continuation value. The division of spoils will, in general, be highly unequal, especially if the parties' discount factors are low.

The Baron-Ferejohn model is attractive for the study of cabinets, as coalition bargaining usually has a strong distributive component, for example control of government portfolios, and provides predictions even when the game lacks a core. Moreover, the model provides both point

and comparative statics predictions about the effects of different institutions on bargaining behavior and outcomes, for example proposer selection and amendment rules. This allows the modeler to assess the effects of different constitutional factors on coalition outcomes.

2. Testing the Baron-Ferejohn Model

Given its status as the canonical model of legislative bargaining, the predictions of the Baron-Ferejohn model were soon tested in controlled laboratory experiments, first by McKelvey (1991). McKelvey uses a three-voter, closed rule, finite alternative version of the Baron-Ferejohn model, which results in mixed strategy equilibria. McKelvey finds that the unique stationary solution to his game at best modestly explains the data: proposers usually offer too much and (the lower) equilibrium proposals predicted by the theory are rejected too frequently.

The McKelvey experiments followed the methodological approach of experimental game theory. Subjects interact anonymously through a computer terminal, are paid depending on performance, and so forth. The goal of this approach is to induce the incentives and knowledge structure specified by the game-theoretic model in the laboratory setting. Yet a faithful implementation of the Baron-Ferejohn model in the laboratory is challenging. These problems are common in experimental game theory and the solutions adopted by McKelvey address the problems to a large extent. Yet some residual concern should remain. First, the Baron-Ferejohn model is of (potentially) infinite duration, but must be implemented in finite time. But participants in a typical experimental session will know or at least reliably estimate the maximal duration of the game (the recruitment flyer will usually state the time they need to make themselves available for the experiment). This may induce endgame effects, especially for later rounds. Second, the model's prediction is based on the assumption of stationarity, a stronger

equilibrium requirement than mere subgame-perfection, which rules out dependence of previous actions and thus eliminates the use of punishment strategies familiar from the study of repeated games. Without it, the Baron-Ferejohn model faces a folk-theorem where all individually rational outcomes can be supported as equilibria. But this means that one cannot test the Baron-Ferejohn model in isolation, but only in conjunction *with* an additional equilibrium refinement (here stationarity). This is important as McKelvey suggests in his discussion of the findings that one way of accounting for the discrepancy between experimental outcomes and theoretical predictions is that subjects may implicitly try to coordinate on a non-stationary equilibria. Third, the unique stationary equilibrium involves randomization. This implies that the model is predicting a distribution over outcomes, not a single outcome, which requires many observations to detect a significant difference between predicted and observed frequencies.

Given the centrality of the Baron-Ferejohn model for formal theories of coalition formation the McKelvey results trigger subsequent experimental work. The first paper in this direction is Diermeier and Morton (2005), which follows the basic set-up of the McKelvey experiment but tries to resolve some of the methodological difficulties of testing the Baron-Ferejohn model. In contrast to McKelvey they use a finite game under weighted majority rule where a fixed pay-off is divided among three actors.^v This leads to a unique subgame perfect equilibrium without assuming additional stationarity. Second, the subgame-perfect equilibrium involves no randomization on the equilibrium path. Third, the weighted majority game allows us to test a rich set of comparative statics, not just point predictions. Comparative statics analyses are of particular interest in testing institutional models where the ability to correctly predict behavioral changes in response to institutional changes is critical.

Diermeier and Morton (2005), however, find little support for either point or comparative static predictions. First, proposers frequently allocate money to all players, not just to the members of the minimal winning coalition. Second, proposers do not seem to select the “cheapest” coalition partner that is to say, the one with the lowest continuation value. Third, proposers offer too much to their coalition partners. Fourth, a significant percentage of first-period proposals above the continuation value are rejected, sometimes repeatedly. Diermeier and Morton’s data, however, do reveal some consistent behavioral patterns. Proposers typically select a subset of players (sometimes containing all other players) and split the money equally among its members. Note that this eliminates the proposer premium. Indeed, players take extreme measures to guarantee equal payoffs among the coalition, even “wasting” small amounts of money to guarantee an equal split.

The Diermeier and Morton findings confirm and sharpen the McKelvey predictions. However, in both cases the main focus is on testing the model’s point predictions. Fréchette, Kagel, and Lehrer (2003) take a different approach by investigating the institutional predictions of the Baron-Ferejohn model. They compare open versus closed rule versions of the Baron-Ferejohn model with five players. As in McKelvey, but in contrast to Diermeier and Morton, play continues until agreement is reached and the focus is on stationary equilibria. In contrast to previous experiments, Fréchette et al. find some qualitative support for the Baron-Ferejohn model. In particular, there are longer delays and more egalitarian distributions under the open rule as predicted. However, some less obvious (but critical) aspects of the Baron-Ferejohn model are not well-supported in the data. For example, in their design proposers should propose minimal winning coalitions in both the open and closed rule case. However, only 4 percent of

proposals correspond to this prediction. Even more troubling, under the open rule, subjects accept proposals that offer them *less* than their continuation value.

A common concern with all experimental investigations of the Baron-Ferejohn model is that the specific games under consideration are cognitively highly demanding. Thus, one explanation for the Baron-Ferejohn models lack of empirical fit may be that subjects do not initially fully understand the game's complex incentives and have insufficient opportunity to learn. In this case subjects may simply revert to an equal sharing heuristic and select coalition partners haphazardly.

Fréchette et al. designed a second experiment to address these cognitive concerns by considering more rounds and by adding a graduate student to the subject pool that used an algorithm to implement the stationary subgame-perfect equilibrium strategy.^{vi} In this experiment proposal behavior more closely resembled the Baron-Ferejohn predictions: allocations are less egalitarian, and equal split proposals (among all players) completely vanish. Nevertheless, play does not converge to the allocation predicted by the Baron-Ferejohn model. Rather, proposers and voters seem to rely on a "fair" reference point of $1/n$ share of the benefits.^{vii} Offers below the reference share are consistently rejected while shares above $1/n$ are usually accepted.^{viii} This focal point interpretation may also account for the odd finding in the open rule case where subjects accepted an amount less than their continuation value, which happened to be significantly higher than the fair reference point.

The hypothesis that subjects are in part motivated by fairness concerns is highly consistent with a related literature in experimental economics on bilateral bargaining games with the ultimatum game as the most well-known example (Güth et al. 1982). In the ultimatum game

one player makes a proposal on the division of a fixed amount of money and the other player must either accept or reject; with rejection implying a zero payoff for both. In experiments on ultimatum games proposers should take (almost) all of the money, yet the divisions are far more equal than predicted. Moreover, if proposers offer less than a certain amount^{ix}, the other player frequently rejects the offer (even if it is a significant amount of money) and receives a pay-off of zero. Experiments on bargaining games with a series of alternating offers result in similar outcomes. Proposers offer more money than suggested by their subgame-perfect strategy and bargaining partners consistently reject offers and forgo higher payoffs (Davis and Holt 1993, Forsythe et al. 1994, Güth et al. 1982, Ochs and Roth 1989, Roth 1995).

Forsythe et al. (1994) investigated this hypothesis by comparing ultimatum and dictator games. The dictator game differs from the ultimatum game in that the proposing player proposes a division between the two players and the other player cannot reject the proposal. In ultimatum games almost sixty percent of the offers observed propose an equal division of pay-offs. While there is still a significant percentage of equal divisions in dictator games (less than twenty percent) the modal division is the subgame perfect allocation where the proposer keeps the entire pay-off. This result suggests that while some of the subjects are primarily motivated by egalitarian notions of fairness, the high percentage of equal divisions in ultimatum games cannot be attributed to a simple desire to be fair.

The Baron-Ferejohn bargaining game is similar to these bargaining games in the sense that proposers are expected to offer their coalition partner his/her continuation value. In the last period of a finite game, this continuation value is zero, as in the ultimatum game. Hence the non-proposing coalition partner in the last period of the Baron-Ferejohn model is like the second

player in an ultimatum game. The relationship between the proposer and non-coalition member in the last period, however, is also similar to the dictator game, since the votes of the non-coalition members are not necessary to pass a proposal.

Diermeier and Gailmard (2006) present an approach to separate cognitive from motivational issues and directly focus on the question of whether and how agents in majoritarian bargaining situations are driven by moral motivations such as fairness. To do this they use a much simpler version of the proposer-pivot game that directly resembles the ultimatum game. Rather than having subjects calculate continuation values, players are directly assigned an *ex ante* known disagreement value, that is a given amount of money he/she will receive if the proposal is rejected. Proposers then make a take-it-or-leave-it offer on how to split a fixed, known amount of money among the players. Disagreement values are, in essence, a “reduced form” representation of continuation values. Alternatively, they can be interpreted as an ultimatum game with competing respondents. By varying the disagreement values as treatment variables, competing motivational theories can be tested. Recall that in a model with self-interested agents, any proposer will select the “cheaper” of the other voters and offer that player her disagreement value (perhaps with a little security margin), while the other (more expensive) voter receives zero. Similarly, voters will accept only offers at or above their respective disagreement values. Note that this optimal behavior (by proposers and voters) does not depend on the *proposer’s* disagreement value, other than in the trivial case where the value is so high that the proposer prefers his or her disagreement value to any possible proposal. Thus, varying the proposer’s reservation value should not have any influence on proposing or voting behavior. That is, the tested theory not only makes certain point and comparative statics predictions, it also

mandates that certain aspects of the game *should not matter*. If they do, the theory simply cannot completely account for the findings.

The results of Diermeier and Gailmard (2006) are at odds not only with the Baron-Ferejohn model, but also with any of the proposed fairness models considered (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). The key feature is the dependency on the reservation value of *the proposer*, which should have no strategic impact whether players are selfish or incorporate fairness consideration in their behavior. But this is not the case. Voters accept lower offers if the proposer has a higher reservation value and proposers proposed significantly more. Interestingly, voters are also more tolerant of higher offers to the other (non-proposing) voter if that voter has a higher reservation value. Diermeier and Gailmard interpret their findings as an entitlement effect. According to this interpretation, experimental subjects interpret their exogenously given reservation value as an entitlement that ought to be respected.^x

3. Alternative Bargaining Protocols

Much of the existing experimental work has concentrated on testing models of coalition formation, such as the Baron-Ferejohn model. Yet, as we discussed in the introduction, some experimental work has focused on institutional analysis instead, with an emphasis on an examination of bargaining protocols. One major influence has been Gamson's (1961) claim that portfolios among cabinet members will be allocated proportionally to the parties' seat shares. This hypothesis has found so much support in the field studies that it has been called "Gamson's Law" (Browne and Franklin 1973, Browne and Fendreis 1980, Schofield and Laver 1985, Warwick and Druckman 2001). Gamson's Law, however, seems to be at odds with the proposer

models as it implies the absence of a proposer premium, a critical implication of the Baron-Ferejohn framework.

This discrepancy has led to the development of alternative bargaining models, demand bargaining (Morelli 1999) and proto-coalition bargaining (Baron and Diermeier 2001, Diermeier and Merlo 2000). In a series of papers Fréchette, Kagel, and Morelli (2005a, 2005b, 2005c) have tested Gamson's Law (interpreted as a proportionality heuristic) in a laboratory setting and compared it to predictions from demand bargaining and the Baron-Ferejohn model.^{xi} In demand bargaining players do not make sequential offers (as in the Baron-Ferejohn model), but make sequential *demands*, that is to say compensations for their participation in a given coalition until every member has made a demand or until a majority coalition forms. If no acceptable coalition emerges after all players have made a demand, a new first demander is randomly selected; all the previous demands are void, and the game proceeds until a compatible set of demands is made by a majority coalition. The order of play is randomly determined from among those who have not yet made a demand, with proportional recognition probabilities.

A simple three-party case provides contrasting predictions generated by the three models. Suppose we have three parties and no party has a majority of seats. With equal proposal power (and sufficiently high discount factor) the equilibrium allocation in the Baron-Ferejohn model will give the proposer about two-thirds of the pie, with one-third given to one other party, and the third party receiving nothing. Demand bargaining, on the other hand predicts a fifty-fifty split between the coalition parties and nothing for the out-party. The Gamson predictions depend on the respective seat share. For example, in the setting used by Fréchette et al (2005a), if one

coalition member has nine votes and the larger forty-five, the larger party would receive about eighty-three percent of the share, *even if the smaller party is the proposer*.

In the laboratory both the Baron-Ferejohn and demand bargaining, however, outperform a proportionality heuristic in the laboratory. Consistent with the results reported in previous experiments, however, the proposer premium is too small compared to the Baron-Ferejohn predictions, and voters reject offers that they consider to be too low, even if acceptance would be in their self-interest, confirming the previous findings. The existence of a proposer premium, however, even if it is too small compared to the model creates a stark contrast with the field research. Fréchette et al. (2005a, 2005b) provide an intriguing explanation that reconciles this apparent conflict. The idea is to apply the same regression approaches used in field studies to the experimental data. The results show that the strong support of proportionality is due to proportional proposer selection as identified in Diermeier and Merlo (2000), not bargaining according to a proportionality heuristic once a proposer has been selected. Interestingly, the same statistical approach is also unable to distinguish between the Baron-Ferejohn model and demand bargaining. In other words, the regression approach used in field data cannot identify the underlying bargaining protocol. In other words, both the Baron-Ferejohn model and the demand bargaining model can explain the striking proportionality regularities in the field data, but existing field data methods cannot distinguish between the competing models.

In summary, experiments on coalitional bargaining suggest that sequential bargaining models offer a promising framework to understand coalition formation, with the Baron-Ferejohn bargaining protocol still the leading contender. Yet, some of the predictions of the Baron-Ferejohn model are consistently rejected in laboratory experiments. While a proposer premium

can consistently be identified, it is too small compared to the Baron-Ferejohn prediction. Correspondingly, voters consistently reject offers that they consider to be unfair.^{xiii} Both findings are consistent with the large literature on ultimatum games and persist in much simplified environments or when learning is possible. This suggests that the motivational profile of experimental subjects needs to be taken seriously, with moral considerations and framing playing an important role. That motivational profile, however, appears to be complex. It appears to include selfish components, fairness concerns, even respect for (arbitrary) entitlements.

The second main finding results from the study of alternative bargaining institutions. These results point to a potential problem of using sequential bargaining models as the formal framework to study coalition formation: they appear to be “too specific” for the task; available methods for studying field data cannot distinguish between competing approaches.

An important task for future work is thus (a) to allow for a richer set of agent motivations that are not captured by monetary incentives, and (b) to try to identify general features of sequential bargaining models that do hold for various model specifications. One main insight from sequential bargaining models is the any *current* agreement depends on the shared expectations of what would happen if that agreement could be reached or sustained, e.g. which *future* agreement would be formed. Notice that such future agreements not only may be less favorable to current coalition partners due a shift in bargaining strength, but, most importantly, future coalitions may consist of different parties, relegating at least some of the current coalition members to the much less desirable role of opposition party. Interestingly, this “fear of being left out” not only sustains current coalitions as equilibria, it may also lead negotiating parties to accept inefficient outcomes out of the fear that the current coalition will be replaced by a new

one and that they may be left out of the final deal. This was formally shown by Eraslan and Merlo (2001).

These questions are difficult to answer in the experimental methodology common in behavioral game theory that has dominated existing laboratory research on coalitions. Rather a more contextualized approach seems necessary. Interestingly, such a tradition already exists, but has been largely ignored by political scientists. It has been developed almost exclusively by psychologists under the name of “multi-party negotiations” and is almost exclusively experimental in nature.

4. Context-Rich Experiments

One first insight from social psychology is that the complexity of multi-party negotiations may be overwhelming to research subjects which may lead them to rely on simple heuristics instead, such as equal sharing (Bazerman, Curhan, Moore, and Valley 2000; Bazerman, Mannix, and Thompson 1988; Messick, 1993; Bazerman et al. 2000). It is instructive to reinterpret the Diermeier and Morton (2005), Diermeier and Gailmard (2006) and Fréchette et al. (2003) results in this context. One possible interpretation of the Diermeier and Morton findings is that subjects were overwhelmed by the complexity of the negotiation task and fell back on an equal-sharing heuristic. Interestingly, the heuristic used appeared to be an equal sharing heuristic, consistent with other results in social psychology, not a proportionality heuristic as suggested by Gamson’s Law. Once a simpler setting (Diermeier and Gailmard 2006) or opportunities for learning were provided (Fréchette et al. 2003), observed behavior more closely resembled predicted behavior, yet evidence for moral motivations could still be clearly detected.

From the point of researchers trained in game theory, much of the psychological multi-person negotiation literature may appear unsatisfactory as too little emphasis is placed on the incentive structure underlying the experiment. Yet, some of the insights potentially can be blended with a more strategically minded approach. In a recent series of papers Swaab, Diermeier, and various co-authors (Swaab et al. 2002; Diermeier et al. 2008; Swaab et al. 2009) have proposed such an approach. They consider the following characteristic function due to Raiffa (1982). The extensive form is intentionally not specified.

In the example in Figure 28-1, any efficient outcome involves the parties reaching a unanimous agreement, but at least some parties, for example, A and B, can form a fairly profitable agreement without including the third party, here C. So, one possible intuition of how the negotiations may proceed is as follows. Parties A and B (or some other “proto-coalition”) may form a preliminary agreement on how to split the pie already available to an AB coalition (here One hundred eighteen thousand) among themselves and then only need to negotiate over the remaining amount with C. The problem with this intuition is, of course, that C will try to break up any proto-coalition between A and B to avoid being left with a pittance. And attractive offers to A or B always exist. That is, for each possible split between A and B, C can propose an allocation that makes either A or B better off. Hence, A or B may be tempted to abandon their proto-coalition and team with C instead. Suppose B now forms a new proto-coalition with C. The A can make a better offer to either B or C and so forth.

In the Baron-Ferejohn model (or any of the sequential bargaining models considered) this problem is avoided by the fact that once a coalition is agreed upon, then game ends. But in the context of coalition government this is a problematic assumption as governing coalitions need to

maintain the confidence of the legislature to remain in power. In other words, during a legislative period the game never “ends” in a strategically relevant sense.

The key for negotiating parties is thus two-fold: (a) they need to settle on a coalitional agreement that includes them, and (b) they need to make sure that the coalition is stable. From a political economy point of view this would require identifying some self-enforcing mechanism that keeps the current coalition in power (Diermeier and Merlo 2000). Psychologists, on the other hand, have focused on alternative means to solve the stability problem. The key notion here is the concept of “trust.” Intuitively parties need to trust their proto-coalition partner to be willing to continue their conversation with the third party, as doing so carries the risk that the third party may be able to break up the current proto-coalition. If such trust cannot be established, parties may be better off refusing to further talk to the out-party to avoid giving it an opportunity to break the current proto-coalition apart, which would lead to an inefficient outcome.

Whether such trust can be established may depend on various factors. For example, players may use non-verbal communication to signal an agreement (they may seek eye contact before speaking to the third party, move together on the same side of the table, and so forth). Thus the degree to which non-verbal factors can be used (for example, in a face-to-face versus computer-based negotiation) may influence the stability of proto-coalitions and negotiation efficiency. The idea underlying the Swaab and Diermeier experiments is that by directly manipulating the communication structure, we can vary the conditions that lead to trust among members of proto-coalitions, which, in turn, influences whether efficient outcomes can be reached. So, the “independent variable” in this approach is the communication structure, the

dependent variable the percentage of efficient coalition outcomes, and the mediating variable trust in the proto-coalition partner.

Diermeier et al (2008) consider three such settings: (a) face-to-face versus computer-mediated decision making, (b) public versus private communication settings, (c) private and secret communication settings. The first finding is unsurprising. Groups negotiating face-to-face were significantly more efficient than groups negotiating via computer mediated communication (seventy percent versus eleven percent). However, face-to-face communication is a complex phenomenon. In addition to the use of non-verbal cues it creates a setting that enables the creation of common knowledge through public communication. To separate these issues, the authors introduced a privacy variable to the negotiation context, allowing parties in both face-to-face and CMC negotiations to access private discussion spaces separate from the third party. The authors expected that the availability of private chat rooms would decrease efficiency in the negotiations regardless of communication style (face-to-face vs. CMC), but that CMC would still lead to less efficient outcomes compared to face-to-face communication. The expectations were partially supported. The ability to communicate privately in the CMC setting lowered efficiency from fifty percent to eighteen percent. In the face-to-face condition the effect was much smaller (eighty percent to seventy-one percent) and not significant at customary levels of statistical significance.

The effects of communication structure can be quite subtle. For example consider the differences between using a private chat room and using instant messaging. In the first case, the content of the conversation is private, but the fact that private communication took place is public (the parties are observed when they “leave “ the common chat room), while in the instant

messaging case, the fact that private communication took place may not been known either, that is, communication is secret. If the intuition that communication structures influence trust between negotiating parties is correct, then secret communication should be particularly destructive, since parties can never be sure that their coalition partner is not secretly trying to double-cross them. Indeed, the mere possibility of secret communication taking place may already undermine trust. Diermeier et al. (2008) find this to be the case. When they compare secret communication to private communication, not a single group is able to reach an efficient outcome in the secret condition.

Once the importance of communication structures has been established, an investigation of communication strategies is a natural next step. This also may help to clarify how negotiators use language to develop a positive common rapport and shared mental models. In unpublished research Swaab, Diermeier, and Feddersen (personal communication) show that merely allowing participants to exchange text messages rather than being restricted to numerical offers doubles the amount of efficient outcomes. Huffaker, Swaab, and Diermeier (In Press) recently proposed the use of tools from computational linguistics to investigate the effect of language more systematically, using the same game-form as above. Taylor and Thomas (2003), for example, have pointed out that matching linguistic styles and word choices improve negotiation outcomes. Similarly, the use of assents (Curhan and Pentland 2007) and positive emotional content is expected to have a positive impact of negotiation success, while the use of negative emotional content may back-fire (Van Beest et al. 2008). Huffaker et al. use zipping algorithms to measure language similarities and text analytics programs to identify assents and emotional content. They

find a significant effect of linguistic similarity and assent, but no support for positive emotions. Negative emotions, however, do lead to fewer efficient coalition outcomes.

These are just some possibilities of including richer psychological models into the study of coalition formation. The investigation of language is perhaps the most promising as it connects to the large literature on framing that is increasingly gaining traction in political science, but other dimensions should be considered. Examples include the explicit study of decision-heuristics, as discussed in the context of Diermeier and Morton (2005) or the use of moral concerns (Fréchette et al. 2003, Diermeier and Gailmard 2006). Yet, some of the methodologies used in this context, for instance, the reliance on unstructured interaction, are alien to most experimental political scientists, though they do connect with a much earlier literature before the non-cooperative revolution. It therefore may be worthwhile to discuss some of these issues in more detail.

5. Comments on Context-Rich Experiments

Political scientists trained in game theory and experimental economists have largely ignored the extensive psychological literature on multi-person negotiations. In part, this may have been due to methodological disagreements. After all, most of the recent work on coalition bargaining fits squarely within the experimental economics research tradition sharing its standards, values, and methods. Yet, psychologists systematically, routinely, and intentionally violate many of the tenets experimental (political) economists hold sacred. Among the many possible violations consider just the following partial list.^{xiii}

1. Psychologists usually do not pay their subjects for performance and use large, but fictitious monetary values.

2. Psychologists mostly do not specify game-forms in all but the most rudimentary fashion.
3. Decision problems are not presented in abstract fashion, but richly contextualized using fictitious context. Face-to-face interactions are common.

For scholars committed to the experimental economics paradigm, it is tempting to dismiss much of the negotiations literature on such methodological grounds. But that would be a mistake. To see why, it is important to recall that the main goal of coalition experiments is to better understand coalition formation in real settings such as the negotiations over forming a new cabinet. It is not to study human behavior in games, even if these games are intended to capture real phenomena.

With this background in mind let us reconsider the approach taken by social psychologists. The issue of how subjects need to be paid to properly induce the desired incentives is, of course, an ongoing concern among experimental economists. Issues include ensuring trust that payments are actually made, the magnitude of the payments, whether they should be paid in cash or lottery tickets, and so forth. Secondly, experimental game theorists take great care to ensure that subjects are not influenced by any other aspect of the decision context than the one specified by the game-form. Subjects interacting on computer terminals are presented with abstract pay-off matrices. The motivation, of course, is to make sure that these other extraneous factors do not influence subjects' decision making processes. Only that, it is argued, will allow us to properly test the predictions of a given model. But there is an underlying assumption here. After all, our intended domain of application is not anonymous agents interacting on a computer screen being paid in lottery tickets. (That would be the proper domain for studying behavior in games). Our intended domain are professional politicians, who know

each other very well, participating in a series of meetings or phone calls over a period of days or weeks and negotiating over extremely high stakes that may determine their professional career.

The assumption (and promise) of the experimental economics approach is that as we move from the very abstract, stripped-down context in an game-theoretic setting to the richly contextualized setting of a cabinet negotiation by professional politicians the main insights gained in the abstract setting *survive*. In other words, having proposal power, for instance, would still be important whether one is deciding on how to split five dollars among players A, B, and C in a computer lab or negotiating on the composition of a ruling coalition. But recent research, ranging from the bias and heuristics literature to evolutionary psychology, suggests that this inference is far more problematic than previously believed. A well-known example is the Wason test (1966), where subjects are asked to turn over cards to determine whether a particular if-then statement is true. Human subjects are notoriously bad at this very elemental logical exercise. Yet performance improves dramatically when the task is presented in a contextualized version as a cheater detection problem (Cosmides and Tooby 1992).

For our discussion the main insight is that contextualized versions of an abstract decision-making problem yield very different predictions than the abstract version. The existence of these differences is, of course, nothing new, and widely documented in the enormous literature on framing effects. Indeed much of the care in designing experiments in the game-theoretic tradition can be interpreted as an attempt to eliminate these factors. But the lesson from the Wason test is quite different. Here subjects perform *much better* in a contextualized setting compared to the abstract one. Moreover, the contextualized version of the Wason test (cheater detection) is highly relevant to coalition formation. Indeed Cosmides and Tooby have argued that the very reason

subjects perform well in cheater detection tasks was the evolutionary need for effective coalition formation in early human society. So, if our goal is to understand reasoning in richly contextualized settings, using very abstract environments may bias results in the *wrong direction*. Indeed, this is precisely what we find when we compare negotiation performance in an abstract setting to a more contextualized setting, for example, one where agents are allowed to exchange text messages. The results by Diermeier et al. (2008) further illuminated this insight as richer communication structures correlate with better bargaining success in a predictable manner.

This approach may open up a potential blend of the experimental economics and social psychology traditions in the context of coalition bargaining experiments. The key challenge will be how to strike the right balance between specifying enough context to avoid the Wason-test trap, but in a fashion to preserve a sufficient level of experimental control. This blend of strategic models and psychological richness may offer some highly promising research directions. Political science with its dual research heritage containing both behavioral and formal traditions seems particularly well-positioned to take advantage of this opportunity.

References

- Baron, David P. 1989. "A Noncooperative Theory of Legislative Coalitions." *American Journal of Political Science* 33: 1048-84.
- Baron, David P. 1991. "A Spatial Theory of Government Formation in Parliamentary Systems." *American Political Science Review* 85: 137-65.
- Baron, David P., and Daniel Diermeier. 2001. "Elections, Governments, and Parliaments in Proportional Representation Systems." *Quarterly Journal of Economics* 116: 933-67.
- Baron, David P., and John A. Ferejohn. 1989. "Bargaining in Legislatures." *American Political Science Review* 89: 1181-206.
- Battaglini, Marco, and Thomas R. Palfrey. 2007. "The Dynamics of Distributive Politics." Social Science Working Paper 1273, California Institute of Technology.

- Bazerman, Max H., Jared R. Curhan, Don A. Moore, and Kathleen L. Valley. 2000. "Negotiation." *Annual Review of Psychology* 51: 279-314.
- Bazerman, Max H., Elizabeth A. Mannix, and L.L. Thompson. 1988. "Groups as Mixed-Motive Negotiations." *Advances in Group Processes* 5: 195-216.
- Bereby-Meyer, Yoella, and Muriel Niederle. 2005. "Fairness in Bargaining." *Journal of Economic Behavior and Organization* 56: 173-86.
- Bolton, Gary E., and Axel Ockenfels. 2000a. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90: 166-93.
- Bolton, Gary E., and Axel Ockenfels. 2000b. "Strategy and Equity: An ERC Analysis of the Güth van Damme Game." *Journal of Mathematical Psychology* 42: 215-26.
- Browne, Eric C., and John Fendreis. 1980. "Allocating Coalition Payoffs by Conventional Norm: An Assessment of the Evidence for Cabinet Coalition Situations." *American Journal of Political Science* 24: 753-68.
- Browne, Eric C., and Mark N. Franklin. 1973. "Aspects of Coalition Payoffs in European Parliamentary Democracies." *American Political Science Review* 67: 453-69.
- Burhans, David. T., Jr. 1973. "Coalition Game Research. A Reexamination." *American Journal of Sociology* 79: 389-408.
- Cosmides, Leda, and John Tooby. 1992. *Cognitive Adaptations for Social Exchange*. New York: Oxford University Press.
- Curhan, Jared R., and Alex Pentland. 2007. "Thin Slices of Negotiation: Predicting Outcomes from Conversational Dynamics Within the First Five Minutes." *Journal of Applied Psychology* 92: 802-11.
- Davis, Douglas D., and Charles A. Holt. 1993. *Experimental Economics*. Princeton: Princeton University Press.
- Diermeier, Daniel. 2006. "Coalition Government." In *Oxford Handbook of Political Economy*, eds. Barry Weingast, and Donald Wittman. New York: Oxford University Press.
- Diermeier, Daniel, and Sean Gailmard. 2006. "Self-Interest, Inequality, and Entitlement in Majoritarian Decision-Making." *Quarterly Journal of Political Science* 1: 327-50.
- Diermeier, Daniel, and Antonio Merlo. 2000. "Government Turnover in Parliamentary Democracies." *Journal of Economic Theory* 94: 46-79.

- Diermeier, Daniel, and Rebecca Morton. 2005. "Experiments in Majoritarian Bargaining." In *Social Choice and Strategic Decisions: Essays in Honor of Jeffrey S. Banks*, eds. David Austen-Smith, and John Duggan. Berlin and New York: Springer.
- Diermeier, Daniel, Hülya Eraslan, and Antonio Merlo. 2003. "A Structural Model of Government Formation." *Econometrica* 71: 27-70.
- Diermeier, Daniel, Hülya Eraslan, and Antonio Merlo. 2007. "Bicameralism and Government Formation." *Quarterly Journal of Political Science* 2: 1-26.
- Diermeier, Daniel, Roderick I. Swaab, Victoria Husted Medvec, and Mary C. Kern. 2008. "The Micro Dynamics of Coalition Formation." *Political Research Quarterly* 61: 484-501.
- Eraslan, Hülya, and Antonio Merlo. 2002. "Majority Rule in a Stochastic Model of Bargaining." *Journal of Economic Theory* 103: 31-48.
- Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114: 769-816.
- Fiorina, Morris P., and Charles R. Plott. 1978. "Committee Decisions under Majority Rule: An Experimental Study." *American Political Science Review* 72: 575-98.
- Fréchette, Guillaume R., John H. Kagel, and Steven F. Lehrer. 2003. "Bargaining in Legislatures: An Experimental Investigation of Open versus Closed Amendment Rules." *American Political Science Review* 97: 221-32.
- Fréchette, Guillaume R., John H. Kagel, and Massimo Morelli. 2005a. "Gamson's Law Versus Non-Cooperative Bargaining Theory." *Games and Economic Behavior* 51: 365-90.
- Fréchette, Guillaume R., John H. Kagel, and Massimo Morelli. 2005b. "Nominal Bargaining Power, Selection Protocol, and Discounting in Legislative Bargaining." *Journal of Public Economics* 89: 1497-517.
- Fréchette, Guillaume R., John H. Kagel, and Massimo Morelli. 2005c. "Behavioral Identification in Coalitional Bargaining: An Experimental Analysis of Demand Bargaining and Alternating Offers." *Econometrica* 73: 893-939.
- Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior* 6: 347-69.
- Gamson, William A. 1961. "A Theory of Coalition Formation." *American Sociological Review* 26: 373-82.
- Gamson, William A. 1964. "Experimental Studies of Coalition Formation." In *Advances in Experimental Social Psychology* Vol. 1, ed. Leonard Berkowitz. San Diego, CA: Academic Press.

- Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 75: 367-88.
- Huffaker, David, Roderick I. Swaab, and Daniel Diermeier. In press. "The Language of Coalition Formation in Online Multiparty Negotiations." *Journal of Language and Social Psychology*.
- Kagel, John H., and Katherine Willey Wolfe. 2001. "Test of Fairness Models based on Equity Considerations in a Three-Person Ultimatum Game." *Experimental Economics* 4: 203-20.
- Laver, Michael, and Norman Schofield. 1990. *Multiparty Government: the Politics of Coalition in Europe*. Oxford, UK: Oxford University Press.
- Mannix, Elizabeth A., and Sally Blount-White. 1992. "The Impact of Distributive Uncertainty on Coalition Formation in Organizations." *Organizational Behavior and Human Decision Processes* 51: 198-219.
- Martin, Lanny W., and Randolph T. Stevenson. 2001. "Government Formation in Parliamentary Democracies." *American Journal of Political Science* 45: 33-50.
- McKelvey, Richard D. 1991. "An Experimental Test of a Stochastic Game Model of Committee Bargaining." In *Laboratory Research in Political Economy*, ed. Thomas R. Palfrey. Ann Arbor: University of Michigan Press.
- Messick, David M. 1993. "Equality as a Decision Heuristic." In *Psychological Perspectives on Justice: Theory and Applications*, eds. Barbara A. Mellers, and Jonathan Baron. New York: Cambridge University Press.
- Morelli, Massimo. 1999. "Demand Competition and Policy." *American Political Science Review* 93: 809-20.
- Ochs, Jack, and Alvin E. Roth. 1989. "An Experimental Study of Sequential Bargaining." *American Economic Review* 79: 355-84.
- Raiffa, Howard. 1982. *The Art and Science of Negotiation*. Cambridge, MA: Belknap.
- Roth, Alvin E. 1995. "Bargaining Experiments." In *The Handbook of Experimental Economics*, eds. John Kagel, and Alvin Roth. Princeton: Princeton University Press.
- Roth, Alvin E., and Ido Erev. 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior* 8: 164-212.
- Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* 81: 1068-95.

- Schofield, Norman, and Michael Laver. 1985. "Bargaining Theory and Portfolio Payoffs in European Coalition Governments, 1945-1983." *British Journal of Political Science* 15: 143-64.
- Shapley, L.S. 1953. "A Value for n-person Games." In *Contributions to the Theory of Games*, Vol. 2, eds. Harold William Kuhn, and Albert William Tucker. Princeton, NJ: Princeton University Press.
- Swaab, Roderick I., Mary C. Kern, Daniel Diermeier, and Victoria Medvec. 2009. "Who Says What to Whom? The Impact of Communication Setting and Channel on Exclusion from Multiparty Negotiation Agreements." *Social Cognition* 27: 385-401.
- Swaab, Roderick I., Tom Postmes, Peter Neijens, Marius H. Kiers, and Adrie C.M. Dumay. 2002. "Multiparty Negotiations Support: The Role of Visualization's Influence on the Development of Shared Mental Models." *Journal of Management Information Systems* 19: 129-50.
- Taylor, Paul J., and Sally Thomas. 2005. *Linguistic Style Matching and Negotiation Outcome*: SSRN.
- Van Beest, Ilja, Gerben A. Van Kleef, and Eric Van Dijk. 2008. "Get Angry, Get Out: The Interpersonal Effects of Anger Communication in Multiparty Negotiation." *Journal of Experimental Social Psychology* 44: 993-1002.
- Volden, Craig, and Clifford J. Carrubba. 2004. "The Formation of Oversized Coalitions in Parliamentary Democracies." *American Journal of Political Science* 48: 521-37.
- Warwick, Paul V., and James N. Druckman. 2001. "Portfolio Salience and the Proportionality of Payoffs in Coalition Governments." *British Journal of Political Science* 31: 627-49.
- Wason, P. C. 1966. "Reasoning." In *New Horizons in Psychology*, ed. Brian Malzard Foss. Harmondsworth: Penguin.

Figure 28-1. Potential Coalitions and their Respective Payoffs

Coalition	Payoff
{A}	0
{B}	0
{C}	0
{A,B}	\$118,000
{A,C}	\$84,000
{B,C}	\$50,000
{A,B,C}	\$121,000

ⁱ The author would like to thank Randy Stevenson, Jamie Druckman, and Skip Lupia for their helpful comments as well as Alison Niederkorn and especially Justin Heinze for his research support and the Ford Motor Company Center for Global Citizenship at the Kellogg School of Management for additional funding. All remaining errors and omissions are my own.

ⁱⁱ For overviews on the development of the field see Laver and Schofield (1990) and Diermeier (2006).

ⁱⁱⁱ For overviews see Gamson (1964) or Burhans (1973). See also Fiorina and Plott (1978) for experiments in spatial settings.

^{iv} Baron and Ferejohn (1989) also consider open rules, where (nested) amendments to a proposal are permitted before the final vote. See also Baron (1989, 1991) for applications to coalition government. There are other variants of the Baron-Ferejohn model but they have played no role in the study of coalition government.

^v Baron and Ferejohn (1989) discuss this case in a footnote.

^{vi} Note, however, that the fact that the presence of a “selfish” player was announced may have changes the nature of the game.

^{vii} The concept of a “fair share” is consistent with Bolton and Ockenfels’s (2000) ERC (“equity, reciprocity, and competition”) theory. While the ERC approach has been successful in explaining two-player bargaining behavior, recent experimental results with three-person games (Kagel and Wolfe 2001; Bereby-Meyer and Niederle 2003) are inconsistent with the ERC approach.

^{viii} An alternative approach has been suggested by Battaglini and Palfrey (2007). The results are not directly comparable as they analyze a bargaining protocol with an endogenous status quo. Battaglini and Palfrey also find a significant number of equal distributions, but in their frame-work they can be explained by risk-aversion.

^{ix} This amount varies from culture to culture. In experiments conducted by Roth et al. (1991) the modal offer varied between 40% of the pay-off in Jerusalem, Israel and 50% in Pittsburgh.

^x The same effects can be found in bilateral bargaining games (Diermeier and Gailmard n.d.).

^{xi} As far as we know there have been no direct tests of proto-coalition bargaining as in the model proposed by Baron and Diermeier (2001), though there have been tests using field data (Carrubba and Volden 2004). Diermeier and Morton (2005) also directly test the proportionality heuristic in a laboratory setting and find no support.

^{xii} Notice that once voter reject “unfair” offers, proposers may act optimally in offering more than prescribed by the Baron-Ferejohn model.

^{xiii} Another important difference is the systematic use of deception. While this is a crucially important topic in many studies it does not play a major role in the research on coalition bargaining.

29. Negotiation and Mediation

Daniel Druckman

Knowledge about negotiation and mediation comes primarily from laboratory experiments. The question asked in this chapter is: What value is added by experiments for understanding processes of elite bargaining? This question is addressed in the sections to follow. After describing the international negotiation context, I provide a brief overview of the experimental approach. Then, key studies on distributive and integrative bargaining are reviewed followed by examples of experiments that capture complexity without forfeiting the advantages of experimental control. The chapter concludes with a discussion of the value added by experiments.

The Context

Negotiating in the international context takes several forms. It occurs from a distance and face-to-face, deals with multiple complex issues, and includes bilateral, multilateral, and global participation. National leaders often make demands or exchange proposals from a distance. Well-known examples include the bilateral exchanges between the United States and the Soviet Union over the 1948-49 blockade of Berlin, between Kennedy and Khrushchev in 1962 over Soviet missile bases in Cuba, and between Carter and Khomeini concerning American hostages in Iran in 1979-80. Leaders and their representatives also confront each other face to face to discuss their interests over security, monetary and trade, or environmental issues. These meetings may take the form of summits, such as the 1986 meeting between Reagan and Gorbachev in Reykjavik, or more protracted meetings, such as the long series of talks between their countries'

representatives over arms control, beginning with the Strategic Arms Limitations Talks (SALT) and winding up with the Strategic Arms Reduction Talks (START).

Many negotiations occur among more than two nations. They may occur between blocs, such as the North Atlantic Treaty Organization (NATO)-Warsaw Pact discussions in the 1970s over mutual and balanced force reductions. They may take the form of three or four-party discussions at which simultaneous bilateral negotiations take place. One example is the discussion among Iceland, Norway, Russia, and the Faroe Islands over fishing rights in the North Atlantic: While Icelandic negotiators rejected the Russo-Norwegian offer, they reached an agreement with the Faroes; the Norwegians protested this agreement. Other examples of simultaneous bilateral talks come from the area of free trade such as NAFTA (Canada, the United States and Mexico) and between Singapore, Australia, and the United States. And, from the area of security comes the example of the 1962-63 partial nuclear test ban talks between the United States, Great Britain, and the Soviet Union. Negotiations also occur in multilateral settings, where representatives from many nations gather for discussions of regional, continental, and global issues. Notable examples are the Uruguay Round of the General Agreements on Tariffs and Trade (GATT), the negotiations establishing the European Community (the Single European Act), the ongoing discussions among members of the Organization for Security and Cooperation in Europe and among members of the UN Security Council, the talks that led to the Montreal Protocol on ozone depletion, and the discussions that resulted in the Rio Declaration on Environment and Development.

These examples share a number of features, including high stakes and high drama, multi-level bargaining at the intersection between intra- and international actors, the need to manage

complexity, accountability to national constituencies, implications for national foreign policies, experienced negotiators, ratification (for treaties), and concern for proper implementation of agreements. Many of these features are captured by detailed case studies of particular negotiations. They are difficult to study in experiments, even when attempts are made to simulate real-world settings. What then can experiments offer? This question is addressed by showing that experiments provide added value to the contributions made by case studies. Knowledge gained from experiments is presented following a discussion of the relevance of the experimental method to the study of elite bargaining.

1. The Experimental Method

Most elite bargainers are career professionals.¹ They differ in many ways from the subjects who serve as role players in negotiation experiments. Among the differences are experience, stakes, issue-area expertise, actors in bureaucratic politics, implementation challenges, and accountability to government agencies or to international organizations. On the other hand, there are some similarities: similar bargaining choice dilemmas, decision-making processes, tactical options, and intra-team or coalition dynamics. A question is whether we emphasize the differences or the similarities. The case for differences is made by Singer and Ray (1966) who pointed out several “critical” dimensions of difference that exist between the small group laboratory where decision-making experiments are conducted and the more complex bureaucracies in which policy-making decisions are made. The argument for similarities was made by Bobrow (1972), “We should move rapidly toward treating phenomena that cross national lines as instances of phenomena that occur in several types of social units. Accordingly, alliances become coalitions; negotiations between nations become bargaining; foreign policy

choices become decision making” (55). Both arguments have merit. An emphasis on differences is reflected in the case-study tradition of research. Similarities are assumed when experimentalists argue for relevance of their findings to the settings being simulated. In this chapter, I will discuss implications of experimental research to elite bargaining in the international setting. The value added by this research reinforces the “similarities” perspective. It does not, however, diminish the importance of the differences listed above. I have argued elsewhere for striking a balance between the respective strengths of case-oriented and experimental research on negotiation, and return to this idea in later sections.

Two of the more vigorous proponents of the experimental method on bargaining argued that: “Abstraction and model building are necessary to reduce the problem to manageable proportions. The experimental method can contribute to the process of identifying critical variables and the nature of their roles in conflict situations” (Fouraker and Siegel 1963, 207). It is this heuristic function of experiments that may be most valuable. It tells us where to look – which variable or cluster of variables accounts for negotiation behavior? By the early 1970s, we had already accumulated a storehouse of knowledge about bargaining from the laboratory (see Rubin and Brown 1975). Spurred on by the early accomplishments, bargaining researchers have added additional storehouses to the “property.” A steady increase of publications, and the founding of several journals and professional associations dedicated to the topic, has resulted in a cross-disciplinary epistemic community of researchers. The list of variables explored has expanded considerably, frameworks and models abound, and innovative methodologies have emerged. An attempt is made to capture these developments without losing sight of the challenge of relevance to elite bargaining.

The experimental literature is organized into two parts. One, referred to as *distributive bargaining*, reviews the findings from a large number of earlier studies conducted primarily during the period from 1960-1980. Another, referred to as *integrative bargaining* or problem solving, discusses a smaller number of experiments conducted more recently. This distinction, suggested originally by Walton and McKersie (1965), has been shown to resonate as well with processes of elite bargaining in the international context (Hopmann 1995). Both sections trace the development of research from the earliest experiments, which provided a spark for later studies. Relevance to elite bargaining is demonstrated with results obtained from analyses of distributive and integrative bargaining processes *in situ*.

2. Distributive Bargaining

Early experiments on negotiation focused primarily on distributive bargaining. This refers to situations in which the interests of the bargainers are in conflict and where each attempts to obtain the largest share of whatever is being contested. These contests often conclude with agreements on outcomes somewhere between the bargainers' opening positions. Bargaining researchers have been concerned with factors that influence: 1) whether an agreement will be reached, 2) the amount of time needed to reach an agreement, 3) the type of agreement reached (as equal or unequal concessions), and 4) the bargainers' satisfaction with the agreement and their willingness to implement it.

Offer Strategies

A large number of experiments were conducted in the 1960s, spurred by Siegel and Fouraker's (1960) findings about levels of aspirations or goals. They found that "... the bargainer who (1) opens negotiations with a high request, (2) has a small rate of concession, (3)

has a high minimum level of expectation, and (4) is very perceptive and quite unyielding, will fare better than his opponent who provides the base upon which these relative evaluations were made” (93). These findings suggest that toughness pays. A question raised is: Do these results apply to a wide range of bargaining situations? The question of generality was evaluated by a flurry of experiments conducted during the 1960s and 1970s. Many of those experiments examined a bargainer’s change in offers made in response to the other’s concession strategy.

The bargaining studies did not support the generality of the Siegel-Fouraker conclusion. They showed that a hard-offer strategy works only under certain conditions: when the bargainer does not have information about the opponent’s payoffs and there is substantial time pressure. The chances of a settlement increased when the opponent used a soft-or intermediate-offer rather than a hard-offer strategy. The best overall strategy for obtaining agreement is matching: it resulted in greater bargainer cooperation than unconditional cooperation, unconditional competition, or partial cooperation (for a review of the findings, see Hamner and Yukl 1977).

These findings support Osgood’s (1962) well-known argument that cooperation will be reciprocated rather than exploited. Referred to as *graduated reciprocation in tension reduction* (GRIT), Osgood reasoned that unilateral concessions would remove the main obstacle to an opponent’s concession making, which is distrust. The initial concession would set in motion a cycle of reciprocated or matched concessions. Support for this hypothesis was obtained by Pilisuk and Skolnick (1968): They found that the best strategy is one that uses conciliatory moves in the beginning and then switches to matching. Referred to also as tit-for-tat, the matching strategy has been shown to be effective in producing agreements over the long term (Axelrod 1980). It has been demonstrated by Crow (1963) in an inter-nation simulation (without

control groups) and in a case of the partial nuclear test ban talks, referred to as the “Kennedy experiment” (Etzioni 1967). The test-ban case was also used as a setting for hypothesis testing.

The distinction between hard and soft bargaining strategies has informed analyses of simulated and actual international negotiations. Using a coding system referred to as bargaining process analysis (BPA), Hopmann and Walcott (1977) showed convergent findings from a simulation and case study of the 1962-63 Eighteen-Nation Disarmament Conference leading to the agreement on the Partial Nuclear Test Ban Treaty. They found that increased tensions in the external environment increased: 1) the amount of hostility in mutual perceptions, 2) the proportion of hard relative to soft bargaining strategies, 3) the employment of commitments, 4) the ratio of negative to positive affect, and 5) the ratio of disagreements to agreements in substantive issues under negotiation. The increase in hostile attitudes and the toughening of positions detracted from arriving at agreements. These results provide additional refuting evidence for the Siegel-Fouraker conclusion that “toughness pays.” More recently, Lytle and Kopelman (2005) showed that better distributive outcomes occurred when bargainers’ threats (hard strategy) were combined with friendly overtures (softer strategy).

Another example of convergence between findings obtained in the laboratory and from a real-world case is provided by Druckman and Bonoma (1976) and Druckman (1986). The former study was conducted with children bargaining as buyers and sellers. The results showed that disappointed expectations for cooperation led bargainers to adjust their concessions, leading to a deadlock. The latter study was conducted with documentation from a military base rights case and analyzed with the BPA coding categories. The results also showed that negotiators adjusted their offer strategy when expectations for cooperation were disappointed: The time series

analysis revealed a pattern of switching from soft to hard moves when the discrepancy between own and other's cooperation increased. The resulting mutual toughness led to an impasse which often produced a turning point in the talks. Referred to as "threshold-adjustment," this pattern has been demonstrated in eight cases of international negotiation (Druckman and Harris 1990). The similar findings obtained from a laboratory study with children and from a case study with professional negotiators bolster the argument for generality of negotiation processes. Taken together, the Hopmann-Wolcott and Druckman studies underscore the relevance of laboratory research for understanding real-world elite bargaining.

The Bargaining Environment

The early bargaining experiments focused primarily on the other bargainer's concession behavior and such features of the setting as time pressure and atmosphere. Bargaining moves or concessions and outcomes were the key dependent variables. Other independent variables studied during this period were group representation, pre-negotiation experience, and orientation. Blake and Mouton's Human Relations Training Laboratory served as a venue for experiments on the impact of representing groups on resolving intergroup disputes (e.g., Blake and Mouton 1961, 1962). They concluded that the commitments triggered by representation are a strong source of inflexibility in negotiation. Although fraught with problems of inadequate controls, their work stimulated a fruitful line of investigation. These studies showed that representation effects on flexibility are contingent on the stakes: High stakes in the form of payoffs or reputations produce stronger effects than low stakes. However the pressures can be offset by the setting and by salient outcomes: Private negotiations (Organ 1971) and fair solutions (Benton

and Druckman 1973) serve to increase a representative's flexibility, leading to agreements rather than impasses.

But, it is also the case that other variables have been found to have stronger impacts on bargaining behavior than representation. One of these variables is a bargainer's orientation as competitive or cooperative: Particularly strong effects were obtained when the orientation was manipulated in instructions (Summers 1968; Organ 1971). Another was pre-negotiation experience: The key distinction is between studying issues (more flexibility) and strategizing (less flexibility) prior to negotiation; whether this activity was done unilaterally (own team members only) or bilaterally (both teams) made little difference (Bass 1966; Druckman 1968). Of the ten independent variables compared in a meta-analysis on compromising behavior, these produced the strongest effect sizes (Druckman 1994). Representation and accountability ranked sixth and seventh respectively.ⁱⁱ

Many of the experimental findings call attention to the importance of reciprocity in bargaining. Regarded as a norm (Gouldner 1960), reciprocal moves reflect a principle distributive principle of equality. Strong support for this principle is found in Deutsch's (1985) experiments on distributive justice. His laboratory subjects showed a strong preference for equal distributions with little variation across subject populations or tasks. These results were explained in terms of the interdependent structure of the tasks and aspirations for cooperation or solidarity. Similar results were obtained by Druckman and Albin (in press) for outcomes of peace agreements. In their comparative study, equality mediated the relationship between the conflict environment and the durability of the agreements. Thus, again, convergent findings were obtained between laboratory and case analyses.

Summary

The discussion in this section reveals contributions made by experiments to our understanding of distributive bargaining processes. Three contributions are highlighted. One concerns the effects of different bargaining strategies: the best strategy is likely to consist of generous opening moves combined with matching or reciprocating concessions. Another deals with the impact of various features of the bargaining situation: bargaining orientation and pre-negotiation have been shown to have the strongest impacts on compromising behavior. A third contribution is to more complex negotiation settings: simulations of international negotiations have demonstrated the deleterious effects of stress, the impasse-producing impact of asymmetric power structures, and opportunities provided by impasses for progress. Relevance of experiments is bolstered further by convergent results obtained from case studies, including the systematic analysis of single cases (Etzioni 1967; Hopmann and Walcott 1977; Beriker and Druckman 1996) and comparative analyses of a relatively large number of cases (Druckman and Harris 1990; Druckman 2001; Druckman and Albin in press). The convergence between findings obtained on equality in the laboratory and from analyses of peace agreements is particularly striking.

3. Integrative Bargaining

Another perspective on negotiation emerged and influenced experimentation beginning in the 1970s. This perspective, referred to as integrative bargaining, describes a situation where parties attempt to jointly enlarge the benefits available to both (or all) so that they may gain a larger value than attained through compromise. The focus is on positive sum rather than non-zero sum (mixed motive) outcomes. Conceived of initially by Follett (1940) and developed

further by Walton and McKersie (1965) and Rapoport (1960), the approach gained momentum with the popular writing of Fisher and Ury (1981), the decision-theory approach taken by Raiffa (1982), and Zartman and Berman's (1982) diagnosis-formula-detail perspective. These theoretical contributions have been complemented by a number of experiments designed to provide empirical foundations for the concept.

The most compelling argument for integrative bargaining comes from experimental findings. Results obtained across many experiments conducted by Pruitt and his colleagues show that the average correlation between joint profits and distributive (integrative) bargaining behavior is inverse (direct) and statistically significant. Yet, despite these findings, bargainers tend to prefer distributive approaches. Why does this occur? An answer to this question is provided by Pruitt and Lewis (1977). Parties tend to imitate each other's distributive behavior: Threats elicit counter-threats and bargainers are less willing to make concessions to the extent that the other's demands are viewed as being excessive. Thus, bargaining may "gravitate toward a distributive approach because it requires only one party to move the interaction in that direction, while the firm resolve of both parties is needed to avoid such movement" (Pruitt and Lewis 1977, 170). Their experiments explored the strategies that encourage this mutual "resolve."

Integrative Strategies

The key finding is that integrative bargaining (and high joint outcomes) depends on flexible rigidity. This approach consists of remaining relatively rigid with respect to goals but flexible with regard to the strategies used to attain these goals. The research reveals how this approach may be achieved. Two strategies have been found to be effective. One is referred to as

heuristic trial and error (HTE): each bargainer seeks the other's reactions to a variety of proposals and options, known also as trial balloons. Another is information exchange: each bargainer asks for and provides information about needs and values. Both strategies convey flexibility; they contrast to the distributive strategies that convey rigidity in the process of seeking favorable outcomes. Their effectiveness depends, however, on maintaining a problem-solving orientation throughout the bargaining process. They also depend on mutual resolve in maintaining high aspirations, referred to as rigidity with respect to goals.

Each of these strategies also has limitations. Further experiments revealed the challenges. The effectiveness of HTE depends on either knowing or constructing the integrative options from available information. When these options are not known, bargainers must re-conceptualize the issues or try new approaches. This requires some form of information exchange. Discussing values and priorities can provide insight into the joint reward structure but can also backfire when the information reveals other incompatibilities. Thus, the new information can either move the process forward or embroil the parties in a continuing impasse. The former consequence is more likely to occur when both bargainers commit to a problem solving orientation. However, that orientation, which also requires mutual resolve, can result in impasses as well. What then can bargainers do to encourage positive and discourage the negative impacts of these strategies? Insights come from the results of more recent experiments.

The systematic construction of alternative offer packages has been shown to be an effective HTE strategy. Referred to as multiple equivalent simultaneous offers (MESOs), this strategy consists of presenting the other bargainer with alternative packages of roughly equal perceived value. It has been found to be more beneficial than single package offers:

Experimental results showed that: 1) more offers were accepted, 2) more satisfaction was expressed with the accepted offers, 3) the presenting bargainer was viewed as being more flexible, and, 4) when both bargainers used MESOs, they were more likely to reach an agreement considered as an efficient outcome (Medvic, Leonardelli, Galinsky, and Claussen-Schulz 2005). The *multiple* and *simultaneous* features of MESOs provide an opportunity to compare alternative packages and then choose one from the menu. This is likely to enhance the perceived value of the choice made by an opposing bargainer. The *equivalent* feature assures that each choice provides the same value to the presenting bargainer: This is especially the case when a well-defined scoring system is used. And, since the priorities of both bargainers must be understood prior to constructing the packages, the chosen offer is likely to be an integrative outcome (Medvic and Galinsky 2005).

The process of constructing MESOs includes developing an understanding of both own and others priorities: Different priorities are a basis for trades, known as logrolling. This is also an element in the information exchange process. But, information exchange goes further. It encourages bargainers to explore each other's underlying interests, values, and needs, which may be regarded as root causes of the conflict. The sensitivities involved in such deep probes can escalate the conflict as Johnson (1967) discovered in his hypothetical court case experiment and Muney and Deutsch (1968) reported in their social issues simulation. The information received by bargainers in a role-reversing condition revealed incompatibilities that led to impasses. However, the information did produce greater understanding of the other's positions: More attitude and cognitive change occurred in the role reversing than in the self-presentation conditions of their experiments (see also Hammond, Todd, Wilkins, and Mitchell 1966). These

findings suggest that short-term bargaining failures may not impede long-term efforts at resolving conflicts: The insights achieved during the information exchange process may be valuable in diagnosing the other's intentions (Van Kleef, van Dijk, Steinel, Harinck, and Van Beest 2008). They may also contribute to future workshops designed to reduce hostility and negative stereotypes (see Rouhana 2000). The diagnoses and reduced hostility may, in turn, pave the way for eventual integrative outcomes.

An interesting chicken-and-egg problem emerges from these experimental findings: Does reduced conflict depend on achieving integrative agreements or do integrative agreements depend on a relaxation of tensions? A way around the dilemma of causality is to assume that the problem is circular: Context and process are intertwined.ⁱⁱⁱ Hopmann and Walcott's (1977) simulation findings show that more agreements occur when tensions are reduced. Thus, context influences outcomes. Integrative processes and outcomes have been shown to improve relationships in case studies on the durability of peace agreements (Druckman and Albin in press), computer simulations that model distributive and integrative negotiations (Bartos 1995), and experimental simulations that compare facilitation with fractionation approaches to negotiation (Druckman, Broome, and Korper 1988). Thus, processes and outcomes influence context. These findings suggest that negotiation processes are embedded in contexts. Integrative bargaining is facilitated by amiable relationships between the bargainers. It is also encouraged when negotiators maintain a problem-solving orientation throughout the bargaining process.

Problem Solving Orientation

Recall that a problem-solving orientation increases the effectiveness of integrative strategies and the chances of obtaining favorable joint outcomes (Pruitt and Lewis 1977). A key

question is how to sustain this orientation. Experiments have provided some clues, including priming, vigorous cognitive activity, and mediation. Results from a meta-analysis of bargaining experiments showed that primed orientations produced stronger effects on outcomes than unprimed (or selected) orientations. The strongest effects were produced by constituent or supervisor communications to adopt either cooperative or competitive strategies (e.g., Organ 1971). The weakest effects occurred when bargainers were selected on pre-negotiation attitudes toward cooperation (negotiation as a problem to be solved) or competition (negotiation as a win-lose contest) (e.g., Lindskold, Walters, and Koutsourais 1983). Thus, explicit communications or instructions help to sustain a problem-solving or competitive bargaining strategy.

A field study conducted by Kressel, Frontera, Forlenza, Butler, and Fish (1994) compared mediators who used either a problem-solving style (PSS) with a settlement-oriented style (SOS) in child custody cases. The former (PSS) approach emphasizes the value of searching for information that can be used to reach an integrative agreement. The latter (SOS) emphasizes the value of efficient compromise solutions. Although SOS was preferred by most mediators, PSS produced better outcomes. It resulted in more frequent and durable settlements, and a generally more favorable attitude toward the mediation experience. A key difference between the approaches is effort. To be effective, PSS requires vigorous cognitive activity that includes three linked parts: persistent question asking, an analysis of sources of conflict, and a plan for achieving joint benefits. Thus, a structured and vigorous approach by negotiators or mediators is needed to sustain and reap the benefits from problem solving.

Progress toward integrative outcomes depends also on the perceived credibility of the mediator. Suggestions made by mediators are more likely to be taken seriously when the

implications for who gives up what are clear and do not favor one bargainer over the other. An experiment by Conlon, Carnevale, and Ross (1994) showed that mediators who suggest compromises (equal concessions by all bargainers) produced more agreements than those who made suggestions that could result in either asymmetrical (favoring one party more than another) or integrative (favoring both parties but complex) outcomes. The fair mediator is given latitude to encourage bargainers to take risks, such as avoiding the temptation to agree on the compromise outcome in favor of information exchange toward the more complex integrative agreement. An implication of these findings is that a mediator's activities can be phased with early suggestions geared toward compromise and later advice oriented toward agreements that provides more joint benefits than a compromise outcome. Thus, a trusted mediator can effectively encourage bargainers to sustain a problem-solving orientation.

The research has provided an answer to the question about the conditions for sustained problem solving. They combine strong communications from constituents or principals with mediator activities that enhance credibility and identify a solution that maximizes joint benefits. But, another question remains: Do the laboratory findings correspond to results obtained from studies of real-world negotiations? This section concludes with a discussion of research that addresses this question.

Problem Solving in situ

In her analyses of thirteen cases of historical negotiations involving the United States, Wagner (2008) found that the sustained use of problem-solving behaviors was strongly correlated with integrative outcomes. This finding corresponds to experimental results showing higher joint profits for bargainers who use problem-solving strategies. But, the case data also

provided an opportunity to refine this result in two ways. By dividing her cases into six stages, Wagner could examine trends in problem-solving behavior. Although sustained problem solving was needed for integrative outcomes, the best outcomes occurred for cases where these behaviors were frequent during the first two-thirds of the talks, particularly in the fourth stage. These outcomes were facilitated as well when negotiators developed formulas during the early stages: For example, identifying the terms of exchange to guide bargaining in a 1942 trilateral trade talk between the United States, United Kingdom, and Switzerland or identifying joint goals for each article in the 1951-1952 United States-Japan Administrative Agreement. These refinements extend the experimental results to processes that are less likely to occur in relatively brief laboratory simulations.

Other correspondences to laboratory results were obtained from Wagner's analyses. The professional negotiators in her cases bargained more than they problem solved: In only one case did the percentages of problem-solving statements exceed forty percent. This finding is echoed by Hopmann's (1995) appraisal of international negotiators' focus on relative gains and competition and concurs with results from comparative cases analyses on negotiations to resolve violent conflicts (Irmer and Druckman 2009). It corresponds to Pruitt and Lewis' (1977) observations about preferences for distributive bargaining among laboratory bargainers and to Kressel et al.'s (1994) tabulation of the relative frequencies of SOS (59% of the cases) to PSS (41%). Her finding that negotiators track each other's behavior by responding in-kind to the other's moves resonate with process findings from the experiments analyzed in De Dreu, Weingart, and Kwon's (2000) meta-analysis. In both the experiments and cases many negotiators reciprocated each other's problem-solving behaviors.^{iv} This sort of reciprocation by amateurs

and professionals was particularly likely for negotiators who understood negotiation strategies. Thus, educating negotiators about strategies – particularly the distinction between distributive and integrative bargaining – may increase their propensity to use approaches that are more likely to lead to better outcomes (see also Odell 2000 on this point).

Summary

The discussion in this section highlights contributions made by experiments to our understanding of integrative bargaining. A challenge for both negotiators and mediators is to resist the temptation to engage in distributive bargaining. Early experiments showed that two strategies are likely to be effective. One, referred to as HTE, consists of seeking the other's reaction to a variety of alternative proposals. Effectiveness is increased when this process is done systematically in the form of MESOs. Another strategy, referred to as information exchange, consists of asking for and providing information about values and needs. Effectiveness depends on the extent to which the new information facilitates the search for integrative solutions; it is reduced when the information reveals additional incompatibilities between negotiators. The effectiveness of these strategies also depend on relationships between the negotiators, their willingness to sustain a problem-solving orientation throughout the process, and the perceived credibility of mediators. These findings come from experiments. They correspond to results obtained from analyses of complex, real-world negotiations. Those analyses also refine the experimental results by capturing trends in problem-solving behavior through stages and calling attention to the usefulness of formulae as guides to bargaining.

4. Capturing Complexity in the Laboratory

The correspondences obtained between laboratory and case findings on distributive and integrative bargaining suggest that these are general processes that occur in a variety of negotiating situations. An example is the importance of a sustained problem-solving orientation throughout the bargaining process: sustained problem solving led to integrative agreements. An advantage of the laboratory is to provide a platform for causal analysis. These analyses do not, however, reveal details of processes that occur in particular real-world negotiations. An advantage of case studies is that they provide an opportunity to record – often through the lens of content analysis categories – the details. Conclusions from these analyses may take the form of such statements as, an increased prevalence of problem-solving behavior in the middle stages (as compared to early and late stages) occurred in cases that resulted in integrative agreements. A challenge for analysts is to find a way of combining the advantages of the experimental laboratory with those of detailed case studies. The discussion in this section addresses that challenge.

The challenge is met by incorporating complexity in laboratory environments without forfeiting the key advantages of experimental design, namely, random assignment and controls. An attempt to address this issue was made by the early research on the Inter-Nation Simulation (INS). This ambitious program of research encompassed a wide variety of studies ranging from abstract models (e.g., Chadwick 1970) to simulation experiments (e.g., Bonham 1971). Yet, despite this variety, researchers shared the goal of producing a valid corpus of knowledge about international relations. Their collective success was documented by ratings of correspondence among INS findings and other sources of data including anecdotal reports, experiments and field

studies: The results were mixed (Guetzkow and Valadez 1981).^v More relevant perhaps for this chapter were the efforts made by the INS researchers to design complex laboratory environments that permitted detailed data collections and analyses. These environments are examples of how complexity can be incorporated in laboratory settings. They served as models for the systematic comparisons performed with negotiation simulations.

Frameworks

Frameworks have been constructed to organize the various influences and processes of international negotiation. These include preconditions, issues, background factors, conditions, processes, outcomes, and implementation of agreements (see Sawyer and Guetzkow 1965; Randolph 1966). The frameworks have been useful for organizing literature reviews (Druckman 1973), chapters in edited books on negotiation (Druckman 1977), case studies (Ramberg 1978), scenario construction (Bonham 1971), teaching and training courses (Druckman 1996, 2006), and as guides for web-based computer-generated advice on impasse resolution (Druckman, Harris, and Ramberg 2002). As organizing devices, these frameworks are primarily synthetic or integrative. The question of interest is how to bridge the gap between frameworks, which capture complexity, and experiments, which investigate causal relations among a few variables. This question is addressed by research on the situational levers of negotiating flexibility.

Situational Levers

This project was an attempt to reproduce the dynamics of actual cases in a randomized experimental design. Key variables from the Sawyer-Guetzkow framework – 16 in all – were incorporated in each of four stages (pre-negotiation planning, setting the stage, the give and take, the endgame) of a conference, referred to as “Cooperative Measures to Reduce the Depletion of

the Ozone Layer” (COMROD).^{vi} Drawing on earlier studies, hypotheses were developed about the timing and effects of each variable on negotiating behavior: For example, issue positions were linked or not linked to political ideologies in the pre-negotiation stage; a deadline did or did not exist in the endgame. Three experimental conditions were compared: all variables in each stage were geared in the direction of hypothesized flexibility (issues not linked to ideologies; deadline); variables geared toward inflexibility (issues linked to ideologies; no deadline), and a mixed condition proceeding from hypothesized inflexibility in the early stages (issues linked to ideologies) to hypothesized flexibility in later stages (a deadline). This was a 3 (flexibility condition) x 4 (stages) experimental design. The simulation was replicated with two samples, environmental scientists at a Vienna-based international organization and diplomats at the Vienna Academy of Diplomacy (Druckman 1993). By bringing elite bargainers into the laboratory, the relevance of the findings for international negotiation is increased.

The analytical challenge presented by this project was to unpack the set of variables in each of the stages. By situating a negotiation process in a complex setting where many variables operate simultaneously, it is difficult to distinguish among them in terms of their relative impacts on negotiating behavior. In technical terms, manipulated variables within the stages are not orthogonal to each other: The design is suited to evaluate the main effects of alternative types of packages and stages, including the interaction between them. Thus, it was necessary to use another analysis strategy. Turning to an earlier literature on psychological scaling, the method of pair comparisons, discussed by Guilford (1950), was appropriate. This method was adapted to the task of comparing pairs of variables in each negotiating stage with regard to their impact on flexibility. The judgment took the form of: Does having an ideology make you more or less

flexible than being your nation's primary representative? A set of computations results in weights for the set of variables in each stage and experimental condition. The weighted variables are then arranged in trajectories, showing the key factors that operated in each negotiating stage and experimental condition by sample (scientists or diplomats).

Similar results were obtained for the scientist and diplomat samples. They suggest the conditions likely to produce flexibility or intransigence. Flexibility is more likely during the early stages when negotiators are in the role of delegate advisors rather than as primary representatives for the delegation. They are likely to be flexible in later stages when the talks are not exposed to media attention and when they have unattractive alternatives. Intransigence was more likely in the early stages when they prepared strategies rather than studying the issues. It was likely in later stages when wide media coverage occurred and when attractive alternatives were available (see also Druckman and Druckman 1996).

Additional experiments provided insights into the timing of moves and the role of mediation (Druckman 1995). Negotiators reached agreement more often when their opponent showed flexibility following a period of intransigence. This finding adds the variable of timing to the idea of firm but flexible behavior (Pruitt and Lewis 1977). Early firmness followed by later flexibility worked best. Suggestions made by mediators had less impact on flexibility than other factors designed into the situation, for example, media coverage, alternatives. Thus, a mediator's advice may be a weaker lever than other aspects of the designed situation.^{vii} It may, however, be the case that advice has more impact when combined with diagnosis and analysis as shown in the next section.

Electronic Mediation

A three-part model of mediation was evaluated in the context of electronic mediation. Referred to as Negotiator Assistant (NA), the web-based mediator implements three functions -- diagnosing the negotiating situation, analyzing causes of impasses, and providing advice to resolve the impasse (Druckman et al. 2002). It was used in conjunction with a simulated negotiation that captured the issues leading to the 2003 war in Iraq. Student role players negotiated seven issues involving weapons inspection, border troops, and terrorism. Three experimental comparisons were performed to assess the impact of NA: compared to no mediation (experiment one), advice only (experiment 2), and a live mediator (experiment three). Results showed that significantly more agreements were obtained in each experiment when negotiators had access to NA between rounds (Druckman, Druckman, and Arai 2004). The e-mediator produced more agreements than a scripted live mediator despite an expressed preference for the latter. These results demonstrate the value of electronic tools for supporting complex negotiations. The study also demonstrates the value of embedding experiments in complex simulations that resemble the types of real-world cases described in the opening section of this chapter. Whether these tools help to resolve impasses in those cases remain to be evaluated.^{viii}

5. Comparing Simulations with Cases

Explicit comparisons of data obtained from simulation and cases were performed by Hopmann and Walcott (1977) and Beriker and Druckman (1996). Results obtained in the former study showed that stress produced similar dysfunctional effects in a laboratory simulation of the partial nuclear test ban talks and in the actual negotiation. Real word-simulation correspondences

were also obtained in the latter study on power asymmetries in the Lausanne Peace negotiations (1922-3). Content analyses of processes recorded in transcripts and generated by simulation role players showed both similarities and some dissimilarities. Both studies support the relevance of laboratory experiments for understanding real-world negotiations. Further support comes from two research streams on other negotiation processes.

Research on the interplay between interests and values illustrates complementary strengths of experiments and case studies. The studies were intended to evaluate propositions derived from the literature on the sociology of conflict (Druckman and Zechmeister 1973). Various experimental simulations (political decision making, prison reform, internal conflict resembling Cyprus, ecumenical councils) were used to evaluate some of the propositions: namely, concerning the link between values and interests (Druckman et al. 1988) and divisions on values within negotiating teams (Jacobson 1981). These propositions described static relationships between variables. Other propositions captured process dynamics and were demonstrated with a case of failed negotiation in the Philippines: namely, converging and diverging values through time (Druckman and Green 1995). The case study complemented the experiments; together, the methods provided a comprehensive assessment of the theory-derived propositions.

More recent research on turning points illustrates the difference between retrospective and prospective analysis. A set cases was used to trace processes leading toward and away from critical departures in each of 34 completed negotiations on security, trade, and the environment (Druckman 2001). A key finding is that crises trigger turning points. This and other findings provided insights into the way that turning points emerged in past cases of elite bargaining. The

findings were less informative with regard to predicting their occurrence. Thus, we designed two experiments to learn about the conditions for producing turning points. Both experiments showed that the social climate (perceptions of trust and power) of the negotiation moderated the effects of precipitating factors on outcomes. The impact of crises on turning points depends on the climate surrounding the negotiation. The experiments identified an important contingency in the emergence of turning points (Druckman, Olekalns, and Smith 2009).

These lines of research demonstrate the value of multi-methods. They highlight complementarities between experiments and case studies. Used together, the methods provide the dual advantages of hypothesis testing and contextual interpretation as well as the strengths of both prospective causal analysis and retrospective comparisons.

Simulations and Cases in Training

Another contribution made by experiments is to skills training for elite negotiators, including diplomats and foreign-service officers. The training procedures emphasize connections between research and practice. This is done, first, by presenting the research-based knowledge in the form of narratives and, second, by conducting a sequence of exercises that are linked to the knowledge. The narratives are summaries of findings on each of 16 themes: for example, emotions, culture, experience, flexibility. Key insights are highlighted with special attention paid to counter-intuitive findings and prescriptions for practice: for example, quick agreements are often sub-optimal; thus, discourage rapid concession exchanges, particularly in negotiations between friends.

The exercises represent each of four negotiating roles: analyst, strategist, performer, and designer. In their roles of analyst and strategist, trainees apply relevant narratives to such real-

work cases as Panama Canal and the Korean Joggers. In their role as performer, they participate in the security issues simulation described above in the section on e mediation. And, as designers, they construct their own scenarios for training exercises.^{ix} The training has been conducted across four continents and may have subtly infused experimental knowledge into professional negotiating practices.^x

6. Conclusion: Experiments as Value-Added Knowledge

A salient finding obtained across negotiating domains is that bargainers prefer to compete for relative gains rather than problem solve for joint gains. This preference was observed in laboratory experiments (Pruitt and Lewis 1977), field studies of mediated child custody cases (Kressel et al. 1994), and both historic (Wagner 2008) and more recent (see Hopmann 1995) cases of international negotiation. Interestingly, it was also shown to occur in cross-cultural bargaining experiments with children, even when higher payoffs would be obtained from cooperative than maximizing differences strategies (McClintoick and Nuttin 1969). An important question is how to change preferences from less to more optimal bargaining strategies. Answers are provided from experimental findings. These answers are shown to have relevance also for negotiating in real-world settings.

Two approaches, based on the idea of flexible rigidity, have been evaluated. One, referred to as heuristic trial and error, consists of gauging the other's reactions to a variety of proposals and options. When this is done systematically, in the form of multiple equivalent simultaneous offers it is often effective. Another, referred to as information exchange, consists of asking for and providing information about needs and values. When guided by a credible mediator, the exchange process is often effective, particularly when the information revealed

helps direct the talks toward integrative outcomes. It is also the case, however, that the effectiveness of both approaches depends on maintaining a problem-solving orientation throughout the negotiation. Sustained problem solving has been shown to be important in the laboratory and *in situ*.

Convergent findings about problem solving attest to the value of experiments as platforms for producing generalized knowledge. They do not, however, attest to their value in capturing context-specific knowledge. Case analyses provided additional information about the frequency of problem solving during different stages and about the value of formulae. A question of interest is whether this sort of contextual detail would be discovered in more complex laboratory simulations. An answer is found in the research on situational levers of flexibility and electronic mediation.

The complex environmental negotiation used to study situational levers provided more specific results on staged processes than other, less complex, experimental platforms. The security issues simulation used to study electronic mediation allowed role players to experience electronic and live mediator functions. Both studies show that a balance can be struck between rigor and relevance. Further, complementary advantages of experiments and cases were evident in the work on turning points, where both retrospective and prospective analyses were performed, and on values and interests, where both hypothesis testing and holistic approaches were used. Thus, experimental knowledge adds to our understanding of the case examples described at the beginning of the chapter. More compelling perhaps are training applications. The gap between experiments with students and cases with professionals is bridged by the use of experimental knowledge in diplomatic training programs. To the extent that these programs

influence the way that diplomats negotiate, experimental findings contribute directly to elite bargaining.

References

- Axelrod, Robert. 1980. "More Effective Choice in the Prisoner's Dilemma." *Journal of Conflict Resolution* 24: 379-403.
- Bartos, Otto J. 1995. "Modeling Distributive and Integrative Negotiations." *The Annals of the American Academy of Political and Social Science* 542: 48-60.
- Bass, Bernard. 1966. "Effects on the Subsequent Performance of Negotiators of Studying Issues or Planning Strategies Alone or in Groups." *Psychological Monographs* 80, whole number 614.
- Benton, Alan, and Daniel Druckman. 1973. "Salient Solutions and the Bargaining Behavior of Representatives and Non-representatives." *International Journal of Group Tensions* 3: 28-39.
- Beriker, Nimet, and Daniel Druckman. 1996. "Simulating the Lausanne Peace Negotiations 1922-1923: Power Asymmetries in Bargaining." *Simulation & Gaming* 27: 162-83.
- Blake, Robert R., and Jane S. Mouton. 1961. "Loyalty of Representatives to Ingroup Positions during Intergroup Competition." *Sociometry* 24: 177-83.
- Blake, Robert R., and Jane S. Mouton. 1962. "Comprehension of Points of Communalities in Competing Solutions." *Sociometry* 25: 56-63.
- Bobrow, Davis B. 1972. "Transfer of Meaning across National Boundaries. In *Communication in International Politics*, ed. Richard J. Merritt. Urbana, IL: University of Illinois Press.
- Bonham, Matthew. 1971. "Simulating International Disarmament Negotiations." *Journal of Conflict Resolution* 15: 299-315.
- Chadwick, Richard W. 1970. "A Partial Model of National Political-Economic Systems." *Journal of Peace Research* 7: 121-32.
- Conlon, Donald E., Peter Carnevale, and William H. Ross. 1994. "The Influence of Third Party Power and Suggestions on Negotiation: The Surface Value of a Compromise." *Journal of Applied Social Psychology* 24: 1084-113.
- Crow, Wayman J. 1963. "A Study of Strategic Doctrines using the Inter-Nation Simulation." *Journal of Conflict Resolution* 7: 580-9.

- De Dreu, Carsten K.W., Laurie R. Weingart, and Seungwoo Kwon. 2000. "Influence of Social Motives on Integrative Negotiation: A Meta-Analytic Review and Test of Two Theories." *Journal of Personality and Social Psychology* 78: 889-905.
- Deutsch, Morton. 1985. *Distributive Justice: A Social-Psychological Perspective*. New Haven, CT: Yale University Press.
- Druckman, Daniel. 1968. "Prenegotiation Experience and Dyadic Conflict Resolution in a Bargaining Situation." *Journal of Experimental Social Psychology* 4: 367-83.
- Druckman, Daniel. 1973. *Human Factors in International Negotiations: Social-Psychological Aspects of International Conflict*. Sage Professional Paper in International Studies Vol. 2, no. 02-020. Beverly Hills and London: Sage.
- Druckman, Daniel, ed. 1977. *Negotiations: Social-Psychological Perspectives*. Beverly Hills CA: Sage.
- Druckman, Daniel. 1986. "Stages, Turning Points, and Crises: Negotiating Military Base Rights: Spain and the United States." *Journal of Conflict Resolution* 30: 327-60.
- Druckman, Daniel. 1993. "The Situational Levers of Negotiating Flexibility." *Journal of Conflict Resolution* 37: 236-76.
- Druckman, Daniel. 1994. "Determinants of Compromising Behavior in Negotiation: A Meta-Analysis." *Journal of Conflict Resolution* 38: 507-56.
- Druckman, Daniel. 1995. "Situational Levers of Position Change: Further Explorations." *The Annals of the American Academy of Political and Social Science* 542: 61-80.
- Druckman, Daniel. 1996. "Bridging the Gap between Negotiating Experience and Analysis." *Negotiation Journal* 12: 371-83.
- Druckman, Daniel. 2001. "Turning Points in International Negotiation: A Comparative Analysis." *Journal of Conflict Resolution* 45: 519-44.
- Druckman, Daniel, and Cecilia Albin. In press. "Distributive Justice and the Durability of Peace Agreements." *Review of International Studies*.
- Druckman, Daniel, and Thomas V. Bonoma. 1976. "Determinants of Bargaining Behavior in a Bilateral Monopoly Situation II: Opponent's Concession Rate and Similarity." *Behavioral Science* 21: 252-62.
- Druckman, Daniel, and James N. Druckman. 1996. "Visibility and Negotiating Flexibility." *Journal of Social Psychology* 136: 117-20.

- Druckman, Daniel, and Noam Ebner. 2008. "Onstage or Behind the Scenes? Relative Learning Benefits of Simulation Role-Play and Design." *Simulation & Gaming* 39: 465-97.
- Druckman, Daniel, and Richard Harris. 1990. "Alternative Models of Responsiveness in International Negotiation." *Journal of Conflict Resolution* 34: 234-51.
- Druckman, Daniel, and P. Terrence Hopmann. 1989. "Behavioral Aspects of Negotiations on Mutual Security." In *Behavior, Society, and Nuclear War, Volume One*, eds. Philip. E. Tetlock, Jo L. Husbands, Robert Jervis, Paul C. Stern, and Charles Tilly. New York: Oxford University Press.
- Druckman, Daniel, and Victor Robinson. 1998. "From Research to Application: Utilizing Research Findings in Negotiation Training Programs." *International Negotiation* 3: 7-38.
- Druckman, Daniel, and Kathleen Zechmeister. 1973. "Conflict of Interest and Value Dissensus: Propositions in the Sociology of Conflict." *Human Relations* 26: 449-66.
- Druckman, Daniel, Benjamin J. Broome, and Susan H. Korper. 1988. "Value differences and Conflict Resolution: Facilitation or Delinking?" *Journal of Conflict Resolution* 32: 489-510.
- Druckman, Daniel, and Justin Green 1995. "Playing Two Games: Internal Negotiations in the Philippines." In *Elusive Peace: Negotiating an End to Civil Wars*, ed. I. William Zartman. Washington, DC: Brookings.
- Druckman, Daniel, Richard Harris, and Bennett Ramberg. 2002. "Computer-Assisted International Negotiation: A Tool for Research and Practice." *Group Decision and Negotiation* 11: 231-56.
- Druckman, Daniel, James N. Druckman, and Tatsushi Arai. 2004. "e-Mediation: Evaluating the Impacts of an Electronic Mediator on Negotiating Behavior." *Group Decision and Negotiation* 13: 481-511.
- Druckman, Daniel, Mara Olekalns, and Philip L. Smith. 2009. "Interpretive Filters: Social Cognition and the Impact of Turning Points in Negotiation." *Negotiation Journal* 25: 13-40.
- Etzioni, Amitai. 1967. "The Kennedy Experiment." *Western Political Quarterly* 20: 361-80.
- Fisher, Roger, and William Ury. 1981. *Getting to Yes: Negotiating Agreement Without Giving In*. Boston: Houghton Mifflin.
- Follett, Mary Parker. 1940. "Constructive Conflict." In *Dynamic Administration: The Collected Papers of Mary Parker Follett*, eds. Henry C. Metcalf, and L. Urwick. New York: Harper & Brothers Publishers.

- Fouraker, Lawrence E., and Sidney Siegel. 1963. *Bargaining Behavior*. New York: McGraw-Hill.
- Gouldner, Alvin W. 1960. "The Norm of Reciprocity: A Preliminary Statement." *American Sociological Review* 25: 161-78.
- Guetzkow, Harold, and Joseph J. Valadez 1981. *Simulated International Processes: Theories and Research in Global Modeling*. Beverly Hills, CA: Sage.
- Guilford, J.P. 1950. *Psychometric Methods*. New York: McGraw-Hill.
- Hammond, Kenneth R., Frederick J. Todd, Marilyn Wilkins, and Thomas O. Mitchell. 1966. "Cognitive Conflict between Persons: Applications of the 'Lens Model' Paradigm." *Journal of Experimental Social Psychology* 2: 343-60.
- Hamner, W. Clay, and Gary A. Yukl. 1977. "The Effectiveness of Different Offer Strategies in Bargaining." In *Negotiations: Social-Psychological Perspectives*, ed. Daniel Druckman. Beverly Hills, CA: Sage.
- Hopmann, P. Terrence. 1995. "Two Paradigms of Negotiation: Bargaining and Problem Solving." *The Annals of the American Academy of Political and Social Science* 542: 24-47.
- Hopmann, P. Terrence, and Charles Walcott. 1977. "The Impact of External Stresses and Tensions on Negotiations." In *Negotiations: Social-Psychological Perspectives*, ed., Daniel Druckman. Beverly Hills, CA: Sage.
- Irmer, Cynthia, and Daniel Druckman. 2009. "Explaining Negotiation Outcomes: Process or Context?" *Negotiation and Conflict Management Research* 2: 209-35.
- Jacobson, Dan. 1981. "Intraparty Dissensus and Interparty Conflict Resolution." *Journal of Conflict Resolution* 25: 471-94.
- Janda, Kenneth. In press. "A Scholar and a Simulation Ahead of their Time." *Simulation & Gaming*.
- Johnson, David W. 1967. "The Use of Role Reversal in Intergroup Competition." *Journal of Personality and Social Psychology* 7: 135-42.
- Kressel, Kenneth, Edward A. Frontera, Samuel Forlenza, Frances Butler, and L. Fish. 1994. "The Settlement Orientation versus the Problem-Solving Style in Custody Mediation." *Journal of Social Issues* 50: 67-84.
- Lindskold, Svenn, Pamela S. Walters, and Helen Koutsourais. 1983. "Cooperators, Competitors, and Response to GRIT." *Journal of Conflict Resolution* 27: 521-32.

- Lytle, Anne L., and Shirli Kopelman. 2005. "Friendly Threats? The Linking of Threats and Promises in Negotiation." Paper presented at the annual meeting of the International Association for Conflict Management, Seville, Spain.
- McClintock, Charles, and J. Nuttin. 1969. "Development of Competitive Behavior in Children Across Two Cultures." *Journal of Experimental Social Psychology* 5: 203-18.
- Medvic, Victoria H., and Adam D. Galinsky. 2005. "Putting More on the Table: How Making Multiple Offers Can Increase the Final Value of the Deal." *Negotiation*. Article Reprint No. N0504B: 3-5.
- Medvic, Victoria H., G.L. Leonardelli, Adam D. Galinsky, and A. Claussen-Schulz. 2005. "Choice and Achievement at the Bargaining Table: The Distributive, Integrative, and Interpersonal Advantages of Making Multiple Equivalent Simultaneous Offers." Paper presented at the annual conference of the International Association for Conflict Management, Seville, Spain.
- Muney, Barbara F., and Morton Deutsch. 1968. "The Effects of Role-Reversal during the Discussion of Opposing Viewpoints." *Journal of Conflict Resolution* 12: 345-56.
- Odell, John. 2000. *Negotiating the World Economy*. Ithaca, NY: Cornell University Press.
- Organ, Dennis. 1971. "Some Variables Affecting Boundary Role Behavior." *Sociometry* 34: 524-37.
- Osgood, Charles E. 1962. *An Alternative to War or Surrender*. Urbana, IL: University of Illinois Press.
- Pilisuk, Marc, and Paul Skolnick. 1968. "Inducing Trust: A Test of the Osgood Proposal." *Journal of Personality and Social Psychology* 8: 121-33.
- Pruitt, Dean G., and Steven A. Lewis. 1977. "The Psychology of Integrative Bargaining." In *Negotiations: Social-Psychological Perspectives*, ed. Daniel Druckman. Beverly Hills CA: Sage.
- Raiffa, Howard. 1982. *The Art and Science of Negotiation*. Cambridge, MA: Harvard University Press.
- Ramberg, Bennett. 1978. *The Seabed Arms Control Negotiation: A Study of Multilateral Arms Control Conference Diplomacy*. Denver: University of Denver Press.
- Randolph, Lillian. 1966. "A Suggested Model of International Negotiation." *Journal of Conflict Resolution* 10: 344-53.
- Rapoport, Anatol. 1960. *Fights, Games, and Debates*. Ann Arbor: University of Michigan Press.

- Rouhana, Nadim N. 2000. "Interactive Conflict Resolution: Issues in Theory, Methodology, and Evaluation." In *International Conflict Resolution After the Cold War*, eds. Paul C. Stern, and Daniel Druckman Washington, DC: National Academy Press.
- Rubin, Jeffrey Z., and Bert R. Brown 1975. *The Social Psychology of Bargaining and Negotiation*. New York: Academic Press.
- Sawyer, Jack, and Harold Guetzkow. 1965. "Bargaining and Negotiation in International Relations. In *International Behavior: A Social-Psychological Analysis*, ed. Herbert C. Kelman. New York: Holt.
- Siegel, Sidney, and Lawrence E. Fouraker. 1960. *Bargaining and Group Decision Making*. New York: McGraw-Hill.
- Singer, J. David, and Paul Ray. 1966. "Decision-Making in Conflict: From Interpersonal to International Relations." *Menninger Clinic Bulletin* 30: 300-12.
- Summers, David A. 1971. "Conflict, Compromise, and Belief Change in a Decision-Making Task." *Journal of Conflict Resolution* 12: 215-21.
- Van Kleef, Gerben A., Eric van Dijk, Wolfgang Steinel, Fieke Harinck, and Ilja Van Beest. 2008. "Anger in Social Conflict: Cross-Situational Comparisons and Suggestions for the Future." *Group Decision and Negotiation* 17: 13-30.
- Wagner, Lynn M. 2008. *Problem-Solving and Bargaining in International Negotiations*. Leiden, the Netherlands: Martinus Nijhoff Publishers.
- Walton, Richard E., and Robert B. McKersie. 1965. *A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System*. New York: McGraw-Hill.
- Zartman, I. William, and Maureen R. Berman. 1982. *The Practical Negotiator*. New Haven, CT: Yale University Press.

ⁱ The career professional designation would apply to civil servants or foreign-service officers but not to political appointees. The latter are usually appointed for relatively brief stints as special envoys or ambassadors. Their term in office typically ends when administrations change.

ⁱⁱ Other variables in the analysis included time pressure, initial position distance, the opponent's strategy, large vs. small issues, framing, and visibility.

ⁱⁱⁱ For a discussion of these issues in the area of arms control, see Druckman and Hopmann (1989).

^{iv} This finding corresponds to the preference for distributive equality obtained in experiments by Deutsch (1985) and in case analyses by Druckman and Albin (in press). These studies were discussed above.

^v Among the strongest correspondences was the arousal of identification with the fictitious nations. Role-players identified with their laboratory groups in a manner similar to decisions makers in the system being simulated. These findings bolster the case for external validity of laboratory studies. They also arbitrated between alternative theories of ethnocentrism. However, it is also the case that these results may be due to the role-players' theories about how they may be expected to behave. Referred to as demand characteristics, the role expectations of simulation

participants is an alternative explanation for the correspondences obtained between simulation and field findings (see Janda in press).

^{vi} This simulation was modeled on the 1992 global environmental declaration on environment and development negotiated in Rio de Janeiro.

^{vii} Stress may, however, play a larger role in real-world negotiation. Results obtained from a random-design field experiment conducted at the DC small claims court showed that contesting parties did not respond to such manipulated aspects of the situation as the configuration of furniture or orientation instructions. These findings were interpreted in terms of the overwhelming effects of emotions on decisions.

^{viii} Evidence for convergent validity of the NA diagnostic function was provided by comparisons of predicted with actual outcomes obtained in nine cases. Computed-diagnosed outcomes corresponded to actual outcomes in eight of the nine international negotiations (Druckman et al. 2002).

^{ix} Recent findings show that designers learn more about negotiation concepts than classmates who role-play those designs (Druckman and Ebner 2008). Learning advantages occur as well for students exposed to the original journal articles used for the narrative summaries (Druckman and Robinson 1998).

^x The complete training package with evaluation results is presented in Druckman and Robinson (1998) and Druckman (2006).

30. The Experiment and Foreign Policy Decision Making

Margaret G. Hermann and Binnur Ozkececi-Taner

Snyder and his colleagues (Snyder, Bruck, and Sapin 1954; see also Snyder et al. 2002) in an influential monograph argued that people and process matter in international affairs and launched the study of foreign policy decision making. They contended that it is policymakers who perceive and interpret events and whose preferences become aggregated in the decision-making process that shape what governments and institutions do in the foreign policy arena. People affect the way that foreign policy problems are framed, the options that are considered, the choices that are made, and what gets implemented. To bolster their claims, Snyder and his associates brought research from cognitive, social, and organizational psychology to the attention of scholars interested in world politics. And they introduced the experiment as a potential methodological tool.

Because it remains difficult to gain access to policymakers and the policymaking process in real time, the experiment has become a tool for simulating “history” and for doing so under controlled conditions. It allows us to explore the causal relationships that occur between the nature of the people involved, the decision-making process, and the decisions that are made. In effect, experiments provide us with access to the temporal sequence that occurs during the decision-making process and help us study how the preferences policymakers bring to the process shape what happens both in terms of the nature of that process and the resulting decisions. The experiment also allows us to compare what happens when a particular problem, process, or type of leader is absent as well as present, there often being few records of instances

in government foreign policymaking that provide the controlled environment that an experiment does.

This tool became even more relevant to the study of foreign policymaking as a result of Simon's (1982, 1985) experiments indicating that decision making was not necessarily rational – that the nature of people's preferences matter and that rationality is bounded by how the people involved process information, what they want, the ways in which they represent the problem, their experiences, and their beliefs. In effect, decision makers “do not have unlimited time, resources, and information” to make choices that maximize their movement toward their goals (Chollet and Goldgeier 2002, 157). They “satisfice,” settling for the first acceptable option rather than pushing for ever more information and an even more optimal choice. “People are, at best, rational in terms of what they are aware of, [but] they can be aware of only tiny, disjointed facets of reality” (Simon 1985, 302). In effect, it becomes important to learn about foreign policymakers' “views of reality” as their preferences, so defined, shape their actions. Indeed, these early experiments showed: 1) that beliefs are “like possessions” guiding behavior and are only reluctantly relinquished (Abelson 1986), 2) that the more complex and important (the more life-like and ill-structured) a problem the less decision makers act like Bayesian information processors (Alker and Hermann 1971), and 3) that prior knowledge about a problem appears to shape cognition and focus decision making (Sylvan and Voss 1998). How decision makers define and represent problems may or may not match how an outside, objective observer views them.

In the rest of this chapter, we are going to examine in some detail a set of experiments that have built on what Snyder and his colleagues and Simon discovered. We will focus on three

important questions where the experiment as a methodology has been and is particularly useful in helping us gain insights into foreign policy decision making. 1) How do policymakers' predispositions shape policy preferences and behavior? 2) How does the way in which a problem is framed influence which options are viewed as relevant? 3) How do individual preferences become aggregated in the decisions of governments, that is, whose positions count and why?

It is important that the reader keep two caveats in mind as we describe experiments that address these questions. 1) In order to achieve some semblance of experimental realism in these experiments, the problems and scenarios that are used are patterned after real historical events with an attempt to have subjects face similar situations to those of actual foreign policymakers even to trying to simulate the time constraints and real-life pressures under which such decision makers work. 2) The experimenters have taken seriously checking that their manipulations work and that the simulated experience has fully engaged the so-called policymakers. Some of the unexpected insights that these experiments have afforded us have come because those participating as subjects have become so caught up in the experience.

1. Predispositions, Preferences, and Decisions

Schafer (1997) was interested in whether or not in conflict situations policymakers' images of their own and the so-called enemy influenced their preferences regarding the other country and their choice of strategy as well as their resulting foreign policy choices. Do policymakers' worldviews matter? By choosing to use the experiment to explore this question, Schafer was able to insure that policymakers' images temporally preceded his assessment of policy preferences and resulting decisions – that he could talk about causation and not just correlation.

Schafer used a 2 x 2 factorial design. His two independent variables were 1) the historical relationship between the countries involved in the conflict – friendly/cooperative vs. unfriendly/conflictual – and 2) the cultural similarity between the countries – similar vs. different. The “policymakers” in the experiment were 76 college students enrolled in an international relations course who were randomly assigned to the four treatment groups. Each was given briefing material concerning an international conflict between their fictitious country and another country. Both countries had similar military resources and the conflict held the possibility for dire consequences for each. The only differences among the briefing materials were the alternative views of the other country – whether their historical relationship had been generally friendly or unfriendly and if their cultures were similar or different. Participants were told that they were valued advisors to the president of their country and that their recommendations were important to the decision-making process. After reading the briefing materials, they were asked to indicate their current attitudes toward the other country, the general strategy that they believed their country should follow toward that other country, and the diplomatic, economic, and military responses their country should take given the situation.

Checks on the manipulation to make sure that those involved picked up on the two different image dimensions that differentiated their country from the other one revealed that they did. The results – using analysis of variance – indicated that both differences in historical relationship and in culture led to parallel types of attitudes toward the “other.” Those with the supposed negative historical relationship tended to view the other country as an enemy and were predisposed toward hostile and conflictual strategies in dealing with the other. Similarly, those where the opposing country was different in culture from them also viewed the other as an

enemy and believed that conflict was the best strategy to choose. When it came to the participants' decisions, there was a significant interaction between historical relationship and similarity of culture. Those in the condition in which participants had both a negative history and a different culture from the other country with which they were in conflict made decisions that were much more conflictual than were made in any other condition – the conflict was exacerbated. In effect, cultural differences mattered more in the policy choices made between perceived enemies than between perceived friends. Only, apparently, when the relationship with another country is perceived to be both historically and currently unfriendly and negative do cultural differences become important in decision making – the other country has to be viewed as both an enemy and “not like us” for attitudes to shape the choices that are made. Schafer's experiment suggests that images matter under certain circumstances.

What if policymakers have experience or expertise in dealing with a particular type of problem; can it predispose them to make different decisions than when such experience is lacking? Consider, for example, that the last three American presidents – Clinton, Bush, and Obama – have had little experience in dealing with foreign policy before coming to office. They have had less background on which to draw in the decision-making process when compared to presidents like Eisenhower or George H.W. Bush who spent their careers leading up to the White House dealing with foreign policy problems. Mintz (2004) sought to explore this issue and to do so with persons actually involved in the foreign policy arena, that is, military officers in the U.S. Air Force. By using such individuals as participants in his experiment, he could be assured that they had some prior experience in making policy decisions and that his experiment worked toward achieving experimental realism and external as well as internal validity.

Seventy-two military officers who were part of the command and instructional staff of the U.S. Air Force Academy were randomly assigned to deal with either a familiar or unfamiliar problem. Those in the experimental condition meant to involve a familiar problem were presented with a scenario asking them to deal with a military dispute that erupted between two small countries over control of a large uranium field leading one to invade the other and hold foreign citizens hostage. Four alternatives were posed to those in this condition; they could use force, engage in containment, impose sanctions, or do nothing. For the unfamiliar problem, participants were faced with choosing the site for a new naval base in the Pacific; four islands unknown to those involved were nominated for such a site and the four islands became the alternatives under consideration. The officers made their decisions using a computerized decision board which provided information regarding the various alternatives as well as recorded the nature of their search for this information and the strategies they used. The information that was presented indicated the alternative's likely political, economic, military, and diplomatic impacts – positive or negative – if selected.

Manipulation checks indicated that the officers dealing with the unfamiliar problem searched for considerably more information than those faced with the familiar problem. Indeed, the officers involved with the familiar problem latched onto a preferred alternative almost immediately and explored its implications before checking out any of the other alternatives. They took a shortcut based on their experience. Those officers dealing with the unfamiliar problem, on the other hand, kept seeking information regarding the implications of all four alternatives on political, economic, military, and diplomatic efforts, comparing and contrasting the effects of each alternative as they considered each domain. The officers used different decision-making

strategies in dealing with familiar and unfamiliar problems, their predispositions having more impact on preferences and choices with the familiar problem.

In the Schafer and Mintz experiments just described, the predispositions of the policymakers were primed by the experimental conditions. Is there any way to examine predispositions that do, indeed, precede participation in the experiment? Beer and his colleagues (Beer, Healy, and Bourne 2004) were interested in exploring this question and used a pre-test to assess the level of dominance-submissiveness of their experimental subjects using scales from the sixteen-personality-factor questionnaire (Institute for Personality and Ability Testing 1979). The subjects, who were students in an introduction to psychology course, were then randomly assigned to one of three experimental conditions where they viewed 1) material reminding them of the terrible costs of war through statistics and graphs out of World War I, 2) material concerning the way in which the appeasement of aggression can lead to war with a focus on the Munich conference and its aftermath, or 3) no introductory material. Following the priming conditions, all three groups read a common scenario with fictionalized countries modeled after the Falklands/Malvinas Islands crisis between Britain and Argentina in 1982. After reading the scenario, subjects were asked to indicate what the country representing Britain was likely to decide to do given their opponent's invasion of the islands; they chose among 15 different alternatives running from a peaceful attempt to resolve the crisis to all-out military retaliation. This decision was then followed by another regarding what each thought the country representing Argentina would do given "Britain's" response.

Beer et al. (2004) report a significant interaction effect between the conflict-focused prime (introductory material) and the personality of the subjects with regard to the type of

decision that was made. Participants high in dominance were more likely to choose to engage in conflict when primed to think about crisis and war regardless of whether the focus was costs or appeasement; when not primed, there was no difference between those high and low in dominance. The priming motivated the more dominant participants to want to take charge.

As can happen in experiments, Beer and his colleagues also discovered an unexpected result that helps us to link the Schafer and Mintz experiments described earlier and provides an important insight regarding the possible interaction between predispositions, preferences, and choices in foreign policy decision making. Forty-five percent of the participants correctly recognized the scenario when asked at the end of the experiment if it reminded them of any recent actual event. There was little difference between the decisions of subjects high and low in dominance for those who correctly identified the event in the scenario – they based their selection of an alternative on their perceptions of the historical scenario. But those high in dominance chose significantly more conflictual actions when they did not recognize the event. Prior knowledge was used when the situation seemed familiar whereas personality shaped decision making without such prior knowledge. In other words, personal predispositions are important in decision making unless the policymaker has prior experience or expertise – in such circumstances experience seems to take precedence in shaping preferences and the generation of alternatives as well as choices.

2. Frames and Expectations

An important part of the decision-making process is the identification and framing of problems. Indeed, this phase is often where people play a key role in shaping what will happen. Consider, for example, how the same event – September 11, 2001 – was identified and framed

differently by leaders in Britain and in the United States. Tony Blair announced at the Labour Party Conference just hours after the Twin Towers collapsed that we had just experienced a “crime against civilization” – police and the courts were the appropriate instruments for dealing with what had happened and justice was the value at stake; George W. Bush framed the event as “an attack against America” and pronounced a war on terror engaging the military and calling forth nationalism. These frames focused the attention of the respective policymaking communities and shaped expectations regarding who would be involved and the nature of the options that were available. Experiments have been usefully used both to demonstrate the importance of frames in defining foreign policy problems as well as their influence on shaping the options considered and resulting decisions. Three illustrate the effects of frames.

The first is an experiment intended to show how relatively easy it is for frames to change the way that a policymaker looks at a situation. Ross and Ward (1995) explored the influence of a frame on students who were recruited to play the Prisoner’s Dilemma game – urged on by their dorm leaders who selected them because of the dorm leaders’ beliefs that these students represented good examples of either “defectors” or “cooperators.” The recruits were randomly assigned to play either a Wall Street or a Community game. Those playing each of these games were given instructions indicating the nature of the game – either that they were members of a Wall Street trading firm or that they were part of a community not-for-profit organization. Subjects then made decisions based on a Prisoner’s Dilemma matrix of payoffs with the largest payoff for both players resulting from cooperation but the largest payoff for a single player coming through defection. Of interest was how the Wall Street and Community frames would shape expectations regarding underlying norms and rules in the early rounds of play.

And, sure enough, the frames had a significant effect on play. Two-thirds of those believing that they were playing the Community Game cooperated in the first round of play while only one-third of those perceiving themselves to be involved in the Wall Street Game cooperated. Indeed, this striking difference only increased over time as the Community Game facilitated more cooperation and the Wall Street Game more defections. In effect, the frames overrode the students' ideas about their own predispositions to cooperate or to defect in a game like the Prisoner's Dilemma. As the authors note, who gets to frame a problem gains influence over the policymaking process.

Building on the large body of non-experimental research suggesting that “democracies do not fight one another” (for an overview, see Chan 1997), Mintz and Geva (1993) designed an experiment to explore the effects of how a country is framed – as a democracy or non-democracy – on decisions regarding the use of force. Does being told that a country is one or the other almost automatically shape responses where the use of force is concerned? To enhance external validity, Mintz and Geva ran the experiment three times – once with U.S. college students, once with U.S. adults who were not college students but members of community organizations, and once with Israeli college students who had been or were members of their country's military. Participants were randomly assigned to one of two conditions – one in which the adversary in a hypothetical crisis was described as democratic and one in which the adversary was described as non-democratic. All read a scenario which was modeled after what happened in the First Persian Gulf War and involved one nation (the adversary) invading another as a result of a conflict over uranium and in the process taking hostages. In the course of the scenario, the adversary was framed as a democracy or a non-democracy. Participants were faced with three alternatives

regarding what the nature of their country's response should be: use force, set up a blockade, or do nothing militarily.

In all three runs of the experiment, participants picked up on the frame for the condition to which they were assigned. Indeed, they viewed the adversary as more dissimilar from their own country when it was the non-democracy than the democracy. Moreover, in each of the runs the participants significantly approved the use of force against a non-democracy more than against a democracy. This significant difference only held, however, for the use of force; for the blockade and "do nothing" alternatives, the frame was not determinative. And, to complement this hesitation regarding the use of force against a democracy, the participants perceived such use of force against a democracy as a foreign policy failure. The frame brought with it not only limitations on who could become a target of military force but how choosing to use force would be evaluated.

Probably the most influential experiments on framing in the literature on foreign policy decision making are those conducted by Kahneman and Tversky and have evolved into what is known as prospect theory (Tversky and Kahneman 1992; Kahneman and Tversky 2000; see also Farnham 1994; McDermott 2001). In essence, their findings indicate that how individuals frame a situation shapes the nature of the decision they are likely to make. If policymakers perceive themselves in a domain of gains (things are going well), they are likely to be risk averse. But if their frame puts them in the domain of losses (things are going poorly), they are likely to be more risk prone or risk seeking. Decisions depend on how the policymaker frames the problem. Critical for determining whether a decision maker finds him/herself in the domain of losses or gains is the individual's reference point or definition of the status quo. Problems arise when

decision makers face situations where there is a discrepancy between what is happening and their reference point. The direction of the discrepancy indicates whether the decision maker interprets the situation as involving gains or losses. Decision makers appear to be more sensitive to discrepancies that are closer to their reference point than to those further away and, perhaps more importantly, to view that “losing hurts more than a comparable gain pleases” (McDermott 2001, 29). As Foyle (1999, 266) has observed about American presidents, their foreign policy choices are much more affected by constituents and context when they fear “losing the public’s support of either the policy or the administration.” In a similar vein, Tomz (2009) talks about this phenomenon in terms of audience costs and has found that British members of parliament who are involved in the foreign policymaking process provide evidence of anticipating such costs as part of their decision making.

Generally the experiments conducted to document prospect theory and the effects of framing involve posing a choice dilemma to subjects with the options under consideration framed either, for example, in terms of how many people will die or in terms of how many will live. This difference in framing leads to taking the risky option when the focus is on how many people will die and the more conservative option when the frame focuses on the people who will live. Kowert and one of the authors (Kowert and Hermann 1997) wondered how generalizable the effects of loss and gain frames were across a variety of different types of problems. What happens, as is often the case in foreign policy decision making, when problems cross domains? In an experimental setting, students in introductory political science courses were faced with choice dilemmas focused around economic issues, political problems, or health concerns.

Interestingly, the results indicated that the framing effect postulated by prospect theory held for political and health problems but not for economic issues. Indeed, three-quarters of the subjects involved were risk-averse when political and health problems were framed as a gain and two-thirds were risk-acceptant for similar types of concerns when they were framed in terms of loss. But roughly two-thirds of the subjects were risk-averse in the economic domain regardless of the frame. They tended to seek out the certain option, even though a little less so when the problem was framed in the domain of losses as opposed to gains (sixty-three percent to seventy-eight percent respectively). Given the students participating in the study were enrolled in a public university, many working full-time jobs to afford an education, economic choice dilemmas may have been more real to them than those dealing with politics or health. This result suggests the importance of determining the reference point in considering the nature of the frame – for these particular students their reference point for economic concerns could have immediately put them in the domain of gains – “I have things under control and let’s not rock the boat.” Of course, it could also be the case that prospect theory has more of an effect for some kinds of problems than for others.

3. Aggregating Individual Preferences into Government Decisions

The studies we have just described have focused on individuals and how they make decisions. But governments are not single individuals nor do they act as a unit. Indeed, consider the wide array of entities that are responsible for making foreign policy, for example, party standing committees, military juntas, leader and advisors, cabinets, interagency groups, parliamentary committees, and loosely structured revolutionary coalitions. And the individuals comprising these entities do not always agree about what should happen with regard to foreign

policy. In fact, an examination of a range of foreign policy decisions (Beasley et al. 2001) has indicated that around seventy-five percent of the time those involved disagree about the nature of the problem, the options that are feasible, or what should happen. As Eulau (1968) argued so long ago: “although concrete political action is invariably the behavior of individual human actors, the politically significant units of action are groups, associations, organizations, institutions, coalitions, and other types of collectivities” (209). Of interest are the rules of aggregation for moving from decision making at the individual level – say, that of a leader – to that of a group or a coalition. Here, too, the experiment has proven to be a useful tool in helping us understand the foreign policymaking process. It has provided us with insights but with one caveat. The laboratory experiments to be reviewed here have generally focused on ad hoc groups, bringing participants together at one point in time, rather than on ongoing groups. More often than not, aggregations of individuals in the foreign policymaking process expect to interact across time.

One set of experiments on the rules of aggregation operating in groups has examined the roles that the members of the group play – in this case, whether or not those involved in the group are leaders and able to make decisions in the group itself such as occurs at summits or delegates who must check back with those they represent in the decision-making process – and the types of decisions and decision process that are likely to result (e.g., Hermann and Kogan 1968; Myers and Lamm 1976; Semmel 1982; Isenberg 1986; Brauer and Judd 1996). To reinforce the assigned roles in these experiments, students are generally randomly divided into groups depending on status either defined by some difference or stipulated in the instructions – for example, in one case upper classmen were designated as the leaders and underclassmen the

delegates and each experienced the other's status in an initial meeting where the leader's position could prevail were there disagreement. In these experiments, leaders and delegates usually met in pairs to start with, coming to some agreement as to what their joint positions were with regard to either a scenario or a set of dilemmas. The leaders and delegates then were formed into separate groups composed either of all leaders or all delegates to again wrestle with the problems. Following this interaction, each participant was asked to indicate if the deliberations with the others playing a similar role had any effect on their original dyadic decisions.

Delegates, having been told that they would be checking back with their leaders later, were found more likely to focus on achieving a compromise so all could gain some and not lose everything while leaders without the same need to report back were more likely to engage in extensive argument and debate and to choose one of their members' positions as well as to show the greatest change in their positions from where they started. Indeed, the delegate groups were more satisfied generally with the intergroup process when they compromised. For the leader groups, decision making with other leaders took precedence over choices made in their dyadic sessions with their delegates. But for the delegates, the dyadic decision making remained important when they met with other delegates.

The changes in position found among both the leader and delegate groups were at first thought to indicate the diffusion of responsibility that making decisions in a group affords – change is feasible because accountability for the decision can be spread and shared with little loss of face (e.g., Kogan and Wallach 1967; Vertzberger 1997). Such behavior has since been viewed as the result of persuasion; group members “with more radically polarized judgments and preferences invest more resources in attempts to exert influence and lead others” (Vertzberger

1997, 284) and, as a result, the discussion includes “persuasive arguments that the typical subject has not previously considered” (Myers and Lamm 1976, 611). Both these explanations seem appropriate to the findings with regard to leaders and delegates. The diffusion of responsibility notion, in effect, provides cover for the groups composed of delegates while persuasion appears a more appropriate explanation for the behavior of the groups comprised of leaders. And, indeed, the delegate groups chose one party’s position *only* when one of the delegates was so highly committed to his/her position that the group risked deadlock and returning to the leader with no decision.

Generally those experimenting with groups have viewed influence among members as going one way – usually the majority overpowering any minority. Groupthink (Janis 1982) is one example of such influence with stress leading to concurrence seeking as members want to remain a part of the group (see also Brown 2000; Hermann et al. 2001; Garrison 2003). But Moscovici (1976), who was interested in understanding social change rather than maintenance of the status quo, sought to understand when minorities could have an influence (see also Wood 2000; Kaarbo 2008). He became even more intent on his enterprise when he found some eight percent of the majorities in his experimental groups went along with the minority when he had a confederate minority declare that the blue color slide all were seeing was instead green – and, as he noted, they went along even though the majority was twice the size of the minority. Moscovici’s experimental research, and that inspired by him, has led to the postulation of the “dual process” model of social influence that indicates there is a difference in the influence process when engaged in by a majority or a minority. Specifically, when the majority is doing the influencing, the conflict is social and the minority only changes its position publicly but does not do so

privately whereas when the minority is doing the influencing, the conflict is cognitive with the members of the majority “trying to comprehend the deviant position” and in the process of such reconsideration changing their original position, even though their conversions have been found to be generally delayed and not done publicly (De Vries and De Dreu 2001, 3).

Minorities in groups who are committed to their positions and resolved that their plan or option is best generally cause the group to re-visit the problem of concern and if consistent and persistent can influence the position of the majority. Minorities also can use procedural manipulations to facilitate making the group more conducive to their positions – say by pushing for a change from majority rule to consensus or unanimity, by pushing for incremental adoption of policies, or by re-framing the problem (e.g., De Dreu and Beersma 2001; Kaarbo 2008). Consider an experiment by Kameda and Sugimori (1993).

These experimenters were interested in studying how the decision rule guiding discussion and policymaking in groups affected the nature of the process and choice as well as the impact of group composition – when opinion was split vs. not split – on decision making. More and more often in foreign policymaking, decisions are being made with a consensus orientation or unanimity decision rule. Arguments are made that this rule facilitates reaching the most feasible solution possible to a problem at the moment and one with which those involved are likely to comply as well as to implement (see Hagan et al. 2001; Hermann et al. 2001). The majority rule is viewed as likely to create a minority ready and willing to continue to push the majority to revisit any decision if it starts to go bad (see Beasley et al. 2001).

Kameda and Sugimori (1993) involved 159 Japanese students in a 2 x 2 between-subjects factorial design (majority vs. unanimity decision rule; opinion in group split vs. not split).

Subjects were presented with a simulated decision task and asked to make an initial decision indicating their preferences for what should be done. They were randomly assigned to groups so that one set of three-person groups had no differences among their initial preferences while the other set of three-person groups had two persons favoring one option and a third person with a different opinion. The groups then were asked to discuss the problem and reach a decision based on the decision rule governing their process. Each group was told that they had sole responsibility for the decision. The decision in the simulated task focused on whether or not the groups should continue to support a policy – or in this case, a person – that was not performing as desired. Would the groups focus on maintaining the original consensus or move to change policy?

The results indicated an interaction between the nature of the decision rule and cohesion around the choice of policy. The groups that had a member with a minority position who were guided by a majority rule were the most likely to move away from the previous consensus and change policy while that same type of group functioning under a unanimity decision rule were the least likely to change positions and, thus, the most likely to continue the original consensus policy regardless of negative feedback. Indeed, the groups with a member with a minority position who were guided by a unanimity rule were the most likely to persist in the presence of negative feedback of all four types of groups in the experiment. And members of these groups were the least likely to view the alternative options the groups had in a positive light – it was almost as if members said “the other options we have to consider are no better than what we are doing and may even lead to worse outcomes.” But, interestingly, the members in the minority evaluated the situation more negatively when they were in the majority rule condition than in the

unanimity condition. Being unable to overturn a vote and in the minority when decisions can be decided by majority rule led those in such positions to dislike the experience and to be dissatisfied with the decision that was made.

Groups governed by a unanimity decision rule appeared to move to maintain the status quo more than those with a majority rule. Moreover, such groups were even more constrained when they had majority and minority views represented among the members of the group. The easiest way to reach consensus was to stick to the original policy. The way to move to have policymakers consider changing policy appears to be through constituting groups with a majority decision rule and some split in the opinion among group members, the latter promoting discussion and breaking immediate consensus. But when the minority in such groups is overruled, they are often dissatisfied with what has happened and can force the group to re-visit the issue again if the new choice is not effective (for an application of these ideas to a set of foreign policy cases, see Beasley et al. 2001). Decision rules appear to differentially affect group process and, in turn, the nature of the decisions made.

An important type of group involved in the foreign policymaking process involves leaders and their advisors. Indeed, all leaders have advisors. We have discovered through case studies that it appears leaders pay more attention to some of their advisors than to others, often depending on how they have structured the setting and their leadership style (e.g., Preston 2001; Mitchell 2005). Redd (2002) sought to explore the influence of advisors where he could have control over the relationship of the advisors to the leader and chose to do so using an experiment. He was interested in “how and in what manner leaders obtain information from advisors” and the effect on policy (Redd 2002, 343). And to study this question, he manipulated the status of

leaders' advisors. What happens when one – in this case, the political advisor – has a higher status or importance in the leader's eyes than the others versus when all are of relatively equal status? Importance of advisors was manipulated in two ways: 1) the participants acting as leaders were randomly assigned to a condition in which the advisors were unequal in status vs. one where they were equal, and 2) the order in which the advisors' positions were presented to the participants (political advisor information being presented first vs. last).

Subjects were undergraduates taking political science courses who were presented with a foreign policy scenario involving “a military dispute between two small countries that erupted over control of a large uranium field” and resulted in one of these nations invading the other and taking foreign citizens hostage (Redd 2002, 345). Participants utilized a decision board in making their decisions, facilitating tracing their decision-making process. The decision board contained four advisors opinions on four options. The four types of advisors focused on political, economic, diplomatic, or military issues. The alternatives that the participants had to choose among were to do nothing, engage in containment, impose international sanctions, or use force. In the condition in which advisors were not equal, participants were told that their chief of staff or political advisor had been found to give them the best advice and the most successful policy recommendations. In the condition where advisors were equal, participants were told that all their advisors have given equally good advice and recommendations.

The results of the experiment showed that when the political advisor was viewed as giving the most relevant advice, participants tended to focus on the nature of that advisor's position and to use it as a guide for what other information was sought. And “decision makers tended not to select the alternative that the political advisor evaluated negatively, regardless of

the overall utility of that alternative” (Redd 2002, 355). Such was particularly the case when the political advisor’s recommendation came early in the process. However, when the advisors were viewed as equal in importance, the participants were forced to consider the recommendations of each advisor for each alternative. There was no shortcut or person to sensitize them for or against a specific option and, as a result, they were more likely to choose the alternative with the highest utility. Trusted advisors appear to have short-circuited their leaders’ search for information by shaping the way in which the problem was defined and the options that were considered. A wider search for information was necessary when advisors were perceived as rather equal in expertise and status.

4. In Conclusion

The illustrative experiments described and discussed here represent how this methodology is being and can be used to study foreign policy decision making. The experiment provides us with a venue to examine processes involved in the making of policy as well as the linkages between these processes and the resulting choices. It enables us to explore the temporal sequence of “what leads to and shapes what” in a way that case studies and observational studies do not. In effect, it provides a controlled environment in which to engage in process tracing – to follow the way in which the decision process evolves. Moreover, the experiment allows us to compare and contrast the presence and absence of the phenomenon under examination and not just to focus on the situations where it was present which is more generally the case in the literature. In effect, without the experiment we know very little about what would have happened if the phenomenon under study had not occurred. Is what actually occurred any different and, if so, how?

Yes, there are problems with experiments as there are with any methodology. Most of the experiments described above involved college students who are not currently policymakers – although a number of the studies did try to replicate or do their research with subjects who were older than college students or actually engaged in certain types of policymaking (such as Air Force officers). And the groups whose decision making was examined were set up in the laboratory and were not going to last longer than the experiment even though an important factor in ongoing policymaking groups is the “shadow of the future” phenomenon where groups have some sense that they will continue to interact in the future and, thus, temper their behavior with this expectation in mind. But even here, there were attempts to bring some flavor of the ongoing group into the laboratory, for instance, by having leader and delegate groups composed of upper- and lower-level students. Moreover, there are always questions regarding experimental realism; for instance, policymakers face more than one problem at a time, they are often not responsible for identifying or framing a problem, and they must delegate decisions to others to implement.

Nevertheless, the three experiments examined in each of the sections in this chapter have focused on questions that have been difficult to explore other than through experiments. They have provided us with new insights into how foreign policy decisions are made and with some intriguing ideas that can be explored further in case and observational studies. They have also laid the groundwork for scholars in the field to work on introducing more of the experimental method into their case and observational studies. Consider, for example, experiments embedded in surveys where manipulations and control conditions are randomly presented to those that are interviewed (e.g., Herrmann et al. 1999; Tomz 2007). And with case studies, there is growing interest in the use of structured, focused comparison where the same questions are systematically

asked of all cases and the cases that are selected to study do and do not exhibit the phenomenon being examined (e.g., Kaarbo and Beasley 1999; Mitchell 2005). Such techniques are intended to bring some internal validity into the field while at the same time facilitating experimental realism.

Because of the difficulty of gaining access to the people and processes involved in foreign policymaking, even in the United States, studying the political aspects of such policymaking has lagged institutional and structural analyses of why governments do what they do in international affairs. The experiment and these new variants of it offer those of us interested in understanding foreign policymaking a way to explore how people and process matter.

References

- Abelson, Robert P. 1986. "Beliefs Are Like Possessions." *Journal for the theory of Social Behavior* 16: 223-50.
- Alker, Henry A., and Margaret G. Hermann. 1971. "Are Bayesian Decisions Artificially Intelligent: The Effect of Task and Personality on Conservatism in Processing Information." *Journal of Personality and Social Psychology* 19: 31-41.
- Beasley, Ryan, Juliet Kaarbo, Charles F. Hermann, and Margaret G. Hermann. 2001. "People and Processes in Foreign Policymaking." *International Studies Review* 3: 217-50.
- Beer, Francis A., Alice F. Healy, and Lyle E. Bourne, Jr. 2004. "Dynamic Decisions: Experimental Reactions to War, Peace, and Terrorism." In *Advances in Political Psychology*, ed. Margaret G. Hermann. London: Elsevier.
- Brauer, Markus, and Charles M. Judd. 1996. "Group Polarization and Repeated Attitude Expressions: A New Take on an Old Topic." In *European Review of Social Psychology* Vol. 7, eds. Wolfgang Stroebe, and Miles Hewstone. Chicester, UK: John Wiley.
- Brown, Rupert. 2000. *Group Processes*. Oxford: Blackwell.
- Chan, Steve. 1997. "In Search of Democratic Peace: Problems and Promise." *Mershon International Studies Review* 41: 59-91.

- Chollet, Derek H., and James M. Goldgeier. 2002. "The Scholarship of Decision Making: Do We Know How We Decide?" In *Foreign Policy Decision-Making Revisited*, eds. Richard C. Snyder, H.W. Bruck, Burton Sapin, Valerie M. Hudson, Derek H. Chollet, and James M. Goldgeier. New York: Palgrave Macmillan.
- De Dreu, Carsten K.W., and Bianca Beersma. 2001. "Minority Influence in Organizations." In *Group Consensus and Minority Influence: Implications for Innovation*, eds. Carsten K.W. De Dreu, and Nanne K. De Vries. Oxford: Blackwell.
- De Vries, Nanne K., and Carsten K.W. De Dreu. 2001. "Group Consensus and Minority Influence: Introduction and Overview." In *Group Consensus and Minority Influence: Implications for Innovation*, eds. Carsten K.W. De Dreu, and Nanne K. De Vries. Oxford: Blackwell.
- Eulau, Heinz. 1968. "Political Behavior." In *International Encyclopedia of the Social Sciences* Vol. 12. New York: Macmillan.
- Farnham, Barbara, ed. 1994. *Avoiding Losses/Taking Risks: Prospect Theory in International Politics*. Ann Arbor: University of Michigan Press.
- Foyle, Douglas C. 1999. *Counting the Public In: Presidents, Public Opinion, and Foreign Policy*. New York: Columbia University Press.
- Garrison, Jean A. 2003. "Foreign Policymaking and Group Dynamics: Where We've Been and Where We're Going." *International Studies Review* 6: 177-83.
- Hagan, Joe D., Philip P. Everts, Haruhiro Fukui, and John D. Stempel. 2001. "Foreign Policy by Coalition: Deadlock, Compromise, and Anarchy." *International Studies Review* 3: 169-216.
- Hermann, Charles F., Janice Gross Stein, Bengt Sundelius, and Stephen G. Walker. 2001. "Resolve, Accept, or Avoid: Effects of Group Conflict on Foreign Policy Decisions." *International Studies Review* 3: 133-68.
- Hermann, Margaret G., and Nathan Kogan. 1968. "Negotiation in Leader and Delegate Groups." *Journal of Conflict Resolution* 12: 332-44.
- Herrmann, Richard K., Philip E. Tetlock, and Penny S. Visser. 1999. "Mass Public Decisions to Go to War: A Cognitive-Interactionist Framework." *American Political Science Review* 93: 553-73.
- Institute for Personality and Ability Testing. 1979. *Administrative Manual for the 16PF*. Champaign, IL: Institute for Personality and Ability Testing.
- Isenberg, Daniel J. 1986. "Group Polarization: A Critical Review and Meta-Analysis" *Journal of Personality and Social Psychology* 50: 1141-51.

- Janis, Irving L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. 2nd ed. Boston: Houghton Mifflin.
- Kaarbo, Juliet. 2008. "The Social Psychology of Coalition Politics." *International Studies Review* 10: 57-86.
- Kaarbo, Juliet, and Ryan K. Beasley. 1999. "A Practical Guide to the Comparative Case Study Method in Political Psychology." *Political Psychology* 20: 369-91.
- Kahneman, Daniel, and Amos Tversky. 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.
- Kameda, Tatsuya, and Shinkichi Sugimori. 1993. "Psychological Entrapment in Group Decision Making: An Assigned Decision Rule and a Groupthink Phenomenon." *Journal of Personality and Social Psychology* 65: 282-92.
- Kogan, Nathan, and Michael A. Wallach. 1967. "Risk Taking as a Function of the Situation, the Person, and the Group." In *New Dimensions in Psychology III*, eds. Michael A. Wallach, and Nathan Kogan. New York: Holt, Rinehart, and Winston.
- Kowert, Paul A., and Margaret G. Hermann. 1997. "Who Takes Risks? Daring and Caution in Foreign Policy Making." *Journal of Conflict Resolution* 41: 611-37.
- McDermott, Rose. 2001. *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy*. Ann Arbor: University of Michigan Press.
- Mintz, Alex. 2004. "Foreign Policy Decision Making in Familiar and Unfamiliar Settings." *Journal of Conflict Resolution* 48: 49-62.
- Mintz, Alex, and Nehemia Geva. 1993. "Why Don't Democracies Fight Each Other? An Experimental Study." *Journal of Conflict Resolution* 37: 484-503.
- Mitchell, David. 2005. *Making Foreign Policy: Presidential Management of the Decision-Making Process*. Burlington, VT: Ashgate.
- Moscovici, Serge. 1976. *Social Influence and Social Change*. Trans. by Carol Sherrard, and Greta Heinz. London: Academic Press.
- Myers, David G., and Helmut Lamm. 1976. "The Group Polarization Phenomenon." *Psychological Bulletin* 83: 602-27.
- Preston, Thomas. 2001. *The President and His Inner Circle: Leadership Style and the Advisory Process in Foreign Affairs*. New York: Columbia University Press.
- Redd, Steven B. 2002. "The Influence of Advisers on Foreign Policy Decision Making: An Experimental Study." *Journal of Conflict Resolution* 46: 335-64.

- Ross, Lee, and Andrew Ward. 1995. "Psychological Barriers to dispute Resolution." In *Advances in Experimental Social Psychology* Vol. 27, ed. M. P. Zanna. San Diego: Academic Press.
- Schafer, Mark. 1997. "Images and Policy Preferences." *Political Psychology* 18: 813-29.
- Semmel, Andrew K. 1982. "Small Group Dynamics in Foreign Policymaking." In *Biopolitics, Political Psychology, and International Politics*, ed. Gerald Hopple. New York: St. Martin's.
- Simon, Herbert A. 1982. *Model of Bounded Rationality*. Cambridge, MA: MIT Press.
- Simon, Herbert A. 1985. "Human Nature in Politics: The Dialogue of Psychology with Political Science." *American Political Science Review* 79: 293-304.
- Snyder, Richard C., H.W. Bruck, and Burton Sapin. 1954. *Decision Making as an Approach to the Study of International Politics*. Foreign Policy Analysis Project Series No. 3. Princeton, NJ: Princeton University Press.
- Snyder, Richard, H.W. Bruck, Burton Sapin, Valerie M. Hudson, Derek H. Chollet, and James M. Goldgeier. 2002. *Foreign Policy Decision-Making Revisited*. New York: Palgrave Macmillan.
- Sylvan, Donald A., and James F. Voss, eds. 1998. *Problem Representation in Foreign Policy Decision Making*. Cambridge: Cambridge University Press.
- Tomz, Michael. 2007. "Domestic Audience Costs in International Relations: An Experimental Approach." *International Organization* 61: 821-40.
- Tomz, Michael. 2009. "The Foundations of Domestic Audience Costs: Attitudes, Expectations, and Institutions." In *Expectations, Institutions, and Global Society*, eds. Masaru Kohno, and Aiji Tanaka. Tokyo: Keiso-Shobo.
- Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5: 297-323.
- Vertzberger, Yaacov. 1997. "Collective Risk Taking: The Decision-Making Group." In *Beyond Groupthink*, eds. Paul 't Hart, Eric K. Stern, and Bengt Sundelius. Ann Arbor: University of Michigan Press.
- Wood, Wendy. 2000. "Attitude Change: Persuasion and Social Influence." *Annual Review of Psychology* 51: 539-70.

IX. Advanced Experimental Methods

31. Treatment Effects

Brian J. Gaines and James H. Kuklinskiⁱ

Within the prevailing Fisher-Neyman-Rubin framework of causal inference, causal effects are defined as comparisons of potential outcomes under different treatments. In most contexts, it is impossible or impractical to observe multiple outcomes (realizations of the variable of interest) for any given unit. Given this fundamental problem of causality (Holland 1986), experimentalists approximate the hypothetical treatment effect by comparing averages of groups or, sometimes, averages of differences of matched cases. Hence, they often use $(\bar{Y}|t = 1) - (\bar{Y}|t = 0)$ to estimate $E[(Y_i|t = 1) - (Y_i|t = 0)]$, labeling the former quantity the *treatment effect* or, more accurately, the *average treatment effect*.

The rationale for substituting group averages originates in the logic of the random-assignment experiment: each unit has different potential outcomes; units are randomly assigned to one treatment or another; and, in expectation, control and treatment groups should be identically distributed. To make causal inferences in this manner requires as well that one unit's outcomes not be affected by another unit's treatment assignment. This requirement has come to be known as the *stable unit treatment value assumption*.

Until recently, experimenters have reported average treatment effects as a matter of routine. Unfortunately, this difference of averages often masks as much as it reveals. Most crucially, it ignores heterogeneity in treatment effects, whereby the treatment affects (or would affect, were it actually experienced) some units differently from others.

This chapter critically reviews how researchers measure, or fail to measure, heterogeneous treatment effects in random assignment experiments, and takes as its integrating theme that these effects deserve more attention than scholars have given them. In multiple ways, such heterogeneity, when not addressed, reduces an experiment's capacity to produce a meaningful answer to the initial research question.

The remainder of this chapter discusses four varieties of heterogeneity that can complicate and derail interpretations of estimated treatment effects.

1. Noncompliance: Calculations intended to correct for measured and unmeasured failures to comply with control-treatment assignment, especially in field experiments, imply possible variance in the treatment's impact on units, and, equally, ignorance of how the treatment would have affected untreated cases. Other, more subtle forms of noncompliance, such as non-response in survey experiments, pose related but distinct problems.
2. Pretreatment: Acknowledging real-world pre-treatment of cases admits the possibility that a random assignment experiment captures not the discrete effect of treatment, but the average marginal effect of additional treatment, conditional on an unmeasured level of real-world pretreatment. At one extreme, a universal, powerful, and enduring real-world effect can ensure that an experimental treatment has no impact, leading to an erroneous conclusion about real-world cause and effect.
3. Selection: Random assignment generates an estimate of the average treatment effect over the whole population. In the real world, individuals often self-select into or out of the treatment. Hence, the experimentally estimated average treatment effect can differ

markedly from the real-world average treatment effect, depending on whether those who selected out would have, in the event of exposure, reacted differently to the treatment from those who selected in. Both of these average treatment effects, experimental and real-world, can be interesting and important, but researchers often fail to discriminate between them, or to identify which is of primary substantive interest.

4. Social Context: Sometimes an individual's social milieu, which is determined by the choices of others, conditions a treatment's impact on him or her. Situations of this sort, in which an individual's outcome depends upon the treatment status of other individuals, encompass various institutional settings and violate the stable unit treatment value assumption. They thereby raise potentially serious complications for the random assignment experiment and observational studies alike.

For any given study, heterogeneity of one or more of these types might be an issue. Sometimes, more than one variety of heterogeneity will obtain; in other cases, none will. All of these sources of heterogeneity can confound efforts to identify an important real-life phenomenon by means of estimating an experimental treatment effect. Unfortunately, most are not directly or fully observable within the experiment itself.

Two crucial implications follow. First, an experimenter should avoid thinking only in terms of a single average treatment effect. One might ultimately choose to compute and report only an average treatment effect, but this decision should be made consciously, and the rationale stated explicitly. Second, prior to designing an experiment, a researcher should seriously consider the plausibility of heterogeneous effects. The temptation, in this regard, is to call for

good theory. While we welcome strong theory, developing thoughtful expectations about the phenomenon under study would be a welcomed first step.

1. The Importance of Distributions to the Study of Treatment Heterogeneity

Social scientists routinely construe treatment effects solely in terms of the difference in means between treatment and control conditions. Because a treatment's effect on each subject is generally unobservable, this difference of group averages does not fully summarize the information that a random assignment experiment generates. The distributions on the dependent variable(s), taken by all randomly constituted groups, can be especially useful when exploring the possibility of heterogeneous treatment effects.

Suppose, for instance, that an experimenter seeks to determine whether requiring people to state and justify their policy positions produces more moderate positions, overall, than requiring them only to state their positions, without rationale. The experimenter randomly assigns subjects to the treatment (state-and-justify) and control (state-only) conditions. He or she measures policy positions on 7-point scales and, on each item, finds no difference in means across conditions.

It would be premature to conclude that the treatment lacks effect. Suppose that in the control condition, where people are not required to justify, policy preferences are uniformly distributed. Moreover, those individuals in the treatment condition who are inclined to take relatively centrist positions in the absence of justification moderate further when asked to defend these positions. Those inclined to take relatively extreme positions, by contrast, become even more extreme when they both state and defend their positions. Members of the control group would thus exhibit approximately uniform preferences, whereas members of the treatment group

would pile up in the middle and at the end points. The means of the two groups would be identical, or nearly so, leading the researcher to miss the substantively interesting dynamics associated with the treatment.ⁱⁱ Generally, a comparison of means is inadequate if treatment produces multiple and offsetting effects (Fisher 1935).

Comparing distributions is more difficult and less familiar than comparing means (or conditional means), but equating “treatment effect” with differences in means is plainly too limiting. Through the remainder of this chapter, we compare means (or proportions), in the interest of simplicity. But looking beyond means, to distributions, is always prudent.

2. Measured and Unmeasured Noncompliance and Treatment Effects

Because field experimenters conduct their research directly in the real world, they typically lose some control over delivery of the treatment, thus failing to implement fully the assignment of units to treatment and control conditions. For this reason, they have devoted more attention to heterogeneous treatment effects than have users of other experimental types. Nevertheless, treatment-effect heterogeneity stemming from noncompliance can pervade all experiments, sometimes in not-so-obvious ways.

To begin, Let T designate treated and C designate controlled, by which we mean untreated. Using * to designate assignment, we can partition units into two mutually exclusive, exhaustive groups according to their assignment status, T* and C*. We can also partition units into another two mutually exclusive and exhaustive groups, according to their actual experience, T and C. To characterize both assignment and actual experience, we designate the four exhaustive, mutually exclusive, and not necessarily observable statuses by listing assignment status first and experience second: T*T, T*C, C*T, C*C. Units in the first and last type can be

described as *compliers*, those in the middle two types as *noncompliers*. Only the two groups T* and C* are randomly constituted, and they provide analytical leverage via their equivalent-in-expectation compositions. Full heterogeneity in parameters across the four types produces nonidentification, so scholars impose assumptions.

Political scientists have focused primarily on one instance of the problem, the observable, unintended nontreatment of some cases assigned to treatment, i.e., the T*C cases (Gerber and Green 2000; Gerber, Green, and Larimer 2008; Hansen and Bowers 2009). In a canonical mobilization experiment, for example, a researcher randomly selects a subset of individuals from a public list of registered voters and exposes them to a mobilization message, such as a postcard containing some information about the election or an argument in favor of voting. A naive estimate of the treatment effect would be the eventually observable difference in actual, real world turnout between those mobilized in this manner (the Ts) and those not (the Cs), ignoring assignments. Alternatively, one might compare the Ts to the C*s.

Recognizing that neither of those differences adjusts for the possibility of systematic differences between compliers and non-compliers, field experimenters typically divide the measured treatment effect by the proportion of assigned-to-treatment cases that accidentally went untreated by virtue of non-delivery of the cards.ⁱⁱⁱ This remedy assumes, often implicitly, that the delivery process measures an otherwise unobservable quality in voters, which might be thought of as being hard-to-reach. The companion hypothesis is that the easy-to-reach (E) and hard-to-reach (H) could differ both in their baseline probabilities of voting (b_E and b_H) and in their responses to mobilization (t_E and t_H).

The goal is to posit an explicit model with sufficient restrictions to permit estimates of interesting parameters. Assuming only two types is a start. Taking the partition of T*s to be a valid measure of type means that, by assumption, the empirical estimate of α , the proportion of the public that is type E, is the number of T*T cases divided by the number of T* cases. Further assuming no accidental treatment of those assigned to control (say, by spillover within households, as when multiple people read the postcard) means no C*T cases. Finally, random assignment creates identical expected mixtures of E and H types in the T* and C* groups. So, if Y is an indicator for turning out to vote, then $E(\bar{Y} | T^*) = \alpha(b_E + t_E) + (1 - \alpha)(b_H)$ and $E(\bar{Y} | C^*) = \alpha(b_E) + (1 - \alpha)(b_H)$. Hence, t_E can be estimated by taking the difference in turnout for the T* and C* groups and dividing by $\hat{\alpha}$, the estimate of the proportion type Es, derived from the T* group.

Note that, under these assumptions, no estimate of t_H is available, and so only one of the two postulated treatment effects is evaluated. The premises of the calculation are that there could be heterogeneity in the effects of the treatment, and that the experiment generates estimates of only some of the potentially interesting parameters. There is, just the same, a tendency among scholars to describe the estimate as “the” treatment effect. Downplaying the unmeasured hypothesized effect is perhaps natural, but it is also imprecise, if not perverse.^{iv}

Treatment is not, of course, delivery of a message to a mailbox, but, rather, exposure of the individual to that message. Accordingly, actual treatment in a field experiment on mobilization is sometimes unobservable. Some cards will be discarded as junk without having been read or processed, some cards will be read, but not by the intended individuals, and so on.

Unmeasured non-treatment of T* cases leads the researcher to overestimate α and underestimate t_E . The resulting bias, therefore, is conservative.

In a departure from the usual approaches, Hansen and Bowers (2009) use sampling theory to generate confidence intervals for treatment effects given noncompliance. In the terminology introduced above, they compute a confidence interval for the mean baseline rate, with no explicit assumptions about compliance types or a noncompliance model, and then use simple arithmetic to translate the observed treatment effect on those reached (t_E) into a confidence interval. Agnostic about types, the method emphasizes average effects, and Hansen and Bowers make a compelling case for its widespread use.

There are other approaches to the noncompliance problem. Albertson and Lawrence (2009), studying the effects of exposure to particular television programs, had to cope with multiple forms of noncompliance, both T*C cases and C*T cases (as measured by unverified self-reports of behavior). They, too, introduce a host of assumptions to simplify the problem to the point of tractability. Positing three types – (1) always-T; (2) compliers, who behave as assigned; and (3) never-T – they next assume a common treatment effect (unobserved, of course, for type-3 subjects) but differing baseline rates (in their case, probabilities of correctly answering factual questions). As a result, they obtain an estimate of the treatment effect from another ratio of a difference and a proportion, the latter estimable on the assumptions that only type-1s are C*T and only type-3s are T*C:

$$E(\bar{Y} | T^*) = \alpha_1(b_1 + t) + \alpha_2(b_2 + t) + (1 - \alpha_1 - \alpha_2)(b_3)$$

$$E(\bar{Y} | C^*) = \alpha_1(b_1 + t) + \alpha_2(b_2) + (1 - \alpha_1 - \alpha_2)(b_3)$$

$$\alpha_2 = \frac{N_{T^*T}}{N_T} - \frac{N_{C^*T}}{N_C} = (\alpha_2 + \alpha_1) - \alpha_1 = \frac{N_{C^*C}}{N_C} - \frac{N_{T^*C}}{N_T} = (\alpha_2 + \alpha_3) - \alpha_3$$

$$\therefore \hat{t} = \frac{(\bar{Y} | T^*) - (\bar{Y} | C^*)}{\hat{\alpha}_2}$$

The approach is clever, although the typology would not work with multiple iterations wherein some initial non-compliers shift to complying even with the same assignment. More to the point, to achieve identification requires a strong assumption of homogenous treatment effects across the posited types, notwithstanding their potentially disparate untreated states.

Nickerson (2005) reviews several alternative assignment strategies, including the placebo-as-control approach, which reduces the need to impose assumptions by creating multiple instances of the contact process. He later (2008) demonstrates this design in a study investigating contagion/spillover effects on the housemates of those treated with mobilizing messages. In two cities, canvassers with either a mobilization message or a placebo environmental message successfully contacted subjects between one-third and one-half of the time. By comparing turnout rates of (a) the two groups of contacted subjects and (b) the two groups of housemates, Nickerson generates estimates of direct and indirect mobilization. But in an otherwise careful presentation, he says little about the unmeasured, unknown effects treatment would have had on those not contacted. He does briefly puzzle over the unexpected lack of difference between both pre- and post-experimental turnout rates of the placebo and control (those assigned to treatment or placebo, but not successfully contacted) groups. But, having noted this surprising similarity in baseline rates, he offers no additional qualifications about possible heterogeneity in treatment effects. The experiment reveals nothing about the susceptibility of the control subjects to the political message. The study, although persuasive and precise, nevertheless exhibits the usual

inattention to the prospect of significant treatment-effect heterogeneity. Because it is highly innovative, we single it out.

Barring mistakes of implementation, laboratory and survey experiments treat all those chosen for treatment, and no one else.^v Generally, therefore, basic compliance with assignment is not an issue. However, when subjects do not understand the treatment or do not take it seriously, these experimenters face problems akin to noncompliance. Laboratory experiments also sometimes lose subjects who depart midway through a study, while surveys usually feature non-response on some items, so that the behavior of interest (the dependent variable) can be missing. Common practice is quietly to drop missing-data cases from analysis, often without even reporting the number of lost observations. In such instances, noncompliance is masked, not overcome.

Suppose that non-response is systematic, such that it wrecks the initial balance that random assignment aims to achieve. A simple fix consists of introducing covariates to account for different rates of non-response in the treatment and control conditions. If, however, non-responders differ from responders with respect to baseline rates or treatment effects, this simple fix will not necessarily suffice. In the event that the treatment itself induces non-response, that unintended effect thwarts measurement of the expected treatment effect, as captured in differences between means. Fortunately, the data sometimes contain evidence of the problem. But solutions are not easy; imputation, for instance, will generally require strong assumptions. Bounds for the effects of non-response on estimates can be computed, but they need not be narrow (Manski 1995, 22-30).

Although laboratory experimenters successfully avoid overt noncompliance most of the time, they must be sensitive to less obvious forms of noncompliance with the treatment, such as a lack of engagement. Survey researchers should view non-responses as a form of noncompliance, and tackle the phenomenon head-on. Because the various forms of noncompliance have design implications, users of all types of experiments would benefit from identifying the extent of noncompliance and making appropriate adjustments. Even when solutions are costly, addressing the problem can only improve the end product.

3. Pretreatment and Treatment Effects

Much of the time, political scientists use experiments to understand ongoing, real world causal relationships. They use survey experiments to reveal how the strategic framing of issues in live debates shapes citizens' views and preferences. They conduct voter mobilization field experiments to determine the real-world impacts of get-out-the-vote campaigns. They experimentally study the effects of negative advertisements in the midst of nasty campaigns.

How to interpret experimental treatment effects depends critically on one's assumptions about the relationship between the two contexts, real world and experimental simulation thereof. Validity concerns abound, of course. In laboratory studies of negative advertisements, for example, scholars strive to mimic the look and tone of real ads and disguise their purpose by embedding the ads in news broadcasts and (temporarily) misleading subjects about the study's goal. When feasible, they even create a home-like setting. Less discussed is how real world pretreatment might complicate the interpretation of any difference in behavior between the experimentally treated and untreated.

Researchers use random assignment to ensure that the treated will be identical to the untreated, in expectation, apart from the experimental treatment itself. Why, then, is nonexperimental pretreatment any different from any other trait that might be pertinent to the behavior of interest but should be balanced, in expectation, across groups? As long as there is some possibility that experimental subjects arrive already having been exposed to the actual treatment being simulated, the experiment estimates not the average treatment effect, but, rather, the average marginal effect of additional treatment (Gaines, Kuklinski, and Quirk 2007).

To be concrete, suppose that viewing negative advertisements decreases everyone's probability of voting such that $p(\text{vote}_i) = b_i(1 - 0.1\sqrt{k_i})$, where k is the number of ads seen in the last j days and b is a baseline, untreated probability, both varying across individuals. The estimated treatment effects from an experiment would then depend on the distributions of b and of k , and on the chance variation in these parameters across treatment and control groups. But the critical point, for present, is that experiments conducted in a relatively pristine population, unexposed to advertisements, will find much larger effects (i.e. differences between reported vote intentions of treatment and control groups) than those performed on subjects who happen to have been inundated with advertisements before they were ever recruited into the study. If everyone's baseline probability of voting is 0.7, and no one has seen any ads, the experiment will estimate a 7 percent demobilization effect from seeing one negative ad (a treatment that might imprecisely be described as being exposed to negative ads). If, instead, subjects arrive at the laboratory having seen, on average, four ads that have already affected their voting plans, the estimated demobilization effect will be less than 2 percent.

All of the foregoing is a bit fast and loose, as it rests on a fuzzy assumption of long-duration treatment effects. The experimenter might offer an ironic rebuttal: pretreatment is a negligible concern because the real-world effect will be short-term, so that subjects will enter the experiment as if they had never been treated. However, this argument implies that the researcher is studying fleeting and, hence, possibly unimportant forces. If political scientists view negative advertisements as important because they affect people's inclinations to vote, then they must surely believe that an experiment wherein they expose some subjects to an ad or two and others to none is generally susceptible to pretreatment effects, unless, perhaps, the experiment is conducted outside of election season.

At minimum, experimenters should consciously consider whether pretreatment has occurred. Whenever the answer is affirmative, they will want to enumerate the implications fully. They should ponder whether the existence of a real world pretreatment biases the experimental estimate downward. Even a simple yes or no, without any estimated magnitude of the bias, represents an improvement over failure to pose the question at all.^{vi}

4. Self-Selection as a Moderator of Treatment Effects

Despite efforts by Tobin (1958), Heckman (1976), and others, selection continues to challenge observational studies. Experimentalists, in contrast, have been able to ignore it, since random assignment, by design, entirely eliminates self-selection into or out of treatment. To achieve this, the random assignment experiment exposes some individuals to a treatment they would never receive in the real world and fails to expose some to a treatment they would receive. Whenever the experimenter's purpose is to estimate real world treatment effects that are shaped by selection processes, therefore, randomization might not be appropriate. The experimenter

begins with one question – what is the (average) real-world treatment effect? – but unwittingly answers another – what would be the (average) treatment effect if everyone were exposed?

Consider, again, the ongoing debate about the effects of exposure to negative campaign ads on voting turnout. Observational studies have generally found negative ads to have no effect, or a small positive effect (see Lau et al. 1999). In contrast, a highly visible study found that exposure to such ads decreased turnout by about five percentage points (Ansolabehere et al. 1994). The authors, who explicitly take prior observational studies as their point of departure, use an experiment to generate their primary findings; and, later, analyze aggregated observational data to confirm their experimental results (Ansolabehere, Iyengar, and Simon 1999).

Putting aside the numerous measurement problems that beset the various observational studies, the observational and experimental results should not be the same, unless everyone in the real world is exposed to campaign ads, or there is no difference in the effects of exposure to these ads between those who do and those who do not experience them in real life. The experiments conducted by Ansolabehere et al., in other words, almost certainly estimate the potential, not the actual, treatment effect.

Is there any way, then, that experimenters can estimate the treatment effects that often seem to be their primary interest, that is, effects moderated by self-selection? To find such an estimation procedure would not only increase the scope of questions an experimenter could address, it would facilitate communication between users of observational and experimental data such they could begin to address genuinely common questions.

One simple and promising approach entails incorporating self-selection into random assignment experiments. Subjects are randomly assigned to treatment, control, or self-selection conditions, and those assigned to the self-selection condition choose to receive or not receive the treatment.^{vii} The experimental simulation of the treatment must be highly valid, with the selection mechanism resembling real world selection.

Analysis of the data created by such an experiment could proceed roughly along the lines of adjusting for non-compliance, discussed earlier. Assume a dichotomous outcome variable, Y , so that the analysis focuses on the difference in proportions (i.e. means of the indicator Y) between control and treatment groups. There are only two behavioral options, and two observable types: those who opt into the treatment given the chance (type o) and those who opt out (type n).^{viii} We let α be the proportion of the former type within the population and assume that types o and n differ in both treatment effect (t) and the baseline (untreated) probability of exhibiting the behavior (b).

The group randomly assigned to control status (RC) does not get the treatment and consists of a mixture of the two types with an expected proportion of:

$$E(\bar{Y}_{RC}) = \alpha b_o + (1 - \alpha) b_n$$

Those randomly assigned to treatment (RT) also comprise a mixture (the same mixture, in expectation), but since they get the treatment, they have an expected proportion of:

$$E(\bar{Y}_{RT}) = \alpha(b_o + t_o) + (1 - \alpha)(b_n + t_n)$$

The third group, randomly assigned to the condition wherein they select whether to be treated or not (RS), has an expected proportion that is a weighted average of the baseline rate for type n 's

and the treatment-adjusted rate for type o's (provided, again, that the experimental treatment selection process perfectly simulates real-world selection):

$$E(\bar{Y}_{RS}) = \alpha(b_o + t_o) + (1 - \alpha)b_n$$

Under these assumptions, the observed rates generate estimates of both postulated treatment effects. A little arithmetic confirms that estimators for the treatment effects on selectors and non-selectors are, respectively:

$$\hat{t}_o = \frac{\bar{Y}_{RS} - \bar{Y}_{RC}}{\hat{\alpha}} \quad \text{and} \quad \hat{t}_n = \frac{\bar{Y}_{RT} - \bar{Y}_{RS}}{1 - \hat{\alpha}}$$

Just as one can accommodate unintended nontreatment in field experiments, creating a simulation of realistic self-selection allows exploitation of otherwise unobservable variance. The design above reveals everything that the random assignment experiment reveals, and more. Most crucially, it adds information useful for studying a phenomenon that seems likely to feature substantial selection effects in the real world. When trying to understand how a treatment works in a world where, first, some units are exposed to it and others are not, and, second, the exposed react differently to the treatment than the unexposed would have reacted, a self-selection module can prove helpful and, arguably, necessary.^{ix}

5. Ecological Heterogeneity in Treatment Effects

People self-select into or out of treatments. Sometimes these decisions do no more than determine whether the isolated individual will or will not receive the treatment. At other times, however, the individual's choice of treatment also determines with whom he or she will interact. If, in turn, such interactions shape the behavior of interest, then the effect of the treatment will be conditional on the social ecology within which an individual is embedded. This additional source

of potential heterogeneity, when not taken into account, will weaken the validity of a random assignment experiment. Examples abound, but for purposes of illustration, consider schools.

School voucher programs have been at the center of intense controversies for decades, with analyses purporting to show very different results on whether they help or hinder educational achievement. The literature is vast: confining attention to one arbitrary year, one can find review essays skeptical that vouchers work (e.g. Ladd 2002), analyses of randomized field experiments reporting strong positive effects (e.g. Angrist et al. 2002; Howell et al. 2002), analyses reporting mixed results (Caucutt 2002), and essays complaining that all studies have been too narrowly focused (Levin 2002).

Several studies of vouchers have exploited a *natural experiment* wherein an actual program implements random selection for choosing awardees. By way of illustrative example, suppose that:

- a school voucher program has more applicants than spots;
- available vouchers are thus given to a randomly chosen subset of the applicants;
- all voucher recipients enroll in private schools;
- all unsuccessful voucher applicants remain in public schools;
- analysts compare subsequent performance by the voucher recipients to subsequent performance by the unsuccessful voucher applicants, to make inferences about the effect of being educated in a private school.

By studying applicants only, the researcher controls for a range of unobservable factors that are likely associated with parental ambition and the child's performance. Studying applicants only also restricts the domain about which one can draw inferences, and thus changes the research question. Still, the use of randomization by the program administrators is felicitous, ensuring that accepted and rejected pupils will be similar in motivation and potential – identical, in

expectation, other than their differences induced by the treatment (attendance at private school, or not). Can this seemingly fool-proof design go wrong?

For simplicity, suppose further that there are only two kinds of pupils, those with low potential (because of cognitive limitations, poor work ethic, lack of family support, etc.) and those with high potential. There are also only two schools, one public and one private, and they are not necessarily the same size. Achievement is measured by a test that is accurate except for random error, and individual students' measured achievements are generated as follows:

$$Y = \beta_0 + \beta_1 H + \beta_2 P + \beta_3 (H \times P) + \beta_4 Z_H + \varepsilon$$

where H and P are indicators for high potential pupils and pupils attending private schools, respectively, Z_H is the proportion of students in the given school who are high potential types, and ε is a standard stochastic disturbance term.

Assume that potential is impossible to measure, and thus acts as a lurking variable.^x Of primary interest is the accuracy of the estimated effect of attending private school generated by the (quasi-)experimental analysis described above, which limits attention to voucher applicants and exploits random assignment. For contrast, we will also consider a naïve estimator, the public-private mean difference.

In the absence of interaction effects ($\beta_3 = 0$ and $\beta_4 = 0$), test scores would be randomly distributed around means as follows: $E(Y|H=P=0) = \beta_0$; $E(Y|H=0, P=1) = \beta_0 + \beta_2$; $E(Y|H=1, P=0) = \beta_0 + \beta_1$; and $E(Y|H=P=1) = \beta_0 + \beta_1 + \beta_2$. Conclusions depend on the signs of the betas. If, for instance, β_2 is negative (private schools perform more poorly than public schools, for pupils of both types) while β_1 is positive, yet high potential students are disproportionately found in the private school, we could obtain an instance of Simpson's paradox, wherein the public school's

overall mean achievement score is lower than the private school's mean score, even though students of both types perform better in the public school. It is probably more plausible that both β_1 and β_2 are positive, and that the ratio of high potential pupils to low potential pupils is higher in the private school than in the public school. Letting λ and δ represent the proportions of private and public school pupils of high potential, respectively, we assume that the private school draws a disproportionate share of high-potential pupils, so that $\lambda - \delta > 0$.

A naive observational study that compares public to private mean achievements would have an expected treatment effect (benefit obtained by attending private school) of:

$$(1-\lambda)(\beta_0 + \beta_2) + \lambda(\beta_0 + \beta_1 + \beta_2) - (1-\delta)\beta_0 - \delta(\beta_0 + \beta_1) = \beta_2 + (\lambda - \delta)\beta_1$$

The estimate is biased upwards, since both means are weighted averages of high and low potential pupils' scores, but the private school puts more weight on the comparatively more numerous high potential students, thereby mixing some β_1 into what was meant to be an estimate of β_2 . A successful selection model might reduce this bias, if there are good (measured) predictors of the unmeasured pupil potential.

For the experimental analysis, the difference between the treatment group (successful voucher applicants, placed in private schools) and control group (unsuccessful applicants, remaining in public schools) does not depend on the mix of low and high potential students in the applicant pool. Random selection ensures, in expectation, recreation of the applicant pool's ratio of high to low potential pupils in both treatment and control. Suppose that α is the proportion of the applicants with high potential. The difference between treatment and control will be $(1-\alpha)(\beta_0 + \beta_2 - \beta_0) + \alpha(\beta_0 + \beta_1 + \beta_2 - (\beta_0 + \beta_1)) = \beta_2$. Hence, random assignment will have

served its purpose and solved the problem of a lurking variable, which caused bias in the observational estimate.

If $\beta_3 > 0$, then test scores are randomly distributed around those same means, except that $E(Y|H=P=1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$. An observational study subtracting the public school mean from the private school mean would again obtain an upwardly biased estimate of β_2 , the private school bonus, namely $\beta_2 + (\lambda - \delta)\beta_1 + \lambda\beta_3$. There are now, by assumption, two distinct private school effects, but this estimate need not lie between β_2 , the true effect for the low potential pupils, and $\beta_2 + \beta_3$, the true effect for the high potential pupils.

The share of the voucher applicants that were high-potential is still α . In the experimental context, the difference between treatment and control is thus $\beta_2 + \alpha\beta_3$. For any α between 0 and 1 (non-inclusive), this value over-estimates the achievement effect of attending private school for pupils with low potential, but underestimates the effect for pupils with high potential. In the extreme case wherein only high-potential pupils apply for vouchers ($\alpha=1$), the treatment-control difference is an accurate estimate of the returns on private schooling for high-potential pupils. Like the observational study, the experimental analysis would be susceptible to the mistake of inferring that all pupils could have their scores boosted by that same amount by being educated in private schools. However, the estimate is usefully bounded by the two real world effects.

Most interesting is the situation wherein there are both pupil-school interaction effects ($\beta_3 > 0$) and pupil-pupil effects ($\beta_4 > 0$). Now achievements are more complicated, depending not only on pupil and school types, but also on the concentration of high-potential pupils within each school. Again, estimating a single private school effect is misleading, since private schooling delivers distinct benefits to pupils with low or high potential. Moreover, all pupils gain from

contact with high potential peers, and this effect can be felt in the private or public school, depending on who applies for vouchers and on the outcome of the lottery after the voucher scheme is oversubscribed. The estimated treatment effect for a quasi-experimental analysis that exploits the randomization will depend on the actual realization of the random selections. The treatment-control difference will contain not only the slight bias noted above, but also a term representing the differences in the schools' distributions of low and high potential pupils.

In short, there are reasons to think that a given pupil's performance might sometimes be affected by the assignment status of other pupils. This represents a violation of Rubin's stable-unit-treatment-value-assumption (1990a, 1990b), in that individual i 's potential outcome, given a school choice, depends on unit j 's assignment status. Because the public and private school populations were not formed entirely by random assignment, the analyst of the natural experiment created by the voucher lottery inherits distinct school ecologies (mixtures of student) that can plausibly condition the treatment effect (the impact on performance of attending private school).

It is not obvious whether a researcher should include, in an estimate of "private schooling effects," these differences in performance attributable to the schools' different ecologies. He or she might or might not wish to credit private schools with recruiting more capable student bodies, when making the comparison of types. However, that move can lead to false inferences about the likely effects of broadening a voucher program. In any case, there is no guarantee that the analyst who seeks to exploit randomization will get a correct or unbiased estimate of private school benefits if there are ecological effects that cannot easily be measured directly.

Actual analyses of vouchers that employ this design often introduce more complicated models that include sensitivity to various pupil traits (e.g. Rouse 1998), so we are not advancing a complaint about the voucher literature in particular. We aim merely to reinforce the point that random assignment is no panacea. When the treatment to which subjects are randomly assigned has complicated effects that interact in some manner with other unobserved (or even unobservable) factors, causal inferences can be biased, not only in observational studies but also in studies that aim to exploit (partial) randomization of treatment assignment. Scholars ignore the issue of treatment effect heterogeneity at their own peril.

6. Beyond Internal and External Validity

The two criteria that scholars have traditionally applied when assessing the random assignment experiment are internal and external validity. Conventional wisdom says that it typically excels at achieving internal validity, while falling short at achieving external validity. How, if at all, do the four varieties of treatment heterogeneity that we have discussed challenge this conventional wisdom?

Most fundamentally, the preceding discussion underlines that assessing the random assignment experiment only in terms of the two types of validity is too limiting. Instead, we propose, when researchers use experiments to make inferences about a world that exists outside of the experimental context, they should think in terms of a complex nexus consisting of research question, real-world context, and experimental context. Often, the research question arises because the experimenter has observed an event or phenomenon about which he or she would seek to make causal inferences. That very observation implies that subjects might enter the experiment already having been pretreated; or that some subjects received the pretreatment while

others did not; or that social interactions generated the event or phenomenon. In turn, each of these possibilities should signal the importance of revisiting the initial research question. Should the study really focus on marginal rather than average treatment effects? Does the researcher really want to know the causal effect when everyone receives the treatment, or, given what he or she has already observed, is the treatment effect amongst only those who likely receive it when carrying on their day-to-day lives of more interest? Should the researcher's question acknowledge the existence of social interactions and their likely effects on the dependent variable of interest? Similarly, given hunches about processes and phenomena outside the experimental context, exactly how should the researcher design the experiment?

Implicit in the preceding discussion is what might be called a “paradox of social-scientific experimental research”: to know how to design an experiment properly for purposes of making inferences about a larger world requires knowledge of how the processes and phenomena of interest work in the real world. Without that knowledge, the researcher cannot know how to design the experiment, or how to interpret the results. In strong form, this paradox implies that conducting experimental research for purposes of making causal inferences about the real world is impossible. We recommend, however, that social scientists not accept the paradox, but use it instead as a reminder that experimental research requires considerable thought and reflection, much more than we all once supposed.

7. Conclusion

By way of conclusion, and to underline various observations we have made throughout this chapter, assume that a team of researchers has been given the opportunity to use a random national sample of individuals in their experiments. The researchers accept the offer, thinking

that the sample will enable them to make more encompassing statements than they could have made otherwise. In other words, they believe they have improved their claim to proper causal inference. They randomly assign their respondents to conditions, and measure average treatment effects.^{xi}

This all sounds ideal, and in the absence of heterogeneous treatment effects, it might be. As the experimental study of politics continues to mature, however, we expect political scientists to find the usually implicit assumption of no heterogeneity less and less tenable. Consequently, they will increasingly question the value of random samples, strictly random assignment, and average treatment effect estimates. What meaning accrues to an experimentally generated average treatment effect estimated on the basis of a national sample of citizens? Does a random assignment experiment suffice when self-selection pervades life? Can experimentalists ignore social interactions when those interactions at least partially determine the magnitude of the treatment effect? These sorts of questions, and the answers that political scientists give, will motivate and shape the next generation of experiments.

The strong possibility of heterogeneous treatment effects will increase the importance of knowing the experimental context. It will also nudge researchers to use experiments more strategically than they have to date. Creative designs, such as Nickerson's two-person-household study, broaden investigation of effects beyond the realm of the isolated individual, in the spirit of the Columbia studies. It is probably true, more generally, that a deliberate focus on experimenting within multiple homogenous populations can be very helpful.

Consider, again, the effects of negative campaign ads on voting turnout. Researchers could deliberately select pools of subjects from distinct environments, chosen according to

expectations about important forms of real-world pre-treatment. They might, for example, implement parallel survey or laboratory experiments in states or districts with and without ongoing negative campaigns, or with various levels of campaign negativity, to explore how real world context affects experimental results. Fortunately, the emergence of internet delivery as a means of conducting survey experiments has already increased the viability of such studies.^{xii} In many respects, we are elaborating a viewpoint that has a kinship with those espoused by Bowers's and Druckman and Kam's chapters in this volume. The foremost goal of the experimenter must be to get the cause and effect story right, and that goal poses more challenges than simple experimental logic suggests. It requires focused attention on the question-context-experiment nexus. Heterogeneity in treatment effects is not, ultimately, a nuisance, but an important hypothesis about the world, to be tested carefully.

References

- Albertson, Bethany, and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 32: 275-300.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-55.
- Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92: 1535-58.
- Ansolahehere, Stephen D., Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88: 829-38.
- Ansolahehere, Stephen D., Shanto Iyengar, and Adam Simon. 1999. "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout." *American Political Science Review* 93: 901-9.

- Braumoeller, Bear G. 2006. "Explaining Variance; Or, Stuck in a Moment We Can't Get Out Of." *Political Analysis* 14: 268-90.
- Caucutt, Elizabeth M. 2002. "Educational Vouchers When There Are Peer Group Effects—Size Matters." *International Economic Review* 43: 195-222.
- Fisher, Ronald Aylmer, Sir. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15: 1-20.
- Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Personal Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94: 653-64.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Vote Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102: 33-48.
- Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to a Cluster-Randomized Get-Out-the-Vote Campaign." *Journal of the American Statistical Association* 104: 873-85
- Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5: 475-92.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-60.
- Hovland, Carl I. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change." *American Psychologist* 14: 8-17.
- Howell, William G., Patrick J. Wolf, David E. Campbell, and Paul E. Peterson. 2002. "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management* 21: 191-217.
- Ladd, Helen F. 2002. "School Vouchers: A Critical View." *The Journal of Economic Perspectives* 16: 3-24.
- Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Assessment." *American Political Science Review* 93: 851-76.

- Levin, Henry M. 2002. "A Comprehensive Framework for Evaluating Educational Vouchers." *Educational Evaluation and Policy Analysis* 24: 159-74.
- Little, Roderick J., Qi Long, and Xihong Lin. 2008. "Comment [on Shadish et al.]." *Journal of the American Statistical Association* 13: 1344-50.
- Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
- Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233-52.
- Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49-76.
- Orbell, John, and Robyn M. Dawes. 1991. "A 'Cognitive Miser' Theory of Cooperators' Advantage." *American Political Science Review* 85: 515-28.
- Rouse, Cecilia Elena. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 113: 553-602.
- Rubin, Donald B. 1990a. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5: 472-80.
- Rubin, Donald B. 1990b. "Formal Modes of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25: 279-92.
- Shadish, William R., M. H. Clark, and Peter M. Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association* 103: 1334-43.
- Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26: 24-36.
- Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17: 143-61.

ⁱ Many individuals have provided invaluable advice since we first began working on this topic. Jake Bowers, Jamie Druckman, Jude Hays, Tom Rudolph, Jasjeet Sekhon, and Paul Sniderman all offered thoughtful comments on related papers, which have shaped this chapter and the project more generally. We are especially indebted to Don Green, whose advice and encouragement at several steps along the way were indispensable. We presented papers related to this chapter at the American Political Science Association and Midwest Political Science Association meetings, as well as at Indiana University, Loyola University (Psychology Department), Florida State University, and the University of Illinois.

ⁱⁱ A careful researcher might note differences in variances, and proceed accordingly (Braumoeller 2006).

ⁱⁱⁱ Equivalently, one can estimate a two-stage OLS linear probability model of voting, with an assignment-to-treatment indicator as an instrument for actually being treated (Angrist, Imbens, and Rubin 1996).

^{iv} The baseline rates, b_E and b_H , are jointly determined, so one could report bounds, given the estimate of t_E . Most analyses, however, ignore them.

^v We assume that experimental treatments successfully simulate the phenomena of interest. Myriad well-known issues concerning validity and strength of treatment arise in experiments, but these are not our primary concern here.

^{vi} Transue, Lee, and Aldrich (2009) demonstrate spillover effects occurring across survey experiments within a single survey. Sniderman's chapter in this volume offers a thoughtful response to the pre-treatment argument.

^{vii} Hovland (1959) first proposed this approach. Since we first wrote on the topic of incorporating self-selection into random assignment experiments, a related article has appeared (Shadish, Clark, and Steiner 2008). It proposes a vaguely similar design, although for a different purpose: to explore whether non-randomized experiments can generate data that, suitably analyzed, replicate results from their randomized counterparts. Unlike us, the authors regard the results of the random assignment experiment as a gold standard, always preferable, not always feasible. In our view, random assignment can deliver fool's gold if the aim is to characterize the real-world phenomenon under study. Commenting on the Shadish et al. article, Little, Long, and Lin (2008) come closer to our view when they argue that "the real scientific value of hybrid designs is in assessing not the effects of selection bias, but rather their potential to combine the advantages of randomization and choice to yield insights about the role of treatment preference in a naturalistic setting" (1344).

^{viii} Heterogeneity of types need not match observable behavior perfectly, of course. If, say, a treatment has distinct effects on type-u people, who usually opt to be treated (but not always), and type-r people, who rarely select treatment (but not never), the analysis proceeds similarly, but with more distracting algebra. Assuming homogeneity within observable groups is merely an analytical convenience.

^{ix} In a very different context, Orbell and Dawes (1991) alter a prior experimental design to permit players of Prisoners' Dilemmas to self-select continuation of play, and thereby uncover theoretically important and otherwise invisible heterogeneity.

^x Inability to observe pupil potential is a strong assumption, employed to simplify the examples.

^{xi} With the exception of some survey experiments, national samples rarely exist in practice. If given their choice, we think, most political scientists would use national random samples to conduct their experiments. In this regard, our discussion addresses current thinking, not necessarily current practice.

^{xii} Taking a cue from field experiments, these researchers should also study not only self-reported vote intent, but also actual turnout, eventually recorded in official records. Doing so should help discriminate between short-lived and long-lived effects of various treatments. Again, felicitously, internet panels make validation easier and less costly.

32. Making Effects Manifest in Randomized Experiments

Jake Bowers¹

Experimentalists desire precise estimates of treatment effects and nearly always care about how treatment effects may differ across subgroups. After data collection, concern may focus on random imbalance between treatment groups on substantively important variables. Pursuit of these three goals — enhanced precision, understanding treatment effect heterogeneity, and imbalance adjustment — requires background information about experimental units. For example, one may group similar observations on the basis of such variables and then assign treatment within those blocks. Use of covariates after data have been collected raises extra concerns and requires special justification. For example standard regression tables only approximate the statistical inference that experimentalists desire. The standard linear model may also mislead via extrapolation. After providing some general background about how covariates may, in principle, enable pursuit of precision and statistical adjustment, this paper presents two alternative approaches to covariance adjustment: one using modern matching techniques and another using the linear model — both use randomization as the basis for statistical inference.

1 What is a manifest effect?

A manifest effect is one we can distinguish from zero. Of course, we cannot talk formally about the effects of an experimental treatment as manifest without referring to probability: a scientist asks, “Could this result have occurred merely through chance?” or “If the true effect were zero, what is the chance that we’d observe an effect as large as this?” More formally, for a frequentist, saying a treatment effect is manifest is saying that the statistic we observe casts a great deal of doubt on a hypothesis of no effects. We are most likely to say that some observed effect casts doubt on the null hypothesis of no effect when we have a large sample and/or when noise in the outcome that might otherwise drown out the signal in our study has been well controlled. Fisher reminds us that while randomization alone is sufficient for a valid test of the null hypothesis of no effect, specific features of a given design allow equally valid tests to differ in their ability to make a treatment effect manifest:

With respect to the refinements of technique [uses of covariates in the planning of an experiment], we have seen above that these contribute nothing to the validity of the experiment, and of the test of significance by which we determine its result. They may, however, be important, and even essential, in permitting the phenomenon under test to manifest itself. (Fisher 1935, 24).

Of course, one would prefer a narrow confidence interval to a wide confidence interval even if both excluded the hypothesis of no effects. As a general rule, more information yields more precision of estimation. One may increase information in a design by gathering more observations and/or gathering more data about each observation. This paper considers covariates as a refinement of technique to make treatment effects manifest in randomized studies. I focus first on simple uses of

covariates in design, and then offer some ideas about their use in post-treatment adjustment.

What is a covariate? How should we use them in experiments?

A covariate is a piece of background information about an experimental unit — a variable unchanged and unchangeable by the experimental manipulation. Such variables might record the groups across which treatment effects ought to differ according to the theory motivating and addressed by the experimental design (say, men and women ought to react differently to the treatment), or might provide information about the outcomes of the experiment (say, men and women might be expected to have somewhat different outcomes even if reactions to treatment are expected to be the same across both groups). Covariates may be used profitably in experiments either *before* treatment is assigned (by creating subgroups of units within which treatment will be randomized during the recruitment and sample design of a study) and/or *after* treatment has been assigned and administered and outcomes measured (by creating subgroups within which outcomes ought to be homogeneous or adjusting for covariates using linear models).

Common Uses (and Potential Abuses) of Covariates in the Workflow of Randomized Experimentation

Every textbook on the design of experiments is, in essence, a book about the use of covariates in the design and analysis of experiments. This chapter ought not, and cannot, substitute for such sources. For the newcomer to experiments, I here summarize in broad strokes, and with minimal citations, the uses to which covariates may be put in the design and analysis of randomized experiments. After this summary, I offer a perspective on the use of covariates in randomized experimentation which, in fundamental ways, is the same as that found in books such as: Fisher (1935, 1925), Cox (1958), Cochran and Cox (1957), and Cox and Reid (2000). I differ from those previous scholars in hewing more closely and explicitly to: 1) the now well-known potential outcomes framework for causal inference (Neyman 1990; Rubin 1974, 1990; Brady 2008; Sekhon 2008) and 2) randomization as the basis for statistical inference.

Covariates allow precision enhancement

Blocking on background variables before treatment assignment allows the experimenter to create sub-experiments within which the units are particularly similar in their outcomes; adjustment using covariates after data have been collected may also reduce non-treatment related variation in outcomes. In both cases, covariates can reduce noise that might otherwise obscure the effects of the treatment.

Of course, such precision enhancements arrive with some costs: Implementing a blocking plan may be difficult if background information on experimental units is not available before recruitment/arrival at the lab (but see Nickerson 2005); care must be taken to reflect the blocking in the estimation of treatment effects to avoid bias and to take advantage of the precision enhancements offered by the design; and, in principle, analysts can mislead themselves by performing many differently adjusted hypothesis tests until they reject the null of no effects even when the treatment has no effect.

Covariates enable subgroup analyses

When theory implies differences in treatment effects across subgroups, subgroup membership must be recorded and, if at all possible, the experiment ought to be designed to enhance the ability of the analyst to distinguish group-differences in treatment effects. Covariates on subgroups may also be quite useful for post-hoc exploratory analyses designed not to cast doubt on common knowledge but to suggest further avenues for theory.

Covariates allow adjustments for random imbalance

All experiments may display random imbalance. Such baseline differences can arise even if the randomization itself is not suspect: recall that one out of twenty unbiased hypothesis tests will reject the null of no difference at the predetermined error rate of $\alpha = .05$ merely due to chance. An omnibus balance assessment such as that proposed by Hansen and Bowers (2008) is immune from this problem, but any unbiased one-by-one balance assessment will show imbalance in $100\alpha\%$ of the covariates tested. Thus, I call this problem “random imbalance” to emphasize that the imbalance could easily be due to chance and need not cast doubt on the randomization or administration of a study (although discovery of extensive imbalance might suggest scrutiny of the randomization and administration is warranted).

Random imbalance in a well-randomized study on substantively important covariates still may confuse the reader. In the presence of random imbalance, comparisons of treated to controls will contain *both* the effect of the treatment *and* the differences due to the random imbalance. One may attempt to remove the effects of such covariates from the treatment effect by some form of adjustment. For example, one may use the linear regression model as a way to adjust for covariates or one may simply group together observations on the basis of the imbalanced covariate. Adjustment on one or a group of observed covariates may, however, produce now-non-random imbalance on unobserved covariates. And, adjustment raises concerns that estimates of treatment effects may come to depend more on the details of the adjustment method rather than on the randomization and design of the study. Thus, the quandary: either risk known confusions of comparisons or risk unknown confounding and bear the burden of defending and assessing an adjustment method. An obvious strategy to counter concerns about cherry-picking results or modeling artifacts is to present both adjusted and unadjusted results and to specify adjustment strategies before randomization occurs.

Randomization is the Primary Basis for Statistical Inference in Experiments

A discussion of manifest effects is also a discussion of statistical inference: statistical tests quantify doubt against hypotheses and a manifest effect is evidence which casts great doubt on the null of no effects. On what basis can we justify statistical tests for experiments?

In national surveys we draw random samples. Statistical theory tells us that the mean in the sample is an unbiased estimator of the mean in the population as long as we correctly account for the process by which we drew the sample in our estimation. That is, in a national survey, often (but not always) our *target of inference* is the population from which the sample was drawn and we are *justified* in so inferring by the sampling design.

In other studies we may not know how a sample was drawn (either we have no well-defined population or no knowledge of the sampling process or both). But we may know how our observed outcome was generated: say we know that at the micro-level our outcome was created from discrete events occurring independently of each other in time. In that case, we would be justified in claiming that our population was created via a Poisson process: in essence we have a population generating machine, data generating process, or model of outcomes as the target of our inference.

Now, what about randomized experiments? Although we might want to infer to a model, or even to a population, the strength of experiments is inference to a counter-factual. The primary targets of inference in a randomized experiment are the experimental treatment groups: we infer from one to another. Randomization makes this inference meaningful. But, randomization also can justify the statistical inference as well: the mean in the treatment group is a good estimator for what we would expect to observe if all of the experimental units were treated: the treatment group in a randomized study is a random sample from the finite “population” of the experimental pool.²

All of the standard textbooks note this fact, but they also point out that estimating causal effects using randomization-based theory can be mathematically inconvenient or computationally intensive and that, thus, using the large-sample sampling theory (and/or Normal distribution models) turns out to provide very good approximations to the randomization-based results most of the time. Since the 1930s, the computational apparatus of randomization-based inference has expanded, as has its theoretical basis and applied reach. In this paper, all of the statistical inference I present will be randomization-based even if most of it also uses large-sample theory: for example, it takes no more time to execute a randomization-based test of the null hypothesis of no effect using mean differences than it does using the linear regression model-based approximation.³

Recently Freedman (2008*c,b,a*) reminded us that the target of inference in randomized experiments was the counterfactual and he noted that linear regression and logistic regression were not theoretically justified on this basis. Green (2009) and Schochet (2009) reminded us that often linear regression can be an excellent approximation to the randomization-based difference of means. The need for this exchange is obvious, even if it is echoed in the early textbooks: those early authors moved very quickly to the technicalities of the approximations rather than dwell on the then uncomputable but theoretically justifiable procedures. As experimentation explodes as a methodology in political science, we are seeing many more small experiments, designs where randomization is merely a (possibly weak) instrument, and theories implying very heterogeneous treatment effects. I expect we will find more of these along with many more large studies with fairly Normal looking outcomes where treatment plausibly just shifts the Normal curve of the treated away from the Normal curve of the controls. Rather than hope that the linear-model approximation works well, this paper presents analyses which do not require that approximation and thereby offers others the ability to check the approximation.

2 Strategies for enhancing Precision Before Assignment

... we consider some ways of reducing the effect of uncontrolled variations on the error of the treatment comparisons. The general idea is the common sense one of grouping the units into sets, all the units in a set being as alike as possible, and then assigning the treatments so that each occurs once in each set. All comparisons are then made within

sets of similar units. The success of the method in reducing error depends on using general knowledge of the experimental material to make an appropriate grouping of the units into sets (Cox 1958, 23).

We have long known that covariates enhance the precision of estimation to the extent that they predict outcomes. This section aims to make this intuition more concrete in the context of a randomized experiment.

An example by simulation

Imagine that we desire to calculate a difference of means. In this instance we are using a fixed covariate x and the purpose of this difference in means is to execute a placebo test or a balance test not to assess the causal effects of a treatment. Imagine two scenarios, one in which a binary treatment $Z_{ib} = 1$ is assigned to $m_b = 1$ subject within each of B pairs $b = 1, \dots, B$; $i = 1, \dots, n_b$, $B \leq n$; $n = \sum_{b=1}^B n_b$ (for pairs $n = 2B$ and thus $n_b = 2$), and another in which a binary treatment $Z_i = 1$ is assigned to $m = n - m = (n/2)$, subjects $i = 1, \dots, n$ without any blocking. Consider the test statistics

$$\begin{aligned} d_{\text{pairs}} &= \frac{1}{B} \sum_{b=1}^B \left(\sum_{i=1}^{n_b} Z_{ib} x_{ib} / m_b - \sum_{i=1}^{n_b} (1 - Z_{ib}) x_{ib} / (n_b - m_b) \right) \\ &= \frac{1}{B} \sum_{b=1}^B \left(\sum_{i=1}^2 (Z_{ib} x_{ib} - (1 - Z_{ib}) x_{ib}) \right) \end{aligned} \quad (1)$$

which reduces to the difference in means of x between treated and control units within pairs summed across pairs and

$$d_{\text{no pairs}} = \sum_{i=1}^n Z_i x_i / m - \sum_{i=1}^n (1 - Z_i) x_i / (n - m) \quad (2)$$

which sums across all units within control and treatment conditions. These two quantities are the same even if one is written as an average over B pairs: because pairs are blocks of equal size and therefore each block-specific quantity ought to contribute equally to the sum.

The theory of simple random sampling without replacement suggests that the variances of these statistics differ. First,

$$\begin{aligned} \text{Var}(d_{\text{no pairs}}) &= \frac{n}{m(n-m)} \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1} \\ &= \left(\frac{4}{n}\right) \left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned} \quad (3)$$

Second,

$$\begin{aligned} \text{Var}(d_{\text{pairs}}) &= \left(\frac{1}{B}\right)^2 \sum_{b=1}^B \frac{n_b}{m_b(n_b - m_b)} \sum_{i=1}^{n_b} \frac{(x_{ib} - \bar{x}_b)^2}{n_b - 1} \\ &= \left(\frac{2}{B^2}\right) \sum_{b=1}^B \sum_{i=1}^2 (x_{ib} - \bar{x}_b)^2. \end{aligned} \tag{4}$$

If pairs were created on the basis of similarity on x , then $\text{Var}(d_{\text{no pairs}}) > \text{Var}(d_{\text{pairs}})$ because $\sum_{i=1}^n (x_i - \bar{x})^2 > \sum_{b=1}^B \sum_{i=1}^2 (x_{ib} - \bar{x}_b)^2$. Any given x_i will be farther from the overall mean (\bar{x}) than it would be from the mean of its pair (\bar{x}_b). Note also that the constants multiplying the sums are $(4/n(n-1))$ in the unpaired case and $8/n^2$ (since $B = (n/2)$) in the paired case. As long as $n > 2$, $(4/(n^2 - n)) < (8/n^2)$ and this difference diminishes as n increases. Thus, benefits of pairing can diminish as the size of the experiment increases as long as within-pair homogeneity does not depend on sample size.

To dramatize the benefits possible from blocking, I include a simple simulation study based roughly on real data from a field experiment of newspaper advertisements and turnout in US cities (Panagopoulos 2006). The cities in this study differed in baseline turnout (with baseline turnout ranging from roughly 10 percent to roughly 50 percent). Panagopoulos paired the cities before randomly assigning treatment, with baseline turnout differences within pair ranging from one to seven percentage points in absolute value. The simulation presented here takes his original eight-city dataset and makes two fake versions, one with thirty two cities and another with 160 cities. The expanded datasets were created by adding small amounts of uniform random noise to copies of the original dataset. These new versions of the original dataset maintain the same general relationships between treatment and outcomes and covariates and pairs as the original, but allow us to examine the effects of increasing sample size. In this simulation the covariate values are created so that the average difference of means within pair is zero. For each of the 1000 iterations of the simulation and for each dataset ($n = 8, 32, 160$), the procedure was:

Create fake data based on original data Each pair receives a random draw from a Normal distribution with mean equal to the baseline outcome of its control group and standard deviation equal to the standard deviation of the pair: For the original dataset the within pair differences on baseline turnout of 1, 4, and 7 translate into standard deviations of roughly .7, 3, and 5. The “true” difference is required to be zero within every pair, but each pair may have a different baseline level of turnout and a different amount of variation — mirroring the actual field experiment.

Estimate variance of treatment effects under null Apply equations (3) and (4).⁴

[Figure 32-1 about here.]

Figure 32-1 shows that estimates of $\sqrt{\text{Var}(d_{\text{no pairs}})}$ were always larger than the estimates for $\sqrt{\text{Var}(d_{\text{pairs}})}$ although the ratio $\sqrt{\text{Var}(d_{\text{no pairs}})}/\sqrt{\text{Var}(d_{\text{pairs}})}$ diminished as the size of the experiment increased. null standard deviation across the simulations for the paired eight city example

(1.8) is close to the average standard deviation for the non-paired tests in the 160 city example (2.3). That is, these simulations show that a paired experiment of eight units could be more precise than an unpaired experiment of 160 units (although of course a paired experiment of 160 units would be yet more precise (the average null sd in that case is .5)). Notice that, in this case, the advantages of the paired design diminish as the size of the experiment increases but do not disappear.

Pairing provides the largest opportunity for enhancement of precision in the comparison of two treatments — after all, it is hard to imagine a set more homogeneous than two subjects nearly identical on baseline covariates (Imai, King, and Nall 2009). And, even if it is possible for an unpaired design to produce differences of means with lower variance than a paired design, it is improbable given common political science measuring instruments.

What do we do once the experiment has run? One can enhance precision using covariates either via post-stratification (grouping units with similar values of covariates) or covariance adjustment (fitting linear models with covariates predicting outcomes). In the first case analyses proceeds as if pre-stratified: estimates of treatment effects are made conditional on the chosen grouping. In the second case linear models “adjust for” covariates — increased precision can result from their inclusion in linear models under the same logic as that taught in any introduction to statistics classes.

3 Balance Assessment: Graphs and Tests

We expect that the bias reduction operating characteristics of random assignment would make the baseline outcomes in the control group comparable to the baseline outcomes in the treatment group. If the distribution of a covariate is similar between treated and control groups we say that this covariate is “balanced”, or that the experiment is balanced with respect to that covariate. Yet, it is always possible that any given randomization might make one or more observed (or unobserved) covariates imbalanced merely by chance. If the imbalanced covariates seem particularly relevant to substantive interpretations of the outcome (as would be the case with outcomes measured before the treatment was assigned), we would not want such differences to confuse treatment-versus-control differences of post-treatment outcomes.

One graphical mode of assessment

Figure 32-2 provides both some reassurance and some worry about balance on the baseline outcome in the thirty two-city data example. The distributions of baseline turnout are nearly the same (by eye) in groups two and three, but differ (again, by eye) in groups one and four. Should these differences cause concern?

[Figure 32-2 about here.]

[Table 32-1 about here.]

The top row of Table 32-1 provides one answer to the question about worries about random imbalance on baseline outcome. We would not be very surprised to see a mean-difference of $d_{\text{blocks}} = -1.2$ if there were no real difference given this design ($p = .2$). Of course, questions about balance of a study are not answerable by looking at only one variable. Table 32-1 thus shows a set of randomization-based balance tests to assess the null of no difference in means between

treated and control groups on covariates one-by-one and also all together using an omnibus balance test. We easily reject the null of balance on the linear combination of these variables in this case ($p=0.00003$), although whether we worry about the evident differences on percent black or median income of the cities may depend somewhat on the extent to which we fear that such factors matter for our outcome — or whether these observed imbalances suggest more substantively worrisome imbalances in variables that we do not observe. Does this mean the experiment is broken? Not necessarily.⁵

Understanding p -values in balance tests.

Say we didn't observe thirty-two observations in sets of four but only eight observations in pairs. What would our balance tests report then? Table 32-2 shows the results for such tests, analogous to those shown in Table 32-1.

[Table 32-2 about here.]

Notice that our p -values now quantify less doubt about the null. Is there something wrong with randomization-based tests, if by reducing the size of our experiment we would change our judgement about the operation of the randomization procedure? The answer here is no.⁶

The p -values reported from balance tests used here summarize the extent to which random imbalance is worrisome. With a sample size of eight, the confidence interval for our treatment effect will be large — taking the pairing into account, a 95% interval will be on the order of ± 3.5 (as roughly calculated on the baseline outcomes in Section 2). For example, both Tables 32-2 and 32-1 show the block-adjusted mean-difference in percent black between treated and control groups to be roughly 14.5 percentage points. In the thirty-two-city example, this difference cast great doubt against the null of balance, while in the eight-city example this difference casts less doubt. Now, evaluating the difference between controls and treatment on actual outcome in the eight-city case gives $d_{\text{pairs}} = 1.5$ and under the null of no effects gives $\sqrt{\text{Var}(d_{\text{pairs}})} = 2.6$ — and inverting this test leads to an approximate 88% confidence interval of roughly $[-7, 7]$: The width of the confidence interval itself is about fourteen points of turnout. Even if percent black were a perfect predictor of turnout (which it is not, with a pair adjusted linear relationship of -0.07 in the eight-city case), the p -value of $.2$ indicates that the relationship with treatment assignment is weak enough, and the confidence intervals on the treatment effect itself would be wide enough, to make any random imbalance from percent black a small concern. That is, the p -values reported in Table 32-2 tell us that random imbalances of the sizes seen here will be small relative to the size of the confidence interval calculated on the treatment effect. With a large sample, a small random imbalance is proportionately more worrisome because it is large relative to the standard error of the estimated treatment effect. Given the large confidence intervals on a treatment effect estimated on eight units, the random imbalances shown here are less worrisome — and the p -values encode this worry just as they encode the plausibility of the null.

Thus, even though our eyes suggested we worry about the random imbalance on baseline turnout we saw in Figure 32-2, that amount of imbalance on baseline outcome is to be expected in both the thirty-two and eight-city cases — it is an amount of bias that would have little effect on the treatment effect were we to adjust for it. Of course, the omnibus test for imbalance on all four covariates simultaneously reported in Table 32-1 does cast doubt on the null of balance — and the

tests using the d -statistics in the table suggest that the problem is with percent black and median household income rather than baseline outcomes or number of candidates.⁷

4 Covariates allow adjustment for random imbalance

Even with carefully designed experiments there may be a need in the analysis to make some adjustment for bias. In some situations where randomization has been used, there may be some suggestion from the data that either by accident effective balance of important features has not been achieved or that possibly the implementation of the randomization has been ineffective (Cox and Reid 2000, 29).

A well-randomized experiment aiming to explain something about political participation showing manifest imbalance on education poses a quandry. If the analyst decides to adjust, she then may fall under suspicion: even given a true treatment effect of zero, one adjustment out of many tried will provide a p -value casting doubt on the null of no effect merely through chance. Without adjustment we know how to interpret p -values as expressions of doubt about a given hypothesis: low p -values cast more doubt than high p -values. Adjustment in and of itself does not invalidate this interpretation: a p -value is still a p -value. Concerns center rather on 1) whether an “adjusted treatment effect” is substantively meaningful and how it relates to different types of units experiencing the treatment in different ways — that is, the concerns center on the meaning of “adjustment” in the context of the adjustment method (a linear model or a post-stratification) and 2) whether some specification search was conducted with only the largest adjusted treatment effect reported representing a particularly rare or strange configuration of types of units. Such worries do not arise in the absence of adjustment. Yet, if the analyst declines to adjust, then she knows that part of the treatment effect in her political participation study is due to differences in education, thereby muddying the interpretation of her study.

One may answer such concerns by announcing in advance the variables for which random imbalance would be particularly worrisome and also provide a proposed adjustment and assessment plan a priori. Also, if one could separate adjustment from estimation of treatment effects, one may also avoid the problem of data snooping. For example, Bowers and Panagopoulos (2009) propose a power-analysis based method of choosing covariance adjustment specifications which can be executed independently of treatment effect estimation, and it is well-known that one may post-stratify and/or match without ever inspecting outcomes. Post-stratification may also relieve worries about whether comparisons adjusted using linear models are artifacts of the functional form (Gelman and Hill 2007, ch.9).

I have noted that there are two broad categories of statistical adjustment for random imbalance: adjustment by stratification and adjustment using models of outcomes. In both cases, adjustment amounts to choice of weights; and in both cases adjustment may be executed entirely to enhance precision even if there is no appreciable evidence of random imbalance. Notice that the “unadjusted” estimate may already be an appropriately weighted combination of block-specific treatment effects — and that to fail to weight (or “adjust”) for block-specific probabilities of treatment assignment will confound estimates of average treatment effects (if the probabilities of assignment differ across blocks) and decrease precision (if the variation in the outcomes is much more homogeneous within blocks than across blocks).

Post-stratification enables adjustment but must respect blocking and design.

Say, for example, that within blocks defined by town, the treated group on average contained too many men (and that although gender was important in the study, the researcher either could not or forgot to block on it within existing blocks). An obvious method of preventing “male” from unduly confusing with estimates of treatment effects is to only compare men to men, within block. Analysis then proceeds using the new set of blocks (which represent both the pre-treatment blocks and the new post-treatment strata within them) as before.

One may also use modern algorithmic matching techniques to construct strata. Keele, McConnaughy, and White (2008) argue in favor of matching over linear models for covariance adjustment and show simulations suggesting that such post-stratification can increase precision. Notice that matching to adjust experiments is different from matching in observational studies: matching here must be done without replacement in order to respect the assignment process of the experiment itself and matching must be full. That is, although common practice in matching in observational studies is to exclude certain observations as unmatchable or perhaps to reuse certain excellent control units, in a randomized experiment every observation must be retained and matched only once. This limits the precision enhancing features of matching (at least in theory) since homogeneity will be bounded first by the blocking structure before random assignment and then again by requiring that all observations be matched.

[Figure 32-3 about here.]

Figure 32-3 shows confidence intervals resulting from a variety of post-stratification adjustments made to the thirty two-city turnout data. In this particular experiment the within-set homogeneity increase resulting from post-stratification did not outweigh the decrease in degrees of freedom occurring from the need to account for strata: the shortest confidence interval was for the unadjusted data (shown at the bottom of the plot).

Did the post-stratification help with the balance problems with the Census variables evident in Table 32-1? Table 32-3 shows the strata-adjusted mean differences and p -value for balance tests now adjusting for the post-stratification in addition to the blocking. Balance on baseline turnout and number of candidates does improve somewhat with the matching but balance on percent black and median household income does not appreciably improve. Notice a benefit of post-stratification here: the post-adjustment balance test shows us that we have two covariates which we could not balance.

[Table 32-3 about here.]

Discussion of the advantages of blocking in Section 2 is, in essence, a discussion about how to analyze blocked (pre- or post-stratified) experimental data. The rest of the paper is devoted to understanding what it is that we mean by “covariance adjusted treatment effects.”

Linear models enable adjustment but may mislead the unwary

Even with carefully designed experiments there may be a need in the analysis to make some adjustment for bias. In some situations where randomization has been used, there may be some suggestion from that data that either by accident effective balance of

important features has not been achieved or that possibly the implementation of the randomization has been ineffective (Cox and Reid 2000, 29).

Cox and Reid's discussion in their Section 2.3 entitled "Retrospective adjustment for bias" echoes Cox (1958, 51–2) and Fisher (1925). What they call "bias," I think might more accurately be called "random imbalance".

Although Fisher developed the analysis of covariance using an asymptotic F test which approximated the randomized-based results, others have since noted that the standard sampling-based infinite-population or Normal-model theory of linear models does not justify their use in randomized experiments. For example, in discussing regression standard errors, Cox and McCullagh (1982) note "It is known . . . that 'exact' second-order properties of analysis of covariance for precision improvement do not follow from randomization theory . . ." (547). In this section, I provide an overview of a randomization-based method for covariance adjustment which can use Normal approximations in the same way as those used for balance assessment and placebo-tests above. This method is not subject to the concerns of Freedman (2008a,b,c), and thus suggests itself as useful in circumstances where the linear model as an approximation may cause concern or perhaps as a check on such approximations.

First, though, let us get clear on what it means to "adjust for" random imbalance on a covariate.

What does it mean to say that an estimate has been "adjusted" for the "covariance" of other variables?

Let us look first at how covariance adjustment might work in the absence of blocking by looking only at the first block of eight units in the thirty two-city dataset. Figure 32-4 is inspired by similar figures in Cox (1958, ch. 4) with dark gray showing treated units and black showing control units. The unadjusted difference of means is 6.6 (the vertical distance between the open squares that are not vertically aligned on the gray vertical line). The thick diagonal lines are the predictions from a linear regression of the outcome on an indicator of treatment and baseline outcome. The adjusted difference of means is the vertical difference between the regression lines, here, five. If there had been no relationship between baseline outcomes and post-treatment outcomes, the regression lines would have been flat and the vertical distances between those lines would have been the same as the unadjusted difference of means (the thin dark gray and black horizontal dashed lines). As ought to be clear here, parallel effects is a required assumption for covariance adjustment to be meaningful. In this case, a regression allowing different slopes and intercepts between the treatment groups shows the treatment slope of .25 and the control group slope of .23, thus, the assumption is warranted.

[Figure 32-4 about here.]

What about with blocked data? Figure 32-5 shows the covariance adjustment for three different covariates: a) baseline outcomes (on the left), b) percent black (in the middle), and c) median household income (in \$1000s, on the right). In each case the data are *aligned* within each block by subtracting the block-mean from the observed outcome (i.e., block centered). A linear regression of the block-mean centered outcome on the block-mean centered covariate plus the treatment indicator is equivalent to a linear regression of the observed outcome on the covariate plus treatment indicator

plus indicator variables recording membership in blocks (i.e., “fixed effects” for blocks).

[Figure 32-5 about here.]

In the leftmost plot, we see that adjustment for baseline outcomes in addition to the blocking structure does very little to change the treatment effect: in fact, these blocks were chosen with reference to baseline outcomes, and so adjusting for blocks is roughly equivalent to adjusting for baseline outcomes (only it does not require what is clearly a dubious parallel lines assumption). If the parallel lines assumption held in this case, however, we might talk meaningfully about an adjusted effect. The average treatment effect in the first block is 6.6, but the treated units were more likely to participate, on average, than the control units even at baseline (a mean difference of 4.1). Some of the 6.6 points of turnout may well be due to baseline differences (no more than 4.1 we assume, and probably less so, since it would only matter to the extent that baseline turnout is also related by chance in a given sample to treatment assignment). In this case, the block-specific relationship between baseline outcomes and treatment assignment is vanishingly small (difference of means is 1.9), so only about two points of the average treatment effect is due to the baseline treatment effect (where “due to” is in a very specific linear smoothed conditional means sense). The adjusted effect is actually closer to 5 than 4.6 because the intuitions here provided with differences of means is not identical to what is happening with an analysis of covariance (although it is close and provides helpful intuition in this case).

The middle and right hand plots show two instances in which the intuitions using differences of means become more difficult to believe. In strata 1, 2, and 4, every control unit has a higher percent black than every treated unit. The unadjusted average treatment effect is 1.8, but after adjustment for percent black the effect is 1.2. The middle plot, however, shows that the assumption of parallel effects is hard to sustain and that there is little overlap in the distributions of percent black between the treatment and control groups. In this case, however, the adjustment makes little difference in the qualitative interpretation of the treatment effect.

The right hand figure is a more extreme case of the middle figure. This time there is no overlap at all between the distributions of median income between the controls (in black, and all on the left side of the plot) and the treated units (in dark gray, and all on the right side of the plot). The adjustment causes the treatment effect to change sign — from 1.8 to -2.1 percentage points of turnout. Is -2.1 a better estimate of the treatment effect than 1.8? Clearly median income has a strong relationship with outcomes and also, via random imbalance, with treatment assignment (recall the test casting doubt on the null of balance in Table 32-1).

What is the problem with covariance adjustment in this way? First, as noted earlier, the assumption of parallel lines is not correct. Second, we begin to notice another problem not mentioned in textbooks like Cox and Reid (2000) or Cox (1958) — random assignment will, in large samples, ensure balance in the distributions of covariates but will not ensure such balance in small samples. This means that the distributions of the covariates in theory, ought to be quite similar between the two groups. However, the theory does not exclude the possibility of random imbalance on one of many covariates, and it is well known that random imbalance can and does appear in practice. Adjustments for such imbalance can be done in such a way that adjusted mean differences are still meaningful representations of the treatment effect (as shown by the adjustment for baseline outcomes in the left plot of Figure 32-5). But, as the distributions of covariates become more

unbalanced, the covariance adjustment can mislead. It is hard to claim, based on these data, that adjusting for median household income is a meaningful operation — one just cannot imagine (in these data) finding groups of treated and control units with the same median household income.⁸

Thus, we can see where some criticisms of covariance adjustment might easily come from: covariance adjustment done with multiple regression without additional diagnostics poses a real problem for diagnosing whether the imbalance is so severe as to provide no “common support” in the distributions of the covariates. In such cases, one really cannot “adjust for” the imbalance and one must be resigned to the fact that treatment versus control comparisons, in the data, reflect something other than treatment assignment even if they would not in a larger sample or across repeated experiments. Of course, as sample size grows, given a finite set of covariates and a treatment with finite variance (i.e., a treatment which only has a few values that does not gain values as sample size grows), we would expect the problem of common support to disappear in randomized studies. Luckily, in many studies, one can assess such problems before treatment is administered.

Randomization alone can justify statistical inference about covariance-adjusted quantities without requiring the common assumptions required to justify standard linear regression tables

A predominant use for covariance adjustment is not to ameliorate random imbalance but to enhance statistical precision. To the extent that a covariate predicts outcomes, one may use it to reduce the noise in the outcome unrelated to treatment assignment, and thus help make treatment effects manifest. Covariance adjustment (whether for precision or for random imbalance) means linear regression. In theory, counter-factual statistical inference using the linear regression model for covariance adjustment estimator is biased (Freedman 2008*b,c,a*) but, in practice, it is often an excellent approximation (Green 2009; Schochet 2009). What should we do when we worry about the approximation: for example when the experiment is small, or there is great heterogeneity in effects and/or variance of effects across blocks, or great heterogeneity or discreteness in the outcome (such that the central limit theorem takes longer to kick in than one would prefer)? Rosenbaum (2002*a*) presents a simple argument which builds on the basics of Fisher’s randomization-based inference. Here I provide some very brief intuition to guide study of that paper. This method of randomization-justified covariance adjustment does not rely on the linear model for statistical inference but does “adjust” using the linear model.

Say an outcome is measured with noise caused in part by covariates. When we randomly assign treatment we are attempting to isolate the part of the variation in the outcome due to the treatment from that due to other factors. Say we are still interested in the difference in mean outcomes between treatment and control groups as assigned. The standard deviations of those means may be large (making the treatment effect hard to detect) or small (making the treatment effect more easily manifest). If part of the noise in the outcome is due to covariates, then the residual from regressing the outcome on the covariates represents a less noisy version of the outcome — the outcome without noise from linear relationships with covariates. This residual e_{ib} (for unit i in block b) is measured in units of the outcome (i.e., “percent turning out to vote” in our running fake data example). The potential outcomes to treatment and control for units i in blocks b , y_{Tib} and y_{Cib} , are fixed, Y_{ib} is random by virtue of its relationship with random assignment Z because

$Y_{ib} = Z_{ib}y_{Tib} + (1 - Z_{ib})y_{Cib}$. A null hypothesis tells us what function of Y_i and Z_i would recover y_{Cib} : that is, if the null hypothesis is correct, then removing the effect (say, τ_{ib}) from the treated units $Y_{ib, Z=1}$ would tell us how the treated units would behave under control. Under a constant, additive model of effects, $y_{Tib} = y_{Cib} + \tau$ and so $Y_{ib} - Z_{ib}\tau_{ib} = y_{Cib}$.⁹ So, considering our null hypothesis for the sake of argument, $H_0 : \tau_0 = \tau_{ib}$, regressing $(Y_{ib} - Z_{ib}\tau_{ib})$ on x_i is regressing a fixed quantity (i.e., y_{Cib}) on another fixed quantity, x_{ib} and so the residuals from that regression are a fixed quantity.¹⁰ Thus one may substitute e for x in (1) and (4). Fisher's style of inference begins with a test of a null hypothesis and inverts the hypothesis for a confidence interval: thus, the method allows for us to infer consistent estimates of the causal effect by testing a sequence of causal hypotheses τ_0 . Very loosely speaking, the point estimate is the causal effect hypothesised by the best-supported hypothesis tested.¹¹ Note that this is a method of hypothesis testing, not of estimation. It would be quite incorrect to interpret the difference means of residuals as an estimate of a treatment effect, because the residuals have already have specific causal hypotheses built into them as just described.

[Figure 32-6 about here.]

Figure 32-6 shows that the Rosenbaum covariance adjustment in this particular data is well approximated by the direct regression-style covariance adjustment in the unadjusted case or when the adjustment is made for baseline turnout and number of candidates — and the version adjusted for baseline turnout and number of candidates (just) excludes zero from its 95% confidence interval. The two approaches differ when the difference of means is adjusted for the Census variables. The most notable difference here is for median household income where the direct adjustment method is based entirely on the linear extrapolation between the groups while the Rosenbaum approach correctly captures the sense in which there is no reasonable way to adjust the treatment effect for this covariate. Since adjustment for percent black also requires much linear extrapolation, the randomization-based confidence interval again reflects the fact that the design itself has little information about what it means to adjust for this variable.

The advantages of this style of covariance adjustment are: 1) that it side-steps Freedman's critiques of covariance adjustment for experiments,¹² 2) although we used large-sample normal approximations to evaluate the differences of means here, neither differences of means nor large-sample normal approximations are necessary (and the large-sample approximations are checkable),¹³ 3) unmodeled heteroskedasticity or incorrect the functional form do not invalidate the statistical inference based on this method as they do for the standard approach. For example, the parallel lines assumption is no longer relevant for this approach since the only job of the linear model is to reduce the variance in the outcome. Finally, this particular data example allows us to notice another side benefit of the statistical property of correct coverage: when there is no (or little) information available in the design, the randomization-based approach will not overstate the certainty of conclusions in the same way as the model-based approach.¹⁴

Best Practices for Regression-Based Adjustment

Adjusted treatment effect estimates always invite suspicion of data snooping or modeling artifacts. None of the techniques discussed here entirely prevents such criticism. Of course, the easy way to avoid such criticism is to announce in advance what kinds of random imbalance are most worrisome,

and announce a plan for adjustment (including a plan for assessing the assumptions of the adjustment method chosen). Covariance adjustment using the standard linear regression model requires that one believe the assumptions of that model. For example, this model as implemented in most statistical software requires a correct model of the relationship between outcomes and covariates among treatment and control units (i.e. a correct functional form), that the heteroskedasticity induced by the experimental manipulation is slight, and that the sample size is large enough to overcome the problems highlighted by Freedman (2008*a,b,c*). As with any use of linear regression, one may assess many of these assumptions. If one or more of these assumptions appear tenuous, however, this paper has shown that one may still use the linear model for adjustment but do so in a way that avoids the need to make such commitments. Readers interested in the Rosenbaum (2002*a*) style of covariance-adjustment should closely study that paper. The code contained in the reproduction archive for this paper may also help advance understanding of that method.

5 The more you know, the more you know

A randomized study which allows “the phenomenon under test to manifest itself” provides particularly clear information and thus enhances theory assessment, theory creation, and policy implementation. Thus, researchers should attend to those elements of the design and analysis that would increase the precision of their results. This paper points to only a very small part of the enormous body of methodological work on the design of experiments.

Random assignment has three main scientific aims: 1) it is designed to balance distributions of covariates (observed and unobserved) such that, across repeated randomizations, assignment and covariates should be independent, 2) it is designed to allow assessment of the uncertainties of estimated treatment effects without requiring populations or models of outcomes (or linearity assumptions), and 3) it is a method of manipulating putatively causal variables in a way that is impersonal — and thus enhances the credibility of causal claims. Political scientists have recently become excited about experiments primarily for the first and third aims but have ignored the second aim. This article has discussed some of the benefits (and pitfalls) of the use of covariates in randomized experiments while maintaining a focus on randomization as the basis for inference.

While randomization allows statistical inference in experiments to match the causal inference, covariate imbalance can and does occur in experiments. Balance tests are designed to detect worrisome imbalances. One ought to worry about random imbalances when they are 1) large enough (and relevant enough to the outcome) that they should make large changes in estimates of treatment effects and 2) large relative to their baseline values such that interpretation of the treatment effect could be confused.

Small studies provide little information to help detect either treatment effects or imbalance. The null randomization distribution for a balance test in a small study ought to have larger variance than said distribution in a large study. The same observed imbalance will cast more doubt on the null of balance in a large study than it will in a small study: the observed value will be farther into the tail of the distribution characterizing the hypothesis for the large study than it will in the small study. The same relationship between small and large studies holds when the test focuses on the treatment effect itself. Thus, a *p*-value larger than some acceptance threshold for the null hypothesis of balance tells us that the imbalance is not big enough to cause detectable changes in assessment. 888

of treatment effects. A p -value smaller than some acceptance threshold tells us that the imbalance is big enough to cause detectable changes when we gauge the effects of treatment.

Given random imbalance, what should one do? Adjustment can help, but adjustment can also hurt. This paper showed (using a fake dataset built to follow a real dataset) a case in which adjustment can help and seems meaningful and a case in which adjustment does not seem meaningful, as well as an intermediate case. One point to take away from these demonstrations is that some imbalance can be so severe that real adjustment is impossible. Just as is the case in observational studies, merely using a linear model without inspecting the data can easily lead an experimenter to mislead herself — and problems could multiply when more than one covariate is adjusted for at a time. Rosenbaum (2002a)'s proposal for a randomization-based use of linear regression models is attractive in that covariance-adjusted confidence intervals for the treatment effect do not depend on a correct functional form for the regression model. In this paper all of the adjustment for median household income depended on a functional form assumption so the randomization-based confidence interval was essentially infinite (signaling that the design of the study had no information available for such adjustment) while the model-based regression confidence interval, while much wider than the unadjusted interval, was bounded. Modern matching techniques may also help with this problem (see Keele, McConnaughy, and White 2008). In this paper precision was not enhanced by matching within blocks but matchings including median household income did not radically change confidence intervals for the treatment effect—and balance tests before and after matching readily indicated that the matchings did not balance median household income.

This paper has not engaged with some of the other circumstances in which covariate information is important for randomized studies. In particular, if outcomes are missing, then prognostic covariates become ever more important in experimental studies given their ability to help analysts build models of missingness and models of outcomes (Barnard et al. 2003; Horiuchi, Imai, and Taniguchi 2007). Thus, I have understated the value of collecting more information about the units one has. The trade-offs between collecting more information about units versus including more units in a study ought to be understood from the perspectives long articulated in the many textbooks on experimental design: simple random assignment of units to two treatments (treatment versus control) can be a particularly inefficient research design.

Notes

¹*Acknowledgements:* Many thanks to Jamie Druckman, Kevin Esterling, Don Green, Ben Hansen, Don Kinder, Jim Kuklinski, Thomas Leeper, Costas Panagopoulos, and Cara Wong. Parts of this work were funded by NSF Grants SES-0753168 and SES-0753164.

²See Bowers and Panagopoulos (2009) and Rosenbaum (2002*b*, ch. 2) for accessible introductions to randomization inference; a mode of inference developed in different yet compatible ways by Neyman (1990) (as a method of estimating mean differences) and (Fisher 1935, ch. 2) (as a method of testing). In this paper, I follow the Fisher-style approach in which causal effects are inferred from testing hypotheses rather than estimated as points. Both methods (producing an plausible interval for causal effects using a point \pm a range or testing a sequence of hypotheses) often produce identical confidence intervals.

³This paper is written in the mixture of R and \LaTeX known as Sweave (Leisch 2002, 2005) and, as such, all of the code required to produce all of the results, tables, and figures (as well as additional analyses not reported) are available for learning and exploration from <http://jakebowers.org>. Thus, I will spend relatively little time discussing the details of the different methods, assuming that those interested in learning more will download the source code of the paper and apply themselves to adapting it for their own purposes.

⁴The actual computations used the `xBalance` command found in the `RItools` library for R (Bowers, Fredrickson, and Hansen 2009) as described in Hansen and Bowers (2008).

⁵Note that it would be strange to see such imbalances in a study run with thirty-two observations rather than an eight observation study expanded to thirty-two observations artificially as done here. These imbalances, however, dramatize and clarify benefits and dangers of post-treatment adjustment as explained throughout this paper.

⁶Echoing Senn (1994) among others in the clinical trials world, Imai, King, and Stuart (2008) provide some arguments against hypothesis tests for balance in observational studies. Hansen (2008) answers these criticisms with a formal account of the intuition provided here pointing out that 1) randomization-based tests do not suffer the problems highlighted by those authors and 2) highlighting the general role that p -values play in randomization based inference.

⁷If we had twenty covariates and rejected the null of balance with $p < .05$, we would expect to falsely see evidence of imbalance in one of twenty covariates. Thus, Hansen and Bowers (2008) urge the use of an omnibus test — a test which assesses balance across all linear combinations of the covariates in the table. Yet, the variable by variable display is useful in the same way that graphs such as Figure 32-2 are useful — in suggesting (not proving) the sources of imbalance.

⁸Notice that this problem also occurred in Figure 32-4 but was not mentioned in order not to detract from the pedagogical task of describing the mechanism of covariance adjustment.

⁹If this model is incorrect, then randomization-based inferences will be conservative but the coverage of the confidence intervals will still be correct as noted independently by (Gadbury 2001) and Robins (2002, § 2.1), and of course, other substantively meaningful models of effects are available Rosenbaum (ch. 5 2002*b*, see) and Rosenbaum (§ 6 2002*a*, and also see) or Rosenbaum

(2010, ch. 2). For example, as Rosenbaum (2002c, 323) notes, if the treatment effect varies by a binary covariate x coding 0 for group 1 and 1 for group 2 (such that the parallel lines assumption is incorrect), we would then specify the potential responses to control as $Y_{ib} - Z_{ib}(\tau_{x=1}x_{ib} + \tau_{x=0}(1 - x_{ib}))$ for treatment effects that differ by group. I use the constant additive effects model in this paper to map most closely onto the causal quantities implied by the choice of a linear regression model for covariance adjustment: indeed, for this very reason I can show how both styles of covariance adjustment can produce identical quantities in Figure 32-6. Interested readers might find the discussion in Rosenbaum (2002c, § 3–6) useful for thinking about the equivalence of estimating an average treatment effect and testing a sequence of hypotheses about individual causal effects.

¹⁰For a single covariate, x and a regression fit $(Y_{ib} - Z_{ib}\tau_{ib}) = \hat{\beta}_0 + \hat{\beta}_1x_{ib}$, $e_{ib} = (Y_{ib} - Z_{ib}\tau_{ib}) - (\hat{\beta}_0 + \hat{\beta}_1x_{ib})$. The residual is written e , not \hat{e} , because the regression fit is not an estimate of an unknown quantity but merely calculating a function of fixed features of the existing data.

¹¹See discussion of the Hodges-Lehmann point estimate in Rosenbaum (2002a) and Rosenbaum (2002b, ch. 2) for more formal discussion of randomization-justified point-estimates of causal effects. In the context of a large, cluster randomized, field experiment with binary outcomes and non-random non-compliance, Hansen and Bowers (2009) show how one may approximate the results of testing sequences of hypotheses with simple calculations of means and associated randomization-based variances, including a method for randomization-based covariance adjustment. Because the Hansen and Bowers (2009) method approximates the results of testing sequences of hypotheses with simple means and variances their method requires an asymptotic justification. That paper (and related reproduction materials) (Bowers, Hansen, and Fredrickson 2008) also provide some tools for assessing the asymptotic justification.

¹²In particular, Freedman (2008b, 189) notes that the Fisher-style covariance adjustment is valid “If $T_i = C_i$ for all i (the “strict null hypothesis”), then $\beta \equiv 0$ and adjustment will help—unless $\overline{\alpha Z} = 0$, i.e., the remaining variation (in C_i) is orthogonal to the covariate.” Another method, elaborated most recently in Hansen and Bowers (2009) also does not directly model the relationship between treatment and outcomes and so also avoids this critique.

¹³Rosenbaum (2002a) focuses on normal approximations to the Wilcox rank sum statistic as his preferred summary of treatment effects (and the normal approximations there are not necessary either, but merely convenient and often correct in large-enough samples with continuous outcomes).

¹⁴The disadvantages of this mode are not on display here (although they will be obvious to those who use the code contained in this paper for their own work). First, remember, this approach produces confidence intervals by testing sequences of hypotheses. It does not “estimate” causal effects as a standard regression estimator would but rather assesses the plausibility of causal effects using tests. Of course, such assessments of plausibility are also implicit in confidence intervals for standard regression estimators. However, the mechanics of the two methods of covariance adjustment are quite different. The randomization-based approach as implemented here builds a confidence interval by direct inversion of hypothesis tests: in this case, we tested hypotheses about the treatment effect from $\tau_0 = -20$ to $\tau_0 = 20$ by .1. This can be computationally burdensome if the number of hypotheses to test is large or if we eschew large-sample approximations. Second, we did not work to justify our choice of mean difference (rather than rank, or other summary of observed outcomes and treatment assignment). The standard linear regression estimator requires

attention to mean-differences as the quantity of interest whereas any test-statistic may be used in the randomization-based method of adjustment shown here.

References

- Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98(462): 299–24.
- Bowers, Jake, and Costas Panagopoulos. 2009. "Probability of What?: A Randomization-based Method for Hypothesis Tests and Confidence Intervals about Treatment Effects."
- Bowers, Jake, Ben B. Hansen, and Mark Fredrickson. 2008. Reproduction archive for: 'Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign'. Technical report University of Illinois and University of Michigan.
- Bowers, Jake, Mark Fredrickson, and Ben Hansen. 2009. *Rltools: Randomization Inference Tools*. R package version 0.1-6 ed.
- Brady, Henry E. 2008. "Causation and Explanation in Social Science." In *Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier.
- Cochran, William G., and Gertrude M. Cox. 1957. *Experimental Designs*. New York: John Wiley & Sons.
- Cox, David R. 1958. *The Planning of Experiments*. New York: John Wiley.
- Cox, David R., and Nancy Reid. 2000. *The Theory of the Design of Experiments*. Boca Raton, FL: Chapman & Hall/CRC.
- Cox, David R., and Peter McCullagh. 1982. "Some Aspects of Analysis of Covariance (with discussion)." *Biometrics* 38: 541–61.
- Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.
- Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Freedman, David A. 2008a. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2: 176–96.
- Freedman, David A. 2008b. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40: 180–93.
- Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23: 237–49.
- Gadbury, G.L. 2001. "Randomization Inference and Bias of Standard Errors." *The American Statistician* 55: 310–3.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. New York: Cambridge University Press.

- Green, Donald P. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science?"
- Hansen, Ben, and Mark Fredrickson. 2010. *Optmatch: Functions for Optimal Matching, Including Full Matching*. R package version 0.6-1 ed.
- Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99(467): 609–18.
- Hansen, Ben B. 2008. "Comment: The Essential Role of Balance Tests in Propensity-matched Observational Studies." *Statistics in Medicine* 27: 2050–4.
- Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23: 219–36.
- Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association* 104: 873–85.
- Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51: 669–87.
- Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24: 29–72.
- Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171: 1–22.
- Keele, Luke J., Corrine McConnaughy, and Ismail K. White. 2008. "Adjusting Experimental Data."
- Leisch, Friedrich. 2002. Dynamic Generation of Statistical Reports Using Literate Data Analysis. In *Compstat 2002 - Proceedings in Computational Statistics*, ed. Wolfgang Haerdle, and Bernd Roenz. Heidelberg, Germany: Physika Verlag.
- Leisch, Friedrich. 2005. *Sweave User Manual*.
- Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (1923)." *Statistical Science* 5: 463–80.
- Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233–52.
- Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment."
- Robins, James M. 2002. "Covariance Adjustment in Randomized Experiments and Observational Studies: Comment." *Statistical Science* 17: 309–21.

- Rosenbaum, Paul R. 2002a. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Statistical Science* 17: 286–327.
- Rosenbaum, Paul R. 2002b. *Observational Studies*. 2nd ed. New York: Springer-Verlag.
- Rosenbaum, Paul R. 2002c. "Rejoinder." *Statistical Science* 17(August): 321–7.
- Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.
- Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5: 472–80.
- Schochet, Peter. 2009. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference." *Journal of Statistical Planning and Inference* 140: 246–59.
- Sekhon, Jasjeet S. 2008. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." In *Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier.
- Senn, Stephen J. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13: 1715–26.

List of Tables

32-1	Balance tests for covariates adjusted for blocking in the blocked thirty two-city Study.	24
32-2	Balance tests for covariates in the blocked 8 City Study.	25
32-3	Balance tests for covariates adjusted for covariates by post-stratification in the blocked thirty two-city study.	26

List of Figures

32-1	The efficiency of paired and unpaired designs in simulated turnout data.	27
32-2	Graphical assessment of balance on distributions of baseline turnout for the thirty two city experiment data.	28
32-3	Post-stratification adjusted confidence intervals for the difference in turnout between treated and control cities in the thirty two-city turnout experiment.	29
32-4	Covariance-adjustment in a simple random experiment.	30
32-5	Covariance-adjustment in a blocked random experiment	31
32-6	Covariance-adjusted confidence intervals for the difference in turnout between treated and control cities in the thirty-two-city turnout experiment data.	32

x	\bar{x}_{Ctrl}	\bar{x}_{TtT}	d_{blocks}	$\sqrt{\text{Var}(d_{\text{blocks}})}$	Std. d_{blocks}	z	p
Baseline Outcome	27.1	25.8	-1.2	0.9	-0.1	-1.4	0.2
Percent Black	17.3	2.6	-14.7	4.2	-1.2	-3.5	0.0
Median HH Income	30.6	46.3	15.7	3.2	2.6	4.9	0.0
Number of Candidates	13.2	10.3	-2.9	2.0	-0.5	-1.5	0.1

Table 32-1: One-by-one balance tests for covariates adjusted for blocking in the blocked thirty two-city study. Two-sided p -values are reported in the p column based on referring z to Normal distribution approximating the distribution of the mean-difference under null of no effects. An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates at $p=0.00003$. Test statistics (d_{blocks} , etc..) are generalizations of equations (1) and (4) developed in Hansen and Bowers (2008) and implemented in Bowers, Fredrickson, and Hansen (2009). Statistical inference here (z and p -values) is randomization-based but uses large-sample Normal approximations for convenience. Other difference-of-means tests without the large-sample approximation, and other tests such as Kolmogorov-Smirnov tests and Wilcoxon rank sum tests, provide the same qualitative interpretations. For example, the p -values on the tests of balance for baseline outcome (row 1 in the table) ranged from $p = 0.16$ and $p = .17$ for the simulated and asymptotic difference of means tests respectively, to $p = .33$ and $p = .72$ for exact and simulated Wilcoxon rank sum and Kolmogorov-Smirnov tests.

x	\bar{x}_{Ctrl}	\bar{x}_{Tt}	d_{blocks}	$\sqrt{\text{Var}(d_{\text{blocks}})}$	Std. d_{blocks}	z	p
Baseline Outcome	26.0	24.8	-1.2	2.0	-0.1	-0.6	0.5
Percent Black	16.8	2.4	-14.4	11.0	-1.1	-1.3	0.2
Median HH Income	29.2	44.5	15.4	8.1	2.5	1.9	0.1
Number of Candidates	13.0	10.0	-3.0	5.1	-0.5	-0.6	0.6

Table 32-2: One-by-one balance tests for covariates in the blocked eight city study. An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates at $p=0.41$. Test statistics (d_{blocks} , etc..) are generalizations of equations (1, 4) developed in Hansen and Bowers (2008) and implemented in Bowers, Fredrickson, and Hansen (2009). Statistical inference here (z and p -values) is randomization-based but uses large-sample Normal approximations for convenience.

x	Post-Hoc Full-Matching on:					
	Blocks		Baseline Turnout		Propensity Score	
	d_{strata}	p	d_{strata}	p	d_{strata}	p
Baseline Outcome	-1.2	0.2	-1.2	0.3	-1.2	0.3
Percent black	-14.7	0.0	-14.7	0.0	-14.7	0.0
Median HH Income	15.7	0.0	15.8	0.0	15.7	0.0
Number of candidates	-2.9	0.1	-2.6	0.3	-2.9	0.3

Table 32-3: One-by-one balance tests for covariates adjusted for covariates by post-stratification in the blocked thirty two-city study. Two-sided p -values assess evidence against the null of no effects. An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates adjusted for Blocks, and two kinds of Post-Hoc full-matching within blocks at $p=0.00003, 0.003, 0.004$. Strata adjusted mean-differences (d_{strata}) are generalizations of equations (1) developed in Hansen and Bowers (2008) and implemented in the `RITools` package for R (Bowers, Fredrickson, and Hansen 2009). Statistical inference here (p -values) is randomization-based but uses large-sample Normal approximations for convenience. Post-hoc stratification results from optimal, full-matching (Hansen 2004) on either absolute distance on baseline turnout or absolute distance on a propensity score with propensity caliper penalty shown in Figure 32-3.

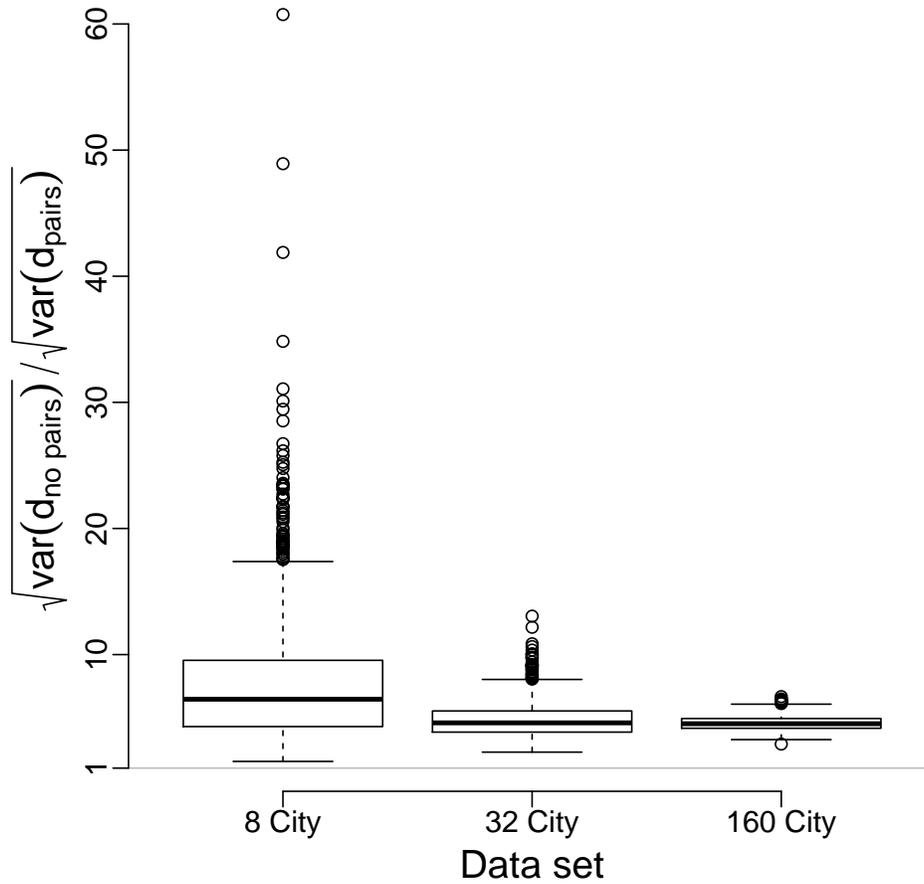


Figure 32-1: Paired designs were always more efficient than unpaired designs in the simulations described in Section 2. The unpaired design dominates the paired design less dramatically as the size of the experiment increases.

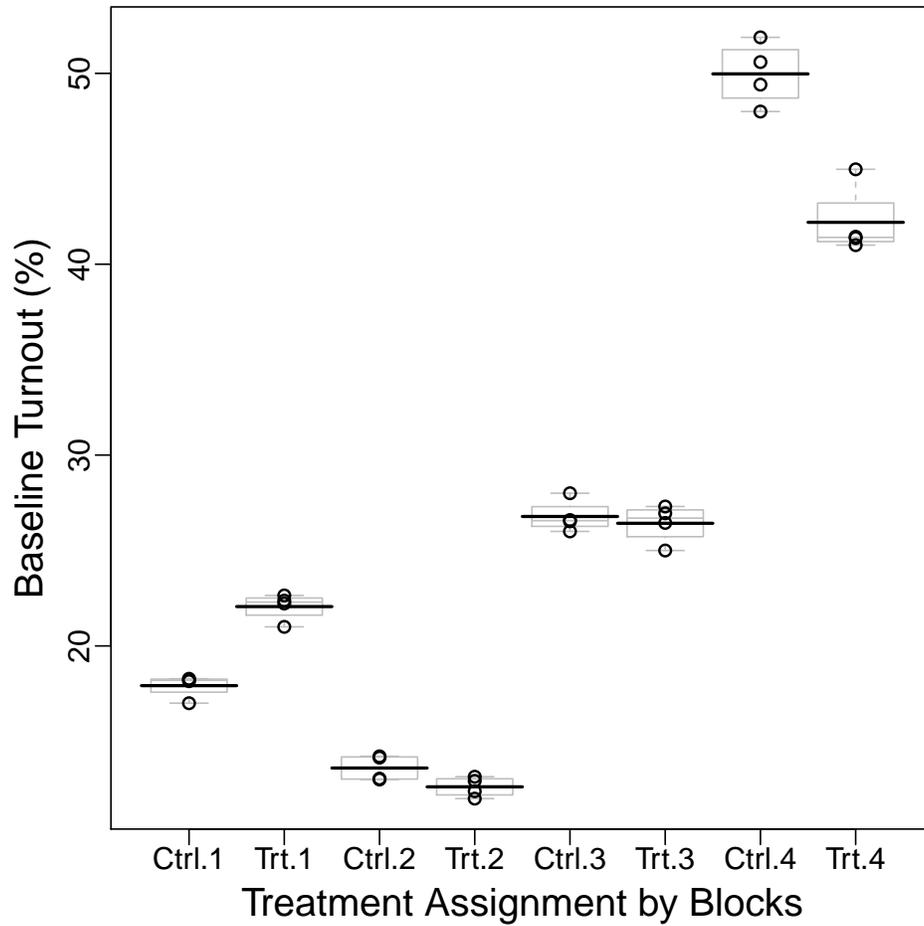


Figure 32-2: Within each of the four strata (1,2,3,4) the baseline outcomes (past turnout) for the thirty two-city data are plotted by assignment condition (Trt=Treatment, Ctrl=Control). Means are long, thick black lines. Boxplots in gray.

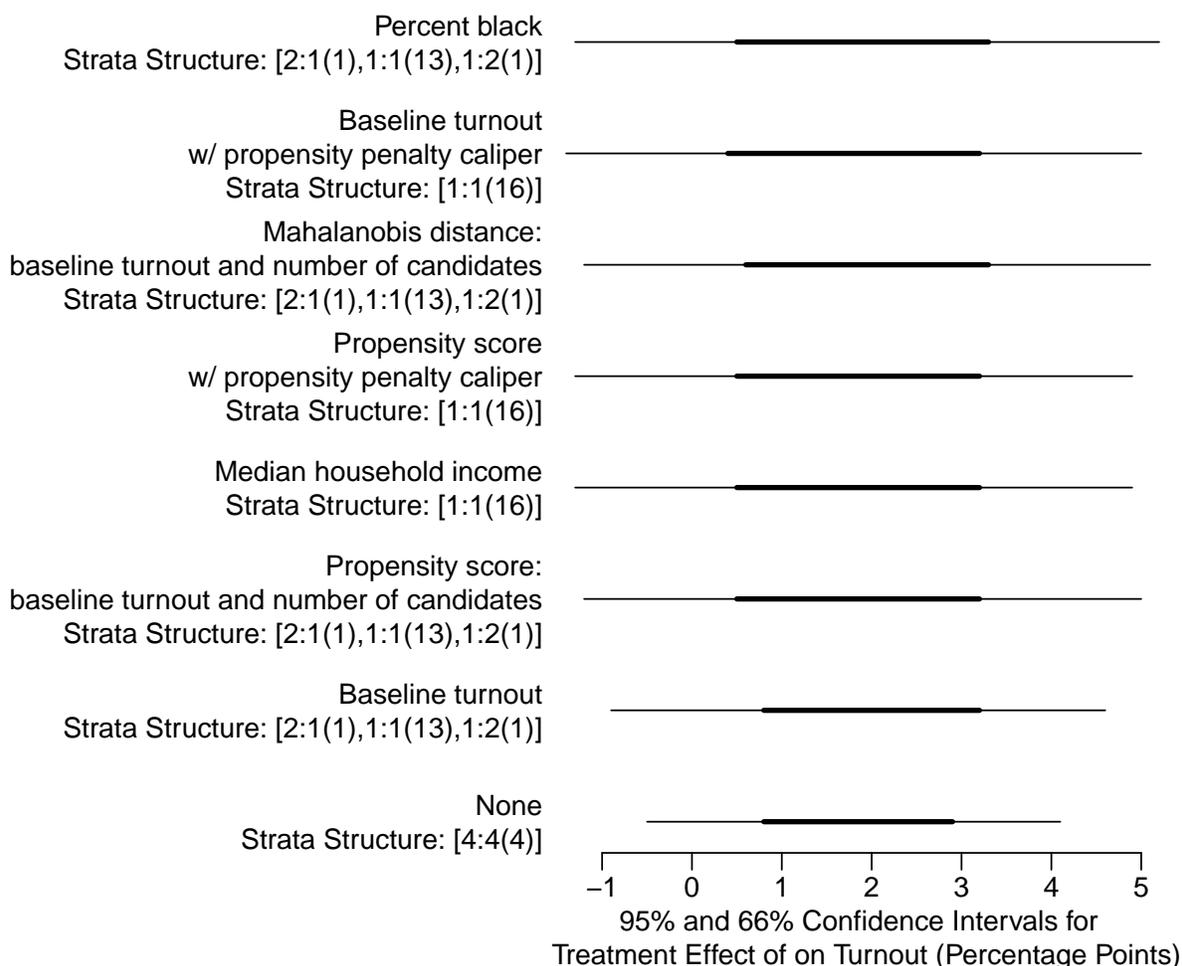


Figure 32-3: Confidence intervals for the difference in turnout between treated and control cities in the thirty two-city turnout experiment data ordered by width from narrowest at bottom to widest at top. Each line represents the interval for the treatment effect after different post-stratification adjustments have been applied within the existing 4 blocks of 8 cities. Thin lines show the 95% intervals. Thick lines show the 66% intervals. The propensity-score was calculated from a logistic regression of treatment assignment on baseline turnout, number of candidates running for office, percent black and median household income (both from the 2000 Census), and block-indicators. All post-stratification and interval estimation was done with full, optimal matching using the `optmatch` (Hansen and Fredrickson 2010) and `RITools` (Bowers, Fredrickson, and Hansen 2009) packages for R. Numbers below the post-stratification label show the structure of the stratification: for example, without any post-stratification the experiment had 4 blocks, each with 4 treated and 4 control cities. The matching on absolute distance on the propensity score with a propensity caliper penalty produced 16 pairs of treated and control cities (1:1(16)). The match on absolute distance on baseline turnout produced 1 set with 2 treated and 1 control (2:1(1)), 13 paired sets (1:1(13)) and 1 set with 1 treated and 2 controls (1:2(1)). Since no observation could be excluded, calipers implied penalties for the matching algorithm rather than excluding observations from the matching (Rosenbaum 2010, ch 8).

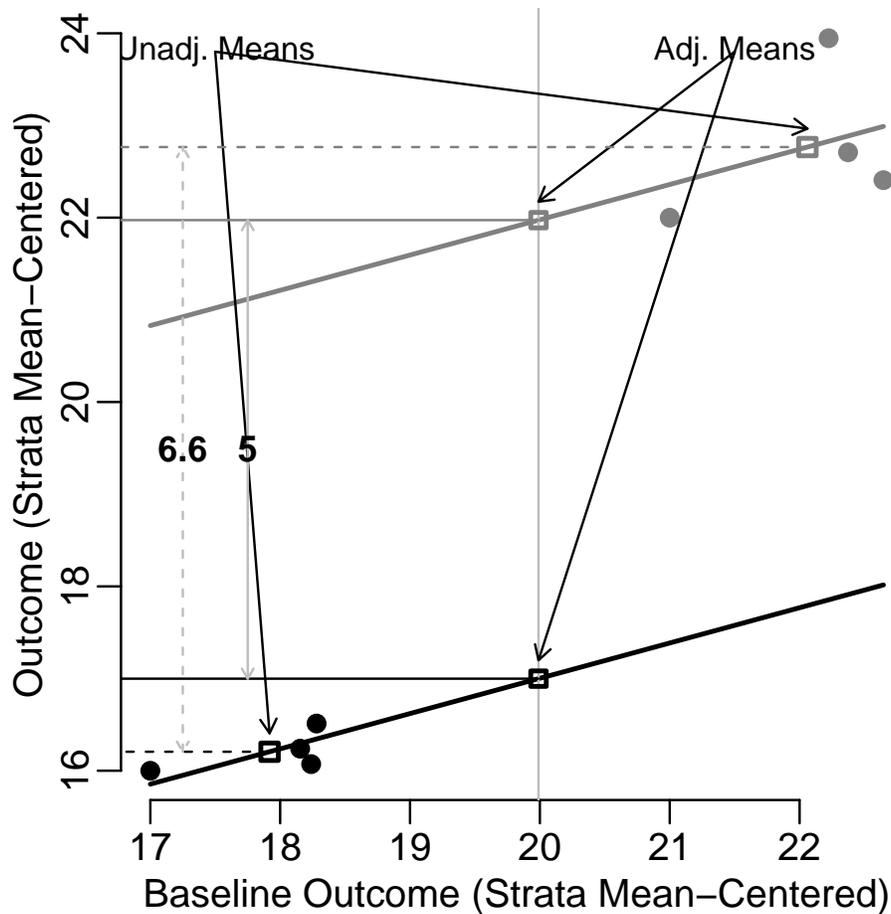


Figure 32-4: Covariance-adjustment in a simple random experiment. Dark gray and black circles show treated and control units respectively. The unadjusted difference of means is 6.6. The thick diagonal lines are: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 y_{i,t-1}$ with $Z_i = \{0, 1\}$ and the adjusted average treatment effect is $(\hat{Y}_i|Z_i = 1, y_{i,t-1} = \bar{y}_{t_1}) - (\hat{Y}_i|Z_i = 0, y_{i,t-1} = \bar{y}_{t_1}) = 5$.

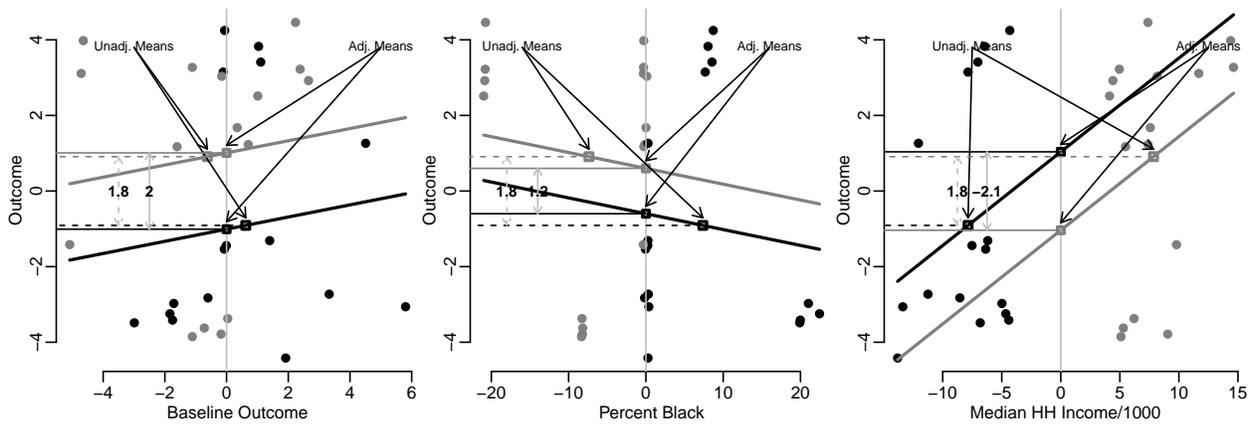


Figure 32-5: Covariance-adjustment in a blocked random experiment (4 blocks). Dark gray and black circles show treated and control units respectively. All variables are block-mean centered.

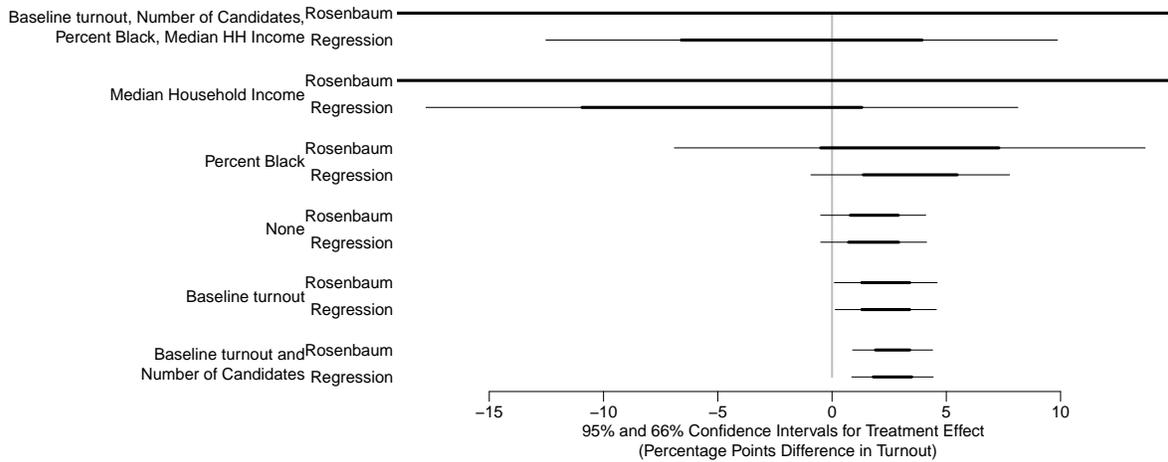


Figure 32-6: Confidence intervals for the difference in turnout between treated and control cities in the thirty-two-city turnout experiment data by type of covariance adjustment using least squares regression ordered by width from narrowest at bottom to widest at top. “Regression”-style adjustment regressed turnout on treatment indicator, block indicators, and covariates and referred to the standard iid+*t*-based reference distribution for confidence intervals. “Rosenbaum”-style adjustment regressed turnout on block indicators, and covariates, and then used the residuals as the basis for tests of the null hypotheses implied by these confidence intervals. The reference distributions for the Rosenbaum-style are large-sample approximations to the randomization distribution implied by the design of the experiment using the *RITools* (Bowers, Fredrickson, and Hansen 2009) package for R. Thin lines show the 95% intervals. Thick lines show the 66% intervals. The randomization based confidence intervals for outcomes after adjustment by functions including median household income are essentially infinite.

33. Design and Analysis of Experiments in Multilevel Populations

Betsy Sinclair ⁱ

Randomized experiments are seen as the most rigorous methodology for testing causal explanations for phenomena in the social sciences and are experiencing a resurgence in political science. The classic experimental design randomly assigns the population of interest into two groups, treatment and control. Ex-ante these two groups should have identical distributions in terms of their observed and unobserved characteristics. Treatment is administered based upon assignment, and by the assumptions of the Rubin Causal Model the average effect of the treatment is calculated as the difference between the average outcome in the group assigned to treatment and the average outcome in the group assigned to control.

Randomized experiments are often conducted within a multilevel setting. These settings can be defined both at the level at which the randomization occurs as well as at the level at which the treatment is both directly and indirectly administered. These indirect effects most often occur as a result of social transmission of the treatment, which is particularly likely when the treatment consists of information. This essay will explore the implications of these multilevel settings to highlight the importance of careful experimental design with respect to random assignment of the population and the implementation of the treatment. There are potential problems with analysis in this context, and this essay suggests strategies to accommodate multilevel settings. These problems are most likely to occur in field settings where control is lacking although can sometimes occur in other settings as well. Multilevel settings have the potential to disrupt the

internal validity of the analysis by generating bias in the estimation of the average treatment effect.

There are two common environments where the standard experiment fails to adjust for a multilevel structure and results in problematic estimates, and both occur where the assignment to treatment is at the individual level but the administration of treatment is not. Both of these instances create problems for analysis. The first problem relates to the selection of groups to receive treatment. Suppose the random assignment is conducted at the individual level, but the implementation of the treatment is directly at the group level, and suppose that the selection of which groups to treat is not random but instead is selected by an organization where the selection is correlated with individual voting probabilities. This produces bias in the inferences that result from this setup.ⁱⁱ An example of this type of experiment would be one in which the randomization assigns individuals to treatment and control but administers treatment to particular ZIP codes. Inferences in this context are problematic because of the selection of particular ZIP codes. This problem is solved with clustered randomization and by making the appropriate adjustment to the standard errors (Arceneaux and Nickerson 2009; Green and Vavreck 2008). This is not the problem addressed in this chapter.

The second problem is in terms of social interactions -- we could have possible spillover effects from one individual to another within the same household for example, or furthermore from one household to another within the same group. That is, the random assignment is conducted at the individual level, but the implementation of the treatment is indirectly at the group level. Again, an example of this type of experiment would be one in which the randomization assigns individuals to treatment and control but administers treatment to

households. Inferences at the individual-level are then problematic because the treatment can be socially transmitted within the household. Social science randomized experiments often rely upon treatments that can be socially transmitted. Social transmission has the potential to result in violation of one of the fundamental assumptions in the analysis of these experiments, the Stable Unit Treatment Value Assumption (SUTVA). SUTVA states that there is no interference between units, such that the assignment of an individual to the treatment group should have no effect on outcomes for other individuals. Many randomized field experiments take place in situations where interference between individuals is likely, such as within social settings where social interaction is expected. If ignored, SUTVA violations have the possibility of adding bias to estimated treatment effects, and it is possible that these biases can go in either a positive or negative direction.

Multilevel randomized experiments rely upon existing social structures, which have the potential to provide solutions to the problems that arise from social transmission. A multilevel randomized field experiment design allows for the opportunity to estimate the treatment effect while acknowledging for the possibility of SUTVA violations as well as an explicit test for the degree to which social interactions take place. Making SUTVA an object of study instead of an assumption has the benefit of providing new insights about interpersonal influence. Multilevel experiments provide an opportunity to better understand social transmission of politics. While theories abound about the structure of social environments, little empirical evidence exists to explicitly validate that these structures influence an individual's politics. Multilevel settings occur when individuals are assigned to treatment but will communicate to each other within different social levels, such as within households or within precincts. Multilevel settings require

specific experimental designs and analyses but allow us to estimate the effects of interpersonal interactions. This essay describes both the advantages of a multilevel randomized field experiment and provides recommendations for the implementation and analysis of such experiments consistent with the current best practices in the literature.

1. Spillovers, Models of Diffusion, and the Reflection Problem

Multilevel contexts highlight the role that social interactions may play in political behavior. Turnout decisions may diffuse through a population in much the same way that disease or trends are also transmitted across individuals who are socially connected. Diffusion processes have been clearly seen in marketing effects (Coleman, Katz, and Menzel 1966; Nair, Manchanda, and Bhatia 2008) and are likely to also be present in political behavior. Formal models of spillover effects in political behavior suggest that the ways in which individuals are socially connected are highly likely to determine their beliefs about candidates (DeMarzo, Vayanos, and Zwiebel 2003; Sinclair and Rogers 2010).

These models are difficult to estimate using observational data due to the reflection problem (Manski 1993). This is an identification problem, where it is difficult to separate the selection into group membership from the effect of the group itself. Thus individual level behavior appears correlated with network peers when in fact network peers do not cause this correlation. Instead, the cause of this correlation could be a common factor that affects all members of the group or characteristics that the members of the group are likely to share. Standard regression models are exposed to the reflection problem, as it is impossible to determine whether or not the individual is having an effect on the group or vice versa. Alternative estimation strategies do exist, which include the use of instrumental variables and the

role of time (Brock and Durlauf 1999; Conley and Udry 2009). Yet observational strategies are not generally sufficient to identify causal effects of social interaction.

By using randomized experiments, we are able to gain some leverage on the reflection problem by providing a stimulus to one set of individuals and then observing the behavior of the group. In particular, by looking for empirical evidence of spillovers in this setting, we are then more able to directly test to what extent models of diffusion are applicable to political behavior. By using a multilevel randomized experiment, we are able to directly evaluate both the effect of the treatment and peer-effects. This occurs via the identification of the appropriate multilevel context and is described in the next section.

With limited exceptions, random assignment is typically done at either the individual or cluster level in randomized field experiments. There are many examples where randomization occurs at the cluster level, ranging from assignment by neighborhood, congressional district, or precinct, often because it is difficult to administer treatment by individual (Arceneaux 2005; King et al 2007; Imai, King, and Nall 2009). For example, in experiments on the effects of campaign advertising, randomization often occurs at the level of the media market (Panagopoulos and Green 2006; Green and Vavreck 2008). Some experiments have failed at the individual level – for example, in experiments on policy such as antipoverty efforts (Adato, Coady, and Ruel 2000) and reductions in class size (Krueger 1999) – because the subjects were able to change their own assignment category within a particular group, thus some types of experiments must be conducted at the cluster level. If randomization occurs at the cluster level, then without additional assumptions the administration of treatment and inferences about treatment efficacy will also be done at the cluster level. One difficulty in evaluating the

treatments in these contexts is that individuals may have communicated to each other about the treatment, thus the observed treatment effect may result from an interaction between the direct and indirect administration of treatment. This has implications for both the external and internal validity of the experiment, as it is not possible to separate the direct and indirect treatment effects.

Randomization occurs at the level of the individual as well, for example, in the bulk of experiments on voter mobilization tactics (for a review of the literature, see Gerber and Green 2008). The majority of this essay will address issues involving the estimation of direct and indirect treatment effects when inferences are drawn about the behavior of the individual. When estimating, one's statistical model must account for the level at which randomization occurs. Inferences are generally drawn about the behavior of the individual and it is possible to create inconsistencies between the assignment of individuals to treatment and the implementation of the treatment.

Implementation of the treatment may occur indirectly as the result of social interaction. Many randomized experiments administer treatments that could be socially transmitted, such as political information, a heightened sense of political interest, or increased social trust. Empirical work on social networks has suggested that many of these treatments can be socially transmitted (Fowler and Christakis 2008; Nickerson 2008; Cacioppo, Fowler, and Christakis 2009). Experiments where the treatment is subject to social transmission are implicitly conducted within a multilevel setting.

2. The Stable Unit Treatment Value Assumption

Within a social setting it is possible that spillovers from individuals assigned to treatment to individuals assigned to control could create biases in the estimation of treatment effects, or vice versa. Here we consider the standard setup for a political interest experiment and examine ways in which these biases can be eliminated. Here our narrative, consistent with work by Sinclair, McConnell, and Green (2010), considers an experimental population where individuals are members of a three-level multilevel setting. Individuals are residents in a household and each household resides within a social group. The collection of groups forms the population. This example could generalize to any number of levels or different settings, so long as each sublevel is fully contained within the previous level.

Our primary concern with the assignment and implementation of randomized field experiments within a multilevel population is violations of the *stable unit treatment value assumption* (SUTVA, as labeled in Rubin 1980). Units here are defined as the unit that is being evaluated (for example, if the treatment is being evaluated at the individual level, then the individual is the unit, whereas if the treatment was being evaluated at the group level, then the group is the unit).

SUTVA states that the potential outcomes for any unit do not vary with the treatments assigned to any other units, and there are no different versions of the treatment (Rubin 1990). The assumption of SUTVA is key to how we draw causal inferences about the efficacy of the treatment. The first part of SUTVA assumes that there is no interference between units; that is, it assumes that there are no spillover effects.ⁱⁱⁱ This essay focuses exclusively on the problem of inference between units as a result of treatment, specifically treatment spillovers.

It is possible that the units interfere with each other in the course of receiving treatment. For example, suppose that some individuals are contacted and given additional political information – it seems likely that they might then discuss this information with the other individuals they know in their household or in their neighborhood. In this case there would be interference between units. In this essay, we will investigate the extent to which we can measure potential spillovers and also design experiments in order to be able to correct for their potential effects. We contend that spillover effects could exist within households or within groups.

We now look at an example where, in the presence of spillovers, it would not be possible to ascertain the exact treatment effect if the randomization is conducted at the individual level and there is communication between individuals within the experimental population. We explore spillovers both within their households and within their groups. That is, suppose we have a violation of SUTVA. We consider the violation in terms of the *intent-to-treat* (ITT) effect.

3. Intent-to-Treat Effect

In our population of n individuals, let each individual i be randomly assigned to treatment, $t = 1$, or control, $t = 0$. We investigate the outcome for each individual Y_i . We want to know what the difference is between treatment and control – the treatment effect – and ideally we would like to calculate $Y_{i,t=1} - Y_{i,t=0}$, yet it is not possible to observe both states of the world at the individual level. However, due to the random assignment the group of individuals who receive treatment are ex-ante identical in terms of their characteristics to those who receive control, and we are able to look at the difference in terms of expected outcomes. We define the ITT effect as $ITT = E(Y_i|t = 1) - E(Y_i|t = 0)$. SUTVA allows us to consider only the assignment of individuals. If we assume that there may be spillovers, however, we define the ITT effect in

terms of the assignment of each individual i and all other individuals k . Formally, we can then observe the ITT effect for those instances where i is socially connected to k others who do not receive treatment – that is, $ITT = E(Y_i|t_i = 1, t_k = 0) - E(Y_i|t_i = 0, t_k = 0)$. When individuals are communicating, we must revise our statements to include the assignments for the other individuals in our sample as well. In the limiting case, we must revise our statement to describe all n individual assignments.

Recall that each individual i is either assignment to treatment, where $t_i = 1$, or control, where $t_i = 0$. We observe the outcome for each individual i as Y_i . To consider a case where individuals are communicating within their households, suppose we consider a case where all individuals live in two-person households and the second person in the household is identified as j . When measuring the expected outcome we have to consider the assignment to the second person in the household as well, $E(Y_i|t_i, t_j)$. Thus in order to estimate the treatment effect, we need a particular group of individuals who have been assigned to treatment where the other individuals in their household have been assigned to control, so that $ITT_{hat} = E(Y_i|t_i = 1, t_j = 0) - E(Y_i|t_i = 0, t_j = 0)$. Yet the standard inference would not have incorporated the treatment assignment of j , so that there is a chance that the inferences could be biased due to communication between individuals. That is, suppose there are four individuals where three are assigned to treatment and one is assigned to control, but that we do not draw inferences based upon a multilevel structure. Then we could misestimate the treatment effect as $ITT_{hat} = 1/6 * (Y_1|t_1 = 1, t_2 = 1) + 1/6 * (Y_2|t_1 = 1, t_2 = 1) + 1/6 * (Y_4|t_4 = 1, t_3 = 0) - 1/2 * (Y_3|t_3 = 0, t_4 = 1)$. For individuals 1 and 2, they may be more likely to change their behavior because they both receive treatment, so this suggests that in fact we could overestimate the treatment effect, yet

individual 3 may also be more likely to change behavior since she shares a household with someone who was also in the treatment group, so this suggests that we could in fact underestimate the treatment effect.

Extending this example to groups, we would then have an even more complicated problem where there could be communication between many households within a group. The true ITT would then need to be written based upon all the instances of communication.

The consequences of these spillovers are such that it is possible that in the presence of communication, the estimated treatment effect can be either an overestimate or an underestimate of the true effect. The direction of the bias will depend on the ways in which communication occurs and the effect of communication on an individual's decision – it may be the case that additional communications between treated individuals, for example, will heighten the probability that they behave in a given way. Violations of SUTVA may produce biased estimates in light of communication about treatment. It is not possible to know whether or not the direction of the bias will be positive or negative prior to conducting the experiment. Key to estimating ITT is to understand which individuals are assigned, either directly or indirectly, to treatment.

4. Identifying a Multilevel Context

We identify two types of multilevel contexts. The focus of this essay is to identify instances where individuals are likely to communicate to each other about the treatment, but other scholars may use this phrase in a different type of situation. First, multilevel contexts are likely to exist where the randomization is conducted at a different level from the level at which treatment is administered. This type of multilevel context has the potential to generate correlations within groups about the treatment. Experiments like these occur most often when the

experimenter is relying upon the multilevel structure in order to implement the treatment. Examples include voter mobilization experiments conducted by precinct or household. The design of these experiments must account for this structure. If it is the case that there is any failure-to-treat – that is, if it is possible that there will be some groups where no attempt is made to administer the treatment – then it is helpful that the order of the attempts to contact each group be randomized.^{iv} This randomization both allows for valid causal inferences and makes it impossible for the selection of particular groups to undermine the randomization. So long as all units will be treated, then the key in drawing inferences in these cases is that if treatment is administered at a different level than that of the randomization, the inferences must adjust for this correlation.^v

The focus of this essay is the second area where multilevel contexts are likely to exist. Multilevel contexts are likely to be present where the treatment consists of something that can be communicated across social ties, such as information. This type of randomized experiment need only be considered where individuals in the population of study are members of the same social structure. That is, instances, where for example, it is possible that an individual assigned to control and an individual assigned to treatment could be residents in the same household. This type of multilevel context requires a very specific design as the treatment has the potential to be indirectly administered at the group level.^{vi} This case has the potential to be extremely problematic for drawing valid causal inferences without additional adjustment. This case has the potential to violate SUTVA.

Empirically, scholars have observed social spillover, which could generate SUTVA violations in the classic experimental framework. For an example of within-household

interference, Nickerson (2008) finds higher levels of turnout in two-person households when one of the individuals is contacted to get out to vote via door-to-door canvassing in a voter mobilization experiment. In this instance, there is interference across within the households. Nickerson finds that sixty percent of the propensity to vote can be passed onto the other member of the household – a precise measurement of treatment spillover. Green, Gerber and Nickerson (2003) find within-household spillover effects from door-to-door canvassing: an increase of 5.7 percentage points for other household members among households of younger voters. In one of the earliest mobilization experiments, the Erie County study reported that while eight percent of Elmira residents had been contacted, turnout increased by ten percent, suggesting that mobilization contact was socially transmitted (Berelson, Lazarsfeld, and McPhee 1954).^{vii} Other scholars have examined spillover effects in contexts unrelated to political behavior (Besley and Case 1994; Miguel and Kremer 2004; Munchi 2004; Duflo et al. 2006). Each of these instances would generate a SUTVA violation.

In order to correctly identify instances where a multilevel structure exists to correct the experimental design to adjust for potential SUTVA violations, it is necessary to have additional information about the population of interest. There are several possibilities for identifying multilevel social structures, whose pragmatism is highly dependent upon the type of experiment being conducted. The first is to explicitly observe the level at which social interactions occur. For example, for researchers conducting experiments where they have clear and explicit knowledge of an individual's full network (for example, if the experiment was conducted via the social-networking website Facebook), then it is possible to explicitly conduct randomizations across separate components of personal networks so as to insure against spillover. However,

most experimental frameworks do not allow for this type of explicit specification of the full network structure. An alternative method for observation is for researchers to rely upon survey results where individuals self-identify their social ties. Randomization can then occur within an individual's self-identified social relationships. Survey data that solicits an individual's discussion partners – people with whom they are likely to communicate with about the treatment and thus where spillovers are likely to occur – has demonstrated that many of these discussion partners are geographically proximate (Huckfeldt, Johnson and Sprague 2004; McClurg 2006). Thus, the final method for observation of an individual's social structure is geography. Relying upon geography requires no additional acquisition of data and also allows the researchers to investigate to what extent there is spillover within an individual's physical context. The choice of each of these methods – explicit observation, survey, and geography – should in large part be based upon the type of treatment administered.

5. Designing a Multilevel Experiment

The problem generated by communication of the treatment with participants in the experiment impels us to generate an alternative experimental design that relies upon our knowledge of an individual's social structure. A multilevel experimental design, where randomization is conducted within an individual's social structure, is the best approach for establishing causal inferences. Our proposed solution is to introduce additional randomization, consistent with recommendations by Sinclair, McConnell and Green (2010). With these additional randomizations, we are able to establish via the Rubin potential outcomes framework a treatment framework that can identify both the spillovers and the direct treatment effect. These additional randomizations are key to designing a multilevel experiment.

The first component of such a design is to identify all levels at which individuals will indirectly interact. For purpose of example, we consider a population where individuals are likely to interact with each other on two levels, household and group. We require that these groups must be subsets of each other, so that each individual in the population is part of exactly one group and one household. The social structure of our population can be seen in Figure 33-1. Key to this analysis is to incorporate the full set of social structures, ranging from those where the most intimate interactions are likely to occur (in this case, households) up to those where the most casual interactions are likely to occur. In our example we model this as the group, but these group level interactions could truly be at the neighborhood level, the school-district level, or the city level. There must exist a group level at which individuals will not interact but levels where the experiment will take place and furthermore each level must be distinct – that is, an individual cannot belong to multiple groups. One way to ensure this is possible is to carefully select the experimental population; if an individual shares a household with individuals who are not part of the experiment, then it is not necessary to account for their presence in the estimation of the treatment effect, so individuals should be eligible to participate in the experimental population if their group memberships can be summarized by the appropriate randomization-levels. It is also possible to increase the number of randomizations to include, as an additional category, those individuals who belong to multiple groups, although this significantly increases the number of individuals necessary to incorporate into the experiment. Identifying these social structures and identifying the relevant experimental population is key to the design of a successful multilevel experiment.

Multilevel experimental design increases both the internal and external validity. By acknowledging the presence of these social structures, we increase the internal validity of our inferences. If it is not possible to identify these social structures, or if it is not possible to conduct an experiment that incorporates variation in these structures, then the researcher needs to think carefully about the type of inference that is drawn from the analysis of this data. Inferences drawn from experimental data which have not explicitly incorporated spillovers but where contagion of the treatment is likely may have greater problems with external validity – the same treatment, administered in a different social context, is unlikely to generate the same effect. However, if inferences are presented at the group level where the potential for spillovers is acknowledged but not incorporated into the experimental design, then in other contexts where the social structures are similar the treatment effects should be similar. Thus the caveat for the researcher should be to acknowledge the potential concern with external validity and to present the group-level treatment effects at the level above that the social interactions have occurred.

[Figure 33-1 Here]

In our example, we randomly assign groups in the population to treatment and control. Within the groups assigned to treatment, we then again repeat the random assignment, randomly assigning households to treatment and control. Finally, within the households assigned to treatment, we again repeat the random assignment, randomly assigning individuals to treatment and control. Treatment is administered at the individual level, to all members of the third stage of the randomization who are assigned to treatment. We conduct randomization at each level where we anticipate social interactions will take place. Sinclair, McConnell, and Green (2010) follow this recommendation for experimental analysis in Los Angeles and Chicago. This process allows

for the identification of both the spillover effects and the true treatment effect while removing the potential bias associated with the spillover. Suppose that the true ITT is α but that we have positive indirect treatment effects from individuals who receive treatment to other individuals within their household that are equal to β and positive indirect treatment effects from individuals who receive treatment to other individuals within their group that are equal to δ . A comparison of individuals from those assigned to each of the categories in Figure 33-1 will enable us to estimate each of these three quantities. The SUTVA violations have become a quantity of interest, allowing for inferences on interpersonal interactions.

6. Empirical Tests of Spillovers and Estimation of the Intent-to-Treat Effect

Here we provide recommendations for estimation of the ITT effect. These recommendations are not statistical corrections in the absence of design approaches, but instead are strategies for situations where the experimental design has explicitly adjusted for spillover. These strategies are simple to adopt and require minimal assumptions.

Where the experiment has incorporated the multilevel context into the experimental design, we recommend two strategies for analysis. First, we use the multiple random assignment variables to detect the presence of any social spillovers within our explicitly specified social contexts. That is, with the random assignment variables, we are able to use Hansen's J-statistic to determine if there is over-identification in the outcome behavior of the individuals in our sample by using each of the levels of random assignment as instrumental variables. This allows us to evaluate whether or not there is enough evidence to reject the null that, for example, both the individual and group level assignment variables are valid instruments. If we can reject this null hypothesis, then this suggests that in fact we do not have spillovers – a clear test which then

implies that we do not need to incorporate the multilevel structure in any additional analyses. If we cannot reject this null, then we recommend a second strategy for analysis. In this second stage, we estimate the quantities resulting from these different random assignments. That is, again suppose that the true ITT is α but that we have positive indirect treatment effects from individuals who receive treatment to other individuals within their household that are equal to β and positive indirect treatment effects from individuals who receive treatment to other individuals within their group that are equal to δ . In the estimation of the ITT, it is then necessary to incorporate parameters to separately estimate α , β and δ – that is, the effects of each level of assignment. A visual demonstration for each of these estimates is included in Table 33-1.

[Table 33-1 Goes Here]

That is, consistent with our example, we then need to estimate the effect of having been assigned to group-level treatment but not household-level or individual-level treatment compared to the group-level control. We also estimate the effect of having been assigned to household-level treatment but not individual-level treatment compared to household-level control and finally the effect of having been assigned to individual-level treatment compared to individual-level control. In Table 33-1, the difference between the two columns produces the quantity of interest. If either of the first two rows produce statistically significant effects, then the final row is likely to be biased. Thus, an estimate of the true treatment effect would subtract each of these spillover effects from the final row to estimate the true ITT.

7. Limitations and Best Practices

There are two limitations that emerge from conducting multilevel experiments. The first is simply the challenge in identifying the levels at which social interactions are likely to take

place. Given that many experiments may take advantage of geography, however, within which to conduct the experiment, this may not be a large hurdle for many types of experiments.

Geography is likely to be a weak proxy for a social environment; the best experiments would be conducted using a pre-existing social network. The second limitation that emerges is that the construction of a multilevel experiment may require a larger experimental population in order to have sufficient statistical power to identify the spillover effects. Thus the additional limitation is the loss of efficiency that emerges with the construction of multiple control groups. In many populations this is not a concern, as there are sufficiently many individuals that the addition of additional control group members does not limit the feasibility of the experiment. Yet there may be contexts within which either it is difficult to locate additional control group participants or where the requirements of additional control groups which are geographically dispersed increases the cost of conducting the experiment.

The best practice when the experiment is limited by the potential size of the control group and thus cannot be reasonably conducted on a multilevel scale is to present the group-level effects and to acknowledge the potential presence of spillovers. Spillovers can take on multiple forms; some individuals may be able to be treated multiple times, and it is possible that both individuals assigned to treatment and control may be subject to spillovers. We anticipate that most spillovers occur when the treatment group contacts the control group, but there are other kinds of situations where the treatment group may also be indirectly treated. This makes it difficult to calculate the appropriate counterfactuals for how large of a potential spillover effect could have been present with observed experimental data. Given the lack of knowledge about the direction of the bias the spillovers could take, it is impossible to ex-ante predict the effects of

potential SUTVA violations. Under certain kinds of spillovers, the estimates could in fact converge to the actual true value of the treatment effect, for example.

One final limitation of the multilevel context, albeit more applicable where there are failure-to-treat cases, is that the bias and loss of efficiency from using instrumental variables when the contacts are made uniformly across all groups is different than when the contacts are concentrated in some groups, as is the potential case in the multilevel context.

8. Recommendations for Experimental Design and Future Research

We recommend that in multilevel contexts, that randomization take place not only at the individual level but also at all appropriate social structure levels. If treatment is then administered at the individual level, it is clear how to draw inferences about spillovers, allowing us great insight into the way in which politics can be socially transmitted and the role of interpersonal influence. In this sense, SUTVA violations have become a quantity of interest.

Most importantly, however, is the detailed exposition of the randomization in multilevel contexts. To the extent that it is possible, researchers must document whether their experimental treatments are administered at a group level and acknowledge that these strategies require shifts in their estimation procedures. Furthermore, if researchers believe that their experiment is operating within a multilevel context, it is key that this be documented so that future researchers can incorporate these facts into future experiments.

While much of this essay has been written from the perspective of field experiments, it animates much of the work done in the survey world (Bowers and Stoker 2002). There are many situations where experiments are conducted within multilevel settings. Voter mobilization experiments, where treatment and randomization occur at the level of the individual, are clearly

prey to potential SUTVA violations. Within existing social and political organizations, there may be hierarchical or geographically distributed groups that are also subject to potential SUTVA violations, such as statewide organizations with local chapters and individual members. Clearly these concerns are relevant for experiments conducted on college campuses, where individual participants may be residents in the same dorm, for example. Experiments conducted using the multilevel randomization design will allow social science to develop an empirical knowledge-base for how much spillover actually does occur and how much potential bias there could be. At this point our collective knowledge regarding about spillover is fairly empty, and we do not know under what conditions spillovers are likely to occur.

Experimenters need to be sure to consider failure-to-treat situations and the ways in which they may further complicate these analyses. This essay has not dealt explicitly with failure-to-treat, but these instances require additional assumptions under which to draw inferences. In particular, many kinds of analyses use the random assignment variable as an instrumental variable in order to estimate the treatment-on-treated effect. This approach is not appropriate in cases where there is social transmission of treatment within the experimental population as the assignment variable fails to capture the indirect treatment.

Multilevel experiments have the potential to yield great insights into the ways in which humans interact; with careful experimental design, the SUTVA violations have the potential to open up new avenues of research previously reliant upon heroic assumptions. Each additional randomization does not add to the technical difficulty of implementing the experiment, as it is still possible for the experimental design to include the same number of participants assigned to be administered the treatment. It is the addition of the new control groups that allows for the

identification of the spillover effects. Researchers should be aware of the statistical power necessary to detect spillover categories, however, and should attempt randomizations so that future meta-analysis will allow us an understanding of spillovers, even if individual studies are inconclusive.

This methodological improvement has the potential to encourage different kinds of inferences in randomized field experiments. This is also a technique that allows the discovery of supportive evidence for individuals who study network analysis via survey data to understand social structure. This method can also be extended to include additional randomizations to study spillover in many directions – for example, we could also include a category where we compared individuals assigned to treatment who were paired with control to individuals assigned to treatment who were paired with treatment to individuals assigned to control – this would allow us to see if in fact the pairing of treated-with-treated would increase the effect of the treatment as well. SUTVA violations have the potential to be extremely interesting quantities of interest. As we develop new and interesting ways to measure spillovers, these quantities will allow us to inform which types of theories are most applicable in the social transmission of politics. We do not yet know whether or not the instigation of political behavior is due to generated conversations, heightened interest, or persuasion. The measurement of spillover will offer one set of illustrations for where our theories should focus.

References

Adato, Michelle, David Coady, and Marie Ruel. 2000. “An Operations Evaluation of PROGRESA from the Perspective of Beneficiaries, Promotoras, School Directors and Health Staff.” Washington, DC: International Food Policy Research Institute.

- Arceneaux, Kevin. 2005. "Using Cluster Randomized Field Experiments to Study Voting Behavior." *The Annals of the American Academy of Political and Social Science* 601: 169-79.
- Arceneaux, Kevin, and David W. Nickerson. 2009. "Modeling Certainty with Clustered Data: A Comparison of Methods." *Political Analysis* 17: 177-90.
- Besley, Timothy, and Anne Case. 1993. "Modeling Technology Adoption in Developing Countries." *American Economic Review* 83: 396-402.
- Berelson, Bernard, Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting*. Chicago: University of Chicago Press.
- Brock, William A., and Steven N. Durlauf. 1999. "A Formal Model of Theory Choice in Science." *Economic Theory* 14: 113-30.
- Bowers, Jack, and Laura Stoker. 2002. "Designing Multilevel Studies: Sampling Voters and Electoral Contexts." *Electoral Studies* 21: 235-67.
- Cacioppo, John T., James H. Fowler, and Nicholas A. Christakis. 2009. "Alone in the Crowd: The Structure and Spread of Loneliness in a Large Social Network." *Journal of Personality and Social Psychology* 97: 977-91.
- Coleman, James S., Elihu Katz, and Herbert Menzel. 1966. *Medical Innovation*. Indianapolis: Bobbs-Merrill Press.
- Conley, Timothy G., and Christopher R. Udry, 2010. "Learning About a New Technology: Pineapple in Ghana." *American Economic Review* 100: 35-69.
- DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel. 2003. "Persuasion Bias, Social Influence and Unidimensional Opinions." *Quarterly Journal of Economics* 118: 909-68.
- Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2006. "Understanding Technology Adoption: Fertilizer in Western Kenya." Unpublished manuscript, Massachusetts Institute of Technology.
- Fowler, James H., and Nicholas A. Christakis. 2008. "Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study." *British Medical Journal* 337: a2338.
- Gerber, Alan S., and Donald P. Green. 2008. *Get Out the Vote!* 2nd ed. Washington DC: Brookings Institution Press.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 94: 653-63.

- Green, Donald P., Alan S. Gerber, and David W. Nickerson. 2003. "Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments." *Journal of Politics* 65: 1083-96.
- Green, Donald P., and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16: 138-52.
- Hansen, Ben B., and Jake Bowers. 2008. "Attributing Effects to a Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association* 104(487): 873-85.
- Huckfeldt, Robert, Paul E. Johnson, and John Sprague. 2004. *Political Disagreement* Cambridge: Cambridge University Press.
- Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24: 29-53.
- King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha Maria Tellez-Rojo, Juan Eugenio Hernandez Avila, Mauricio Hernandez Avila, and Hector Hernandez Llamas. 2007. "A 'Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program." *Journal of Policy Analysis and Management* 26: 479-509.
- Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114: 497-532.
- Lazarsfeld, Paul, Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice*. New York: Columbia University Press.
- Manski, Charles. 1993. "Identification of Exogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60: 531-42.
- McClurg, Scott. 2006. "Political Disagreement in Context: The Conditional Effect of Neighborhood Context, Discussion, and Disagreement on Electoral Participation." *Political Behavior* 28: 349-66.
- Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72: 159-217.
- Munshi, Kaivan. 2004. "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution." *Journal of Development Economics* 73: 185-213.
- Nair, Harikesh, Puneet Manchanda, and Tulikaa Bhatia. 2008. "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders." SSRN working paper 937021.

- Nickerson, David. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49-57.
- Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 57 (371): 591-3.
- Rubin, Donald B. 1986. "Which Ifs Have Causal Answers? Discussion of Holland's 'Statistics and Causal Inferences'." *Journal of the American Statistical Association* 81: 961-2.
- Rubin, Donald B. 1990. "Formal Modes of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference* 25: 279-92.
- Panagopoulos, Costas, and Donald P. Green. 2006. "The Impact of Radio Advertisements on Voter Turnout and Electoral Competition." Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL.
- Sinclair, Betsy and, Brian Rogers. 2009. "Political Networks: The Relationship Between Candidate Platform Positions and Constituency Communication Structures." Unpublished manuscript, University of Chicago.
- Sinclair, Betsy, Margaret McConnell, and Melissa R. Michelson. 2010. "Strangers vs Neighbors: The Efficacy of Grassroots Voter Mobilization." Unpublished manuscript, University of Chicago.
- Sinclair, Betsy, Margaret A. McConnell, and Donald P. Green. 2010. "Detecting Spillover in Social Networks: Design and Analysis of Multilevel Experiments." Unpublished manuscript, University of Chicago.

Figure 33-1: Multilevel Experiment Design

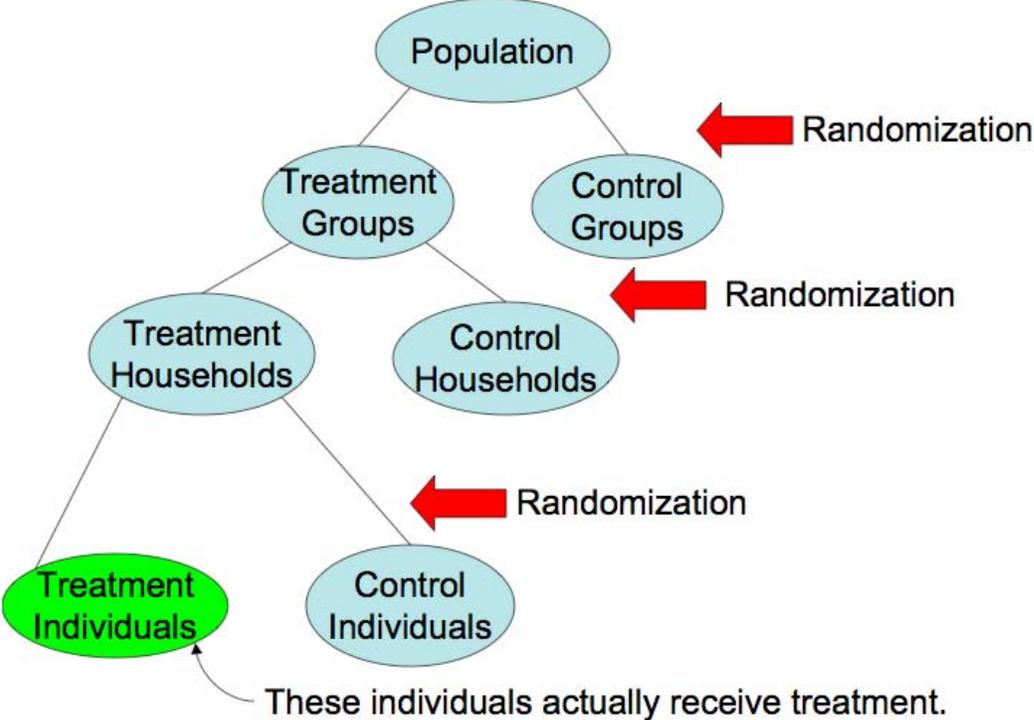


Table 33-1: ITT Effects

Assignment (Group, Household, Individual)	Assignment (Group, Household, Individual)	Quantity of Interest
Control, Control, Control	Treatment, Control, Control	Group-level Spillover
Treatment, Control, Control	Treatment, Treatment, Control	Household-level Spillover
Treatment, Treatment, Control	Treatment, Treatment, Treatment	Treatment Effect (Potentially Biased)

ⁱ The author would like to thank Donald Green, Jamie Druckman, Thomas Leeper, Jon Rogowski, John Balz, Alex Bass, Jaira Harrington, and participants of the West Coast Experimental Political Science Conference for their helpful comments in improving this manuscript.

ⁱⁱ As oftentimes not all individuals who are assigned to receive the treatment will actually be treated, there is also a loss of efficiency if the contact rates differ greatly across groups.

ⁱⁱⁱ The second part of SUTVA assumes that the treatment is the same for each unit: “SUTVA is violated when, for example there exist unrepresented versions of treatments or inference between units” (Rubin 1986, 961).

^{iv} Failure-to-treat problems present challenges for analysis, some of which can be mitigated via randomization inference (Hansen and Bowers 2008).

^v If there exist inconsistencies between the random assignment and the administration of the treatment, we recommend two strategies for analysis. Suppose that an experiment has been conducted where the randomization occurred at the individual level but the treatment was administered at the group level. In this case, we first recommend clustering the standard errors at the group level when estimating the ITT. This clustering explicitly acknowledges the correlation that is likely to exist within the group as a result of the administration of the treatment and adjusts for the lack of independence of all observations within the group (Green and Vavreck 2008; Arceneaux and Nickerson 2009). Note, however, that this adjustment is not sufficient to account for the potential biases that have occurred as a result of social interactions. Correlation in the standard errors does not account for the possibility that individuals who are assigned to treatment, for example, may have been indirectly treated multiple times from other individuals assigned to treatment nor the possibility that individuals assigned to control may have been indirectly treated from individuals assigned to treatment. Our second recommendation, if there are a sufficient number of group-level observations, is to conduct analysis either via a hierarchical linear model or via meta-analysis. Meta-analysis, for example, does not require homoskedasticity across groups, which is a necessary assumption with the inclusion of group-level fixed-effects, which assumes that there is a group-specific effect (Gerber, Green, and Larimer 2008; Sinclair, McConnell and Michelson 2010). While this is more often a concern when there are failure-to-treat instances, there is still likely to be group-level variation that is not properly accounted for via fixed-effects.

^{vi} If we conduct both our randomization at the group level and our analysis on the group level, then this case requires no additional shifts in experimental design and in fact is eligible for the block-group randomization (King et al 2007).

^{vii} The assumptions about social transmission of political information and positive and significant effects of peers on individual political behavior date back to the Erie County Study of 1940 and the Elmira Community Study of 1948, some of the earliest quantitative work in political science (Lazarsfeld, Berelson and Gaudet 1948; Berelson et al. 1954).

34. Analyzing the Downstream Effects of Randomized Experiments

Rachel Milstein Sondheimerⁱ

The work in this volume provides profound evidence of the value of randomized experimentation in political science research. Laboratory and field experiments can open up new fields for exploration, shed new light on old debates, and answer questions previously thought to be intractable. While acknowledgment of the value of experimentation in political science is becoming more commonplace, significant criticisms remain. The oft-repeated shortcomings of experimental research tend to center on the practical and ethical limitations of randomized interventions. I begin this chapter by detailing some of these criticisms and then explore one means of extending the value of randomized interventions beyond their original intent to ameliorate some of these same perceived limitations.

One of the most prominent critiques of this genre is that randomized experiments tend to be overly narrow in scope in terms of timeframe and subject matter and high in cost. While short term experiments may incur costs similar to observational research, they often focus on a single or just a few variations in an independent variable, seemingly limiting their applicability to a breadth of topics that a survey could cover. The high cost associated with long term data collection and the necessity of maintaining contact with the subjects involved impedes the likelihood of gathering information on long-term outcomes. There are also few incentives to conduct interventions in which the impacts may only be determined years down the road. Such studies are not amenable to dissertation research unless graduate students extend their tours of duty even longer nor do they suit junior faculty trying to build their publication records. The

isolation of long-term effects necessitates long-term planning, long-term maintenance, and long-term funding, all of which tend to be in short supply for many researchers.

The second practical critique of randomized experiments stems from the difficulty of such studies to capture the influence of variables of interest to many political scientists. We cannot purposefully alter political culture to see its effect on democratization nor can we alter the family background of a set of individuals to investigate its influence on political socialization. Perhaps, some might argue, experiments are only useful in studying narrow topics of political behavior in highly controlled settings, mitigating their applicability to broad real world questions.

Randomized interventions also face ethical limitations and critiques. Prohibitions on potentially harmful interventions need little discussion here. We must consider experiments testing political variables of interest that may fall into an ethical grey area. There are more possibilities of these than one might initially assume. As in medical testing, controversy exists concerning the practice of denying a known good to a subject as an evaluative technique. Subjecting participants to varying treatment regimens might indicate an underlying assumption that one such regimen is “better” than another. If this is indeed the case, practitioners are left open to the charge that they are somehow adversely affecting one group of people over another by failing to provide them with the “better” treatment option. For example, intentionally boosting some individuals’ levels of educational attainment in comparison to others in an effort to examine the ramifications of additional years of schooling on political and behavioral outcomes seems unethical given the widespread belief in the positive externalities associated with schooling.

Also ethically dubious is the notion that some interventions could have the long-term consequence of influencing social outcomes. Artificially enhancing randomly selected candidates' campaign coffers to test the effects of money on electoral success could affect electoral outcomes and the drafting and passage of legislation. Randomly assigning differential levels of lobbying on a particular bill could sway the drafting and passage of legislation. Testing the effectiveness of governmental structures through random assignment of different types of constitutions to newly formed states could result in internal strife and international instability.

The list of interventions that, although useful for research and pedagogical purposes, would be simply impractical or unethical is seemingly endless while the universe of interventions that are both feasible and useful appears somewhat limited in scope. Does this mean that experiments will only be useful in answering questions where it is practical and ethical to manipulate variables of interest? The answer is no for myriad reasons discussed in this volume and elsewhere. Here I focus on just one: we can expand the utility of initial interventions beyond their original intent through examination of the long-term, and sometimes unforeseen, consequences of randomized interventions. Using downstream analysis, political scientists can leverage the power of one randomized intervention to examine a host of causal relationships that they might otherwise have never been able to study through means other than observational analysis. While political scientists may not randomly assign some politicians to receive more money than others, some other intervention, natural or intended, may produce such an outcome. Researchers can exploit this variation achieved through random or near-random assignment to examine the consequences of this resulting discrepancy in campaign finances on other outcomes like electoral success.

In the next section, I further define and detail the key assumptions associated with downstream analysis of randomized experiments. I then highlight research that uses downstream analysis and outline some potential research areas that may benefit from this outgrowth of randomized interventions. I conclude with a discussion of the methodological, practical, and ethical challenges posed by downstream analysis and offer some suggestions for overcoming these difficulties.

1. Extending the Benefits of Randomized Experiments

For most researchers, a randomized experiment ends once the treatment is applied and the outcome measures are collected. This seeming finality belies the possibility that many, although perhaps not all, treatments produce ramifications that extend well beyond the original timeframe and purpose of the study. The unintended consequence of these interventions is that they often provide an exogenous shock to an outcome not normally amenable to classic randomization. This allows researchers to extend the positive externalities of experimentation to different fields of study and interest; this is especially useful for fields in which experiments are often untenable for practical and ethical reasons. Green and Gerber (2002) define these downstream benefits as “knowledge acquired when one examines the indirect effects of a randomized experimental intervention” (394).

Analyzing the second order consequences of randomized experiments can help justify some of the perceived limitations often associated with such endeavors. Downstream analysis opens up the possibility of extending a narrowly construed topic or outcome to apply to a broader range of fields. Experiments on the utility of direct mail as a mobilization tool can become investigations into the causes of habitual voting (Gerber, Green, and Shacher 2003). Random

assignment of a local political office reserved for women can become investigations into the long-term effects of breaking the electoral glass ceiling for women (Bhavnani 2009). This type of application is best seen through an example involving the effects of schooling on participation.

The positive effect of formal schooling on political and civic participation is widely reported in observational research (e.g. Berelson, Lazarsfeld, and McPhee 1954; Campbell et al. 1960; Converse 1972; Delli Carpini and Keeter 1996; Miller and Shanks 1996; Nie, Junn, and Stehlik-Barry 1996; Verba and Nie 1972; Wolfinger and Rosenstone 1980), but some recent research questions the underpinnings of this relationship (Tenn 2007; Kam and Palmer 2008; Sondheimer and Green 2010). Many scholars tend to view this link as causal despite failure to clarify the causal mechanisms that produce this strong relationship and an inability to isolate educational attainment from unobserved factors like cognitive ability and family background. Observational analysis of the topic is faced with diminishing returns; random assignment of varying levels of educational attainment to otherwise similar subjects would advance current knowledge in this field but is impossible for moral and practical reasons. It is possible to still leverage the benefits of experimental methods in this field. Randomized trials to test different education techniques and policies often produce differential levels of schooling between treatment and control cohorts. The second order effects of these interventions, if observational analysis is correct, should also produce differential rates of political participation as a result of the boost given to the average years of schooling among the treatment cohort. While such studies were designed to examine the value of programs like public preschool and small class size, political scientists can use the exogenous shocks to years of schooling brought about by

randomized interventions to examine the effects of high school graduation on voting behavior and other political and civic participation outcomes.

In the next section I detail how we can estimate such effects continuing with the example of the downstream possibilities for using randomized educational interventions to untangle the causal relationship between educational attainment and political participation.

2. A Model and Estimation of Downstream Effects of Randomized Experiments

Estimation of the Local Average Treatment Effect

The explication of and assumptions underlying the Rubin Causal Model and the use of instrumental variables (IV) estimation to overcome unobserved heterogeneity is laid out by Sovey and Green (2009) and adapted here for the downstream analysis of experiments. The IV estimator consists of a two-equation model and is written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \lambda_1 Q_{1i} + \lambda_2 Q_{2i} + \dots + \lambda_K Q_{Ki} + \Lambda + u_i \quad (1)$$

$$X_i = \gamma_0 + \gamma_1 Z_i + \delta_1 Q_{1i} + \delta_2 Q_{2i} + \dots + \delta_K Q_{Ki} + \Lambda + e_i \quad (2)$$

In the first equation, we assume that the individual voter turnout (Y_i) is the dependent variable, educational attainment (X_i) is the regressor of interest, $Q_{1i}, Q_{2i}, \dots, Q_{Ki}$ are covariates, and u_i is an unobserved disturbance term. The second equation holds that the endogenous regressor (X_i), educational attainment in this case, is a linear function of an instrumental variable (Z_i), the covariates, and an unobserved disturbance term (e_i). Random assignment of Z allows us to isolate exogenous variation in X , that is, the piece of X that is independent of other factors that might influence Y , overcoming concerns of omitted variable bias in our estimation of the effect of X on Y .

In order to see what IV is estimating, it is useful to introduce the Rubin Causal Model as presented by Angrist, Imbens, and Rubin (1996). I apply this model to our running example of using a randomized educational intervention to isolate the effects of graduation from high school on individual voter turnout. Here, we make use of three dichotomous variables coded as either 0 or 1: 1) individual assignment to the treatment or control cohort of the intervention (Z_i), 2) whether or not an individual graduated from high school (X_i), and 3) whether or not an individual voted in a given election (Y_i).

The IV estimator above appears to calculate the causal effect of X on Y, but as we will see below, closer examination shows that the estimator isolates the causal effect for X on Y for those subjects affected by the initial intervention – subjects whose X outcome changed due to the introduction of Z. Angrist and Krueger (2001) define this estimand as the *local average treatment effect* (LATE).ⁱⁱ Speaking in terms of LATEs alerts us to the fact that IV reveals the causal influence of the intervention for a subset of the population. We need not assume that the initial intervention affected all treatment subjects in the same way or that the intervention determined the outcome for all subjects in the treatment cohort.

The first step in estimating the LATE model is to conceptualize the effects of randomized assignment in the intervention on educational outcomes. In discussing estimation techniques for randomized experiments, Imbens and Rubin (1997) group subjects into four categories based on how the treatments they receive depend on the treatments to which they are assigned. In the case of downstream analysis of randomized interventions, the concept of compliance is applied somewhat differently. In the case of downstream analysis of educational interventions, we define compliance in terms of whether people graduate high school in response to being assigned to the

treatment group. In the context of a downstream analysis, Imbens and Rubin's four groups are as follows:

- 1) *Compliers* graduate from high school if and only if they are assigned to the treatment group ($z_i = 1, x_i = 1$) or ($z_i = 0, x_i = 0$);
- 2) *Never-takers* do not graduate from high school no matter the group to which they are assigned ($z_i = 0, x_i = 0$) or ($z_i = 1, x_i = 0$);
- 3) *Always-takers* graduate from high school no matter the group to which they are assigned ($z_i = 0, x_i = 1$) or ($z_i = 1, x_i = 1$); and
- 4) *Defiers* graduate from high school if and only if they are assigned to the control group ($z_i = 0, x_i = 1$) or ($z_i = 1, x_i = 0$).

Note that we cannot look at a given individual and classify her into one of these mutually exclusive groups because we are limited to only observing one possible outcome per individual. In other words, if a subject assigned to the treatment group graduates from high school, we cannot discern whether she is a Complier who could have only graduated if assigned to the treatment group or an Always-taker who would have graduated regardless of random assignment in the intervention.

In order to draw causal inferences from an instrumental variables regression based on downstream data, one must invoke a series of assumptions. The first is independence: potential outcomes of the dependent variable must be independent of the experimental group to which a person is assigned. This criterion is satisfied by random assignment.

The second is the exclusion restriction. The instrumental variable, Z , only influences Y through its influence on X . In other words, Y changes because of variation in X and not because of something else. In this case, we assume that random assignment in the original experiment has no influence on voting outcomes aside from that mediated by graduation from high school. Evaluating whether or not a given instrument satisfies this condition rests on understanding the nature of the relationship between an observed variable (Z_i) and an unobserved variable (u_i).

Theoretically, random assignment can ensure the statistical independence of Z and u . However the nature of the randomized intervention may lead to violations of the exclusion restriction – for example, subjects who realize that they are being studied may be influenced by the simple fact that they are part of an experimental treatment group. When evaluating the exclusion restriction, the researcher must consider causal pathways that might produce changes in Y through pathways other than X .

Third, we invoke the stable unit treatment value assumption (SUTVA), which holds that an individual subject's response is only a function of her assignment and educational attainment and is not affected by the assignment or outcomes of other subjects. Fourth, we assume monotonicity, that is, the absence of Defiers. Finally, we must assume that Z exerts some causal influence on X . Low correlation between Z and X can lead to small sample bias, a problem discussed in Section 4.

We are primarily interested in the relationship between graduation from high school and its subsequent effect on the likelihood of voting. As such, we can say that the dependent variable for each subject i can be expressed as either y_{i1} if the individual graduated from high school or y_{i0} if the subject did not graduate from high school. The mechanics of downstream analysis can be illustrated by classifying subjects' outcome measures based on assignment and response to the educational intervention, producing four categories of subjects:

- 1) Individuals who voted regardless of whether or not they graduated from high school ($y_{i1}=1, y_{i0}=1$);
- 2) Individuals who voted if they graduated from high school but did not vote if they did not graduate from high school ($y_{i1}=1, y_{i0}=0$);
- 3) Individuals who did not vote if they graduated from high school but did vote if they did not graduate from high school ($y_{i1}=0, y_{i0}=1$); and
- 4) Individuals who did not vote regardless of whether or not they graduated from high school ($y_{i1}=0, y_{i0}=0$).

As detailed in Table 34-1, based on the four compliance possibilities and the four classifications of outcome measures mediated by educational attainment, we can create a total of 16 mutually exclusive groups into which subjects may fall. The total subject population share of each group is expressed as π_j such that $\sum_j^{16} \pi_j = 1$. To estimate the causal effect between our variables of interest, we must further assume monotonicity (Angrist, Imbens, and Rubin 1996, 444-55) – that the treatment can only increase the possibility of graduating from high school. This assumption holds that there are no Defiers or that $p_{13}=p_{14}=p_{15}=p_{16}=0$.

Randomization allows us to leverage the variation brought about by an exogenous shock to an otherwise endogenous measure to isolate the influence of that measure on another outcome of interest. As such, we can only identify the influence of schooling on the outcomes of a limited number of the groups of subjects classified in Table 34-1. We cannot evaluate the effects of schooling on the voting behavior of the Never-takers who will not graduate from high school regardless of assignment or the Always-takers who will graduate from high school regardless of assignment. These outcomes are determined by factors that might be endogenous to our voting model. To isolate the effects of educational attainment on voting, we can only estimate the schooling's effects on the outcomes of the Compliers who graduate from high school if assigned to the treatment but do not graduate if assigned to the control. This *complier average causal effect* (CACE) is expressed as:

$$E[Y_{11} - Y_{01} | D = \text{Compliers}] = \frac{\pi_6 - \pi_7}{\pi_2 + \pi_3 + \pi_7 + \pi_8} \quad (4)$$

A randomized experiment and application of the Rubin Causal Model allows us to estimate the LATE given four previously discussed assumptions: 1) SUTVA, 2) the exclusion

restriction, 3) an observed effect of assignment on graduation rates, and 4) monotonicity. As we will see, if these assumptions are met, LATE estimation will also provide us with a consistent estimator of the CACE.

The first step is to estimate the voting rates, V , for the treatment and control groups. As the number of observations in the control group approaches infinity, the observed voting rate in

the assigned control group ($\hat{V}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} y_{i0}$) can be expressed as

$$p \lim_{N_c \rightarrow \infty} \hat{V}_c = \pi_3 + \pi_4 + \pi_7 + \pi_8 + \pi_{10} + \pi_{12}. \quad (5)$$

Similarly, the observed voting rate in the assigned treatment group can be expressed as

$$p \lim_{N_t \rightarrow \infty} \hat{V}_t = \pi_3 + \pi_4 + \pi_6 + \pi_8 + \pi_{10} + \pi_{12}. \quad (6)$$

Next, we can find a consistent estimator of the proportion of Compliers (α), subjects who only graduated from high school if and only if assigned to the treatment regimen, by subtracting the graduation rate of the control group from the graduation rate of the treatment group:

$$\begin{aligned} p \lim_{N \rightarrow \infty} \hat{\alpha} &= p \lim_{N_t \rightarrow \infty} (\hat{\pi}_5 + \hat{\pi}_6 + \hat{\pi}_7 + \hat{\pi}_8 + \hat{\pi}_9 + \hat{\pi}_{10} + \hat{\pi}_{11} + \hat{\pi}_{12}) - p \lim_{N_c \rightarrow \infty} (\hat{\pi}_9 + \hat{\pi}_{10} + \hat{\pi}_{11} + \hat{\pi}_{12}) \\ &= \pi_5 + \pi_6 + \pi_7 + \pi_8. \end{aligned} \quad (7)$$

Finally, we can combine the estimators presented in equations 5, 6, and 7 to provide a consistent LATE estimate that is the same as the CACE presented above:

$$p \lim_{N \rightarrow \infty} \frac{\hat{V}_t - \hat{V}_c}{\hat{\alpha}} = \frac{\pi_6 - \pi_7}{\pi_5 + \pi_6 + \pi_7 + \pi_8}. \quad (8)$$

The estimator for the LATE also estimates the CACE, the effect of educational attainment among those who would not have graduated save for the intervention (the Compliers).

In conclusion, the IV estimator estimates CACE. We are not estimating the effect of high school for everybody, just the effect for those who are influenced by a high school inducing

program. Of course one can generalize beyond compliers but this must be done cautiously and through replication unless one is willing to make strong assumptions. The downstream analysis of an individual intervention might be best interpreted as a LATE estimate, but, if a persistent pattern emerges through the downstream analysis of multiple interventions with different target populations, we can then begin to extrapolate the results to a more broad population.

3. Downstream Analysis in Practice

“Downstream experimentation” was a term originally coined by Green and Gerber in 2002. The concept of using existing randomized and natural experiments to examine second order effects of interventions was slow to build due, in large part, to the relative dearth of suitable experiments in political science. Now that experiments are becoming more widespread and prominent in political science literature, scholars are beginning to cull the growing number of interventions to test theories seemingly untestable through traditional randomization. In this section, I touch on just a few such examples and offer avenues for further exploration using similar first order experiments. This discussion is meant to encourage those interested to seek out these and other works to explore the possibilities of downstream analysis.

As I discussed above, an interesting stockpile of experiments well-suited for downstream analysis are interventions designed to test public policy innovations, programs in education in particular. While a key variable of interest to many is the influence of education on a wide array of political and social outcomes, one’s level of educational attainment is itself a product of numerous factors, potentially impeding our ability to isolate schooling’s causal effect through standard observational analysis (Rosenzweig and Wolpin 2000). At the same time, it is nearly impossible and potentially unethical to randomize the educational attainment of individuals or

groups to gauge its effect. Sondheimer and Green (2010) examine two randomized educational interventions, the High/Scope Perry Preschool project examining the value of preschool in the 1960s and the Student Teacher Achievement Ratio program testing the value of small classes in Tennessee in the 1980s, in which the treatment groups witnessed an increase in years of schooling in comparison control group. They used these differential levels of schooling produced by the randomized interventions to isolate the effects of educational attainment on likelihood of voting, confirming the strong effect often produced in conventional observational analysis. In addition to examinations of voter turnout, downstream analysis of educational interventions can isolate the effects of years of schooling on a range of outcomes including views on government, party affiliation, civic engagement and social networking.

As discussed throughout this volume (see, in particular, Michelson and Nickerson's chapter), experimentation is proliferating in the field of voter mobilization. Scores of researchers conduct randomized trials to estimate the effectiveness of different techniques aimed at getting people to the polls. In doing so, these studies create the opportunity for subsequent research on the second order effects of an individual casting a ballot when she would not have done so absent some form of intervention. Gerber et al. (2003) use an experiment testing the effects of face-to-face canvassing and direct mail on turnout in a local election in 1998 to examine whether voting in one election increases the likelihood of voting in another election. Previous observational research on the persistence of voting over time is unable to distinguish between the unobserved causes of an individual voting in the first place from the potential of habit formation. Gerber et al. find that the exogenous shock to voting produced within the treatment group by the initial mobilization intervention in 1998 endured somewhat in the 1999 election, indicating a

persistence pattern independent of other unobserved causes of voting. Further extension of mobilization experiments could test the second order effects of casting a ballot on attitudes (e.g. internal and external efficacy), political knowledge and the likelihood of spillover into other forms of participation.

Laboratory settings and survey manipulation offer fruitful ground for downstream analysis of randomized experiments. Holbrook discusses experiments, predominantly performed in laboratories or lab-like settings or in survey research, that seek to measure either attitude formation or change as the dependent variable (see Holbrook's chapter in this volume). As she notes, understanding the processes of attitude formation and change is central to research in political science because such attitudes inform democratic decision making at all levels of government and politics. Experiments seeking to understand the causes of attitude formation and change can use downstream analysis to examine the second-order effects of these exogenously induced variations on subsequent beliefs, opinions, and behaviors. For example, Peffley and Hurwitz (2007) use a survey experiment to test the effect of different types of argument framing on support for capital punishment. Individual variation in support for capital punishment brought about by this random assignment could be used to test how views on specific issues influence attitudes on other political and social issues, the purpose and role of government generally, and evaluations of electoral candidates.

Natural experiments provide further opportunity for downstream examination of seemingly intractable questions. In this vein, scholars examine the second- and third-order effects caused by naturally occurring random or near-random assignment into treatment and experimental groups. Looking to legislative research, Kellerman and Shepsle (2009) use the

lottery assignment of seniority to multiple new members of Congressional committees to explore the effects of future seniority on career outcomes like passage of sponsored bills in and out of the jurisdiction of the initially assigned committee and reelection outcomes. Bhavnani (2009) uses the random assignment of seats reserved for women in local legislative bodies in India to examine whether the existence of a reserved seat, once removed, increases the likelihood of women being elected to this same seat in the future. The original intent of the reservation system is to increase the proportion of women elected to local office; Bhavnani exploits the random rotation of these reserved seats to examine the “next-election” effects of this program once the reserved status of a given seat is removed and the local election is again open to male candidates. Bhavnani focuses his analysis to the subsequent elections in these treatment and control wards but one could imagine using this type of natural randomization process to examine a host of second-order effects of the forced election of female candidates ranging from changes in attitudes towards women to shifts in the distribution of public goods in these wards.

As the education experiments indicate, randomized policy interventions used to test new and innovative ideas provide fascinating opportunities to test the long-term ramifications of changes in individual and community level factors on a variety of outcomes. Quasi-experiments and natural experiments that create as-if random placement into treatment and control groups provide additional prospects. Many programs use lotteries to determine which individuals or groups will receive certain new benefits or opportunities. Comparing these recipients to non-recipients or those placed on a waiting list evokes assumptions and opportunities similar to those of randomized experiments (Green and Gerber 2002). Numerous scholars in a range of disciplines (e.g. Katz, Kling, and Liebman 2001; Ludwig, Duncan, and Hirschfield 2001;

Leventhal and Brooks-Gunn 2003; Sanbonmatsu et al. 2006) have examined the Moving to Opportunity (MTO) program that relocates participants from high poverty public housing to private housing in either near-poor or non-poor neighborhoods.

Political scientists can and have benefited from this as-if random experiment as well. Observational research on the effects of neighborhoods on political and social outcomes suffers from self-selection of subjects into neighborhoods. The factors that determine where one lives are also likely to influence one's proclivities toward politics, social networking tendencies, and other facets of political and social behavior. Downstream research into a randomly assigned residential voucher program allows political scientists the opportunity to parse out the effects of neighborhood context from individual level factors that help determine one's choice of residential locale. Political scientists are just beginning to leverage this large social experiment to address such questions. Gay's (2010) preliminary work on the MTO allows her to examine how an exogenous shock to one's residential environment affects political engagement in the form of voter registration and turnout. She finds that subjects who received vouchers to move to new neighborhoods voted at lower rates than those who did not receive vouchers, possibly due to the disruption of social networks that may result from relocation. Future research in this vein could leverage this and similar interventions to examine how exogenously induced variations in the residency patterns of individuals and families affects social networks, communality, civic engagement and other variables of interest to political scientists.

Other opportunities for downstream analysis of interventions exist well beyond those discussed here. As this short review shows, however, finding these downstream possibilities often entails looking beyond literature in political science to other fields of study.

4. Challenges

Downstream analysis of existing randomized interventions provides exciting possibilities for researchers interested in isolating causation. We can expand the universe of relationships capable of study using assumptions generated by randomization and experimental analysis. While downstream analyses can be used to overcome some of the limitations of randomized interventions, they do pose their own set of challenges. In this section, I discuss the methodological, practical, and ethical challenges faced by those who wish to perform downstream analysis.

Methodological Challenges

Two key impediments to downstream analysis of randomized experiments stem from two of the three conditions for instruments to maintain the conditions for instrumental variable estimation, specifically finding instruments that meet the exclusion restriction and provide a strong relationship between assignment and the independent variable of interest. First, a suitable instrument must meet the exclusion restriction in that it should be exogenous to the regression of interest not exerting an independent influence on the dependent variable. Randomization of subjects into treatment and control cohorts is not sufficient to meet the exclusion restriction because the specific nature of the treatment regimen can still violate this condition. Returning to our investigation of educational interventions as a means of isolating the influence of educational attainment on political participation, we can imagine myriad situations in which the intervention may influence participation, independent of the effects of schooling. If the intervention works through a mentoring program, mentors may discuss politics and the importance of political involvement with subjects in the treatment group, increasing the likelihood of voting later in life.

Similarly, it is possible that an intervention intended to boost the educational attainment of a treatment group also influences the family dynamics of the subjects' home lives. If this occurs and the exclusion restriction is violated, researchers will be unable to isolate the causal influence of variations in years of schooling on political participation independent of family background.

Another example of the potential violation of the exclusion restriction exists concerning Gerber et al. (2003) work on voting and habit formation. Recall that Gerber et al. use a randomized mobilization intervention to find that, all else equal, voting in one election increases the likelihood of voting in subsequent elections indicating that electoral participation is habit forming. This result hinges on the assumption that no other changes took place for individuals in the initial experiment other than assignment to the control or treatment group. A violation of the exclusion restriction would occur if the outcome induced due to assignment influenced other factors thought to influence subsequent voting. For example, if political parties and other campaign operatives tend to reach out to those that already seem likely to vote (Rosenstone and Hansen 1993), voting in the first election may increase the likelihood of being subject to increased mobilization in subsequent elections, increasing the likelihood of voting in subsequent elections. If this occurs and the exclusion restriction is violated, Gerber et al. would still be correct in arguing that voting in one election increases the likelihood of voting in subsequent elections but this result may not be due to habit but to another factors such as outreach to likely voters.

There is no way to test whether or not this condition is met. Instead we must make theoretical assumptions about the relationship between the instrument and potential unobserved causes of the variation in the dependent variable. In-depth knowledge of the original intervention

and its possible effects on numerous factors relating to the outcome variable of interest is necessary to uncover possible violations.

The second methodological difficulty posed by researchers wishing to perform downstream analysis relates to the second condition necessary for consistent estimates using instrumental variables – that the instrument must be correlated with the endogenous independent variable. In downstream analysis, meeting this condition entails finding randomized experiments that “worked” in so far as random assignment produced variation in the treatment group as compared to the control. As Sovey and Green (2009) and Wooldridge (2009) discuss, use of a weak instrument – an exogenous regressor with a low partial correlation between Z_i and X_i – can cause biased IV estimates in finite samples. Fortunately, weak instrumentation can be diagnosed using a first-stage F-statistic as outlined by Stock and Watson (2007).

Finding a suitable instrument to meet this condition is a delicate matter because while researchers need to find an experiment with a strong result, we also need to be wary of interventions with too strong results, an indication that something might be awry. As Green and Gerber (2002, 394-402) discuss, such results might indicate a failure of randomization to create similar groups prior to the intervention. Randomized experiments with large numbers of subjects and replicated results provide the best potential pool of studies for downstream analysis.

Practical Challenges

The most straightforward challenge of downstream analysis of randomized experiments is that, in most circumstances, the analyst has no control over the course of the initial intervention. There are some opportunities to improve on the analysis of an original intervention, but, in general, the internal validity of downstream analysis cannot be much better, if any, than

that of the original experiment. Downstream analysis cannot recover the value of a botched experiment any more than it can mar the results of a well-executed intervention.

In some cases scholars can alter the original analysis of results to either narrow down or expand the reach of the intervention, such as considering the effects of an intended treatment rather than the effects of the treatment on the treated. The ability of downstream analysts to make this type of decision is dependent on the depth and clarity of the description of the original intervention. In examining the original Perry intervention, Schweinhart, Barnes, and Weikart (1993) detailed their decision to move two subjects from the treatment group to the control group because of the subjects' unique family concerns. Such documentation allowed Sondheimer and Green (2010) to move these subjects into the original intent-to-treat group for use in their own downstream analysis of the program. This example speaks to the necessity of meticulous recordkeeping and documentation of decisions made prior to, during, and following a randomized intervention.

A second and related practical challenge is that of subject attrition. Depending on the topic of the investigation, there can be a long lag time between the initial intervention and the collection of new data for downstream analysis. The longer the window between the initial intervention and the downstream analysis, the more likely one is to face high levels of attrition. This loss of data becomes increasingly problematic if it is nonrandom and associated with factors relating to the initial intervention.

Concerns over attrition can be ameliorated if those performing the original intervention collect and maintain contact and other identifying information on their original subjects. If this data is lost or never collected, the possibilities for downstream analysis dissipate. Even if

researchers do not foresee the necessity of maintaining such information at the time of the initial intervention, the ability to reconnect with subjects cannot be undervalued. In the aftermath of his famous obedience studies, Milgram (1974) followed up with his subjects to evaluate their lingering responses and reactions to the intervention. Davenport et al. (2010) recently studied the enduring effects of randomized experiments testing the benefits of applying social pressure as a voter mobilization tool by analyzing registration rolls in subsequent elections. Collecting and keeping track of subject contact information may slightly increase the cost of the original experiment but has the potential to increase the payoffs in the long-term.

Both of these practical challenges can be overcome through cooperation among researchers. This advice may seem prosaic but, in the case of experiments, there is usually only one opportunity to perform an intervention and optimal execution of said experiment will influence future scholars' ability to build off of the results. Sharing experimental designs prior to implementation can open up important dialogue among scholars, dialogue that can improve initial interventions and heighten prospects for future downstream possibilities. Moreover, the maintenance of subject contact information may not be a first-order priority for researchers but may provide long-term benefits to one's own research team and others. The sharing of results among a broad swath of the research community can also increase the likelihood of extending the findings beyond their original intent. As seen above, many interventions give way to downstream analysis in entirely different subject areas. Cross-disciplinary collaboration on randomized experiments will help scholars approach previously intractable puzzles. This will be easier to do if researchers cooperate from the outset.

Ethical Challenges

While downstream analysis of preexisting randomized interventions provides researchers the opportunity to study exogenous changes in independent variables of interest normally deemed off-limits to outside manipulation, this type of analysis does pose some interesting ethical considerations with reference to subject consent and the use of deception. Concerns in both realms derive from the unending nature of experiments subject to future downstream analysis. The issues raised with regard to consent and deception should not limit scholars seeking to perform downstream analysis. Rather they are useful in considering how researchers planning new interventions ought to proceed to ensure that their contemporary research has value beyond its original intent.

In terms of consent, we should consider whether or not it is problematic for participants to be subjected to tests and data collection post hoc if they did not accede to further examination at the time of the original intervention. Such downstream research might be unobtrusive and not involve interaction with the original subjects but these concerns remain the same. Moreover, encouraging researchers to share data on experiments to allow for later analysis may violate some institutional Internal Review Board guidelines stipulating that research proposals detail the names of any investigators who will have access to data containing names and other identifiable information of participants in the original study.

Issues raised over deception and full disclosure closely parallel concerns over consent. While consent focuses on what the original researcher should do prior to the intervention, challenges concerning deception center around behavior following the initial intervention. The American Psychological Association's "Ethical Principles of Psychologists and Code of Conduct" (2003) stipulate that researchers using deception provide a full debriefing to

participants as soon as possible following the intervention but “no later than the conclusion of the data collection” (American Psychological Association 2003, Section 8.07). Ideally this disclosure should occur immediately following the conclusion of the intervention but researchers are permitted to delay the debriefing to maintain the integrity of the experiment. However this stipulation mandating a debriefing at the conclusion of the data collection is problematic in the face of downstream analysis because researchers might never be sure when, if ever, data collection is completed. Scholars must consider whether a full or even limited debriefing will impede the ability of future researchers to conduct downstream analysis of the original experiment and what the proper course of behavior should be given this potentiality.

5. Conclusion

Researchers conducting randomized experiments ought to consider the possibility for potential long- and short-term downstream analyses of their initial interventions. Doing so expands our estimates of the costs and benefits associated with randomized experimentation and provides unique research prospects for those who are unable to feasibly conduct such interventions. Awareness of these additional research opportunities can help structure the intervention and data collection process in ways amenable to downstream analysis. Downstream analysis is only possible if data are maintained and made available to others. Moreover, consideration of downstream possibilities prior to implementing a particular protocol will help researchers brainstorm the range of measures collected at the outset of a project. This will expand the value of the experiment to the individual researcher in the short-term and to the broader reaches of the research community in the long-term.

References

- American Psychological Association. 2003. *Ethical Principles of Psychologists and Codes of Conduct*.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444-55.
- Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15: 69-85.
- Berelson, Bernard, Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.
- Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work After They are Withdrawn? Evidence From a Natural Experiment in India." *American Political Science Review* 103: 23-35.
- Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.
- Converse, Philip E. 1972. "Change in the American Electorate. In *The Human Meaning of Social Change*, eds. Angus Campbell, and Philip E. Converse. New York: Russell Sage Foundation.
- Davenport, Tiffany C., Alan S. Gerber, Donald P. Green, Christopher W. Larimer, Christopher B. Mann, and Costas Panagopoulos. "The Enduring Effects of Social Pressure: Tracking Campaign Experiments Over a Series of Elections." Working Paper.
- Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know About Politics and Why it Matters*. New Haven: Yale University Press.
- Gay, Claudine. 2010. "Moving to Opportunity: The Political Effects of a Housing Mobility Experiment." Working Paper.
- Gerber, Alan S., Donald P. Green, and Ron Shacher. 2003. "Voting May be Habit-Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 27: 540-50.
- Green, Donald P., and Alan S. Gerber. 2002. "The Downstream Benefits of Experimentation." *Political Analysis* 10: 394-402.
- Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62: 467-75.
- Imbens, Guido W., and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25: 305-27.

- Kam, Cindy, and Carl L. Palmer. 2008. "Reconsidering the Effects of Education on Civic Participation." *Journal of Politics* 70: 612-31.
- Katz, Lawrence F., Jeffery R. Kling, and Jeffery B. Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics* 116: 607-54.
- Kellermann, Michael, and Kenneth A. Shepsle. 2009. "Congressional Careers, Committee Assignments, and Seniority Randomization in the US House of Representatives." *Quarterly Journal of Political Science* 4: 87-101.
- Leventhal, Tama, and Jeanne Brooks-Gunn. 2003. "Moving to Opportunity: An Experimental Study of Neighborhood Effects on Mental Health." *American Journal of Public Health* 93: 1576-82.
- Ludwig, Jens, Greg J. Duncan, and Paul Hirschfield. 2001. "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment." *Quarterly Journal of Economics* 116: 655-79.
- Milgram, Stanley. 1974. *Obedience to Authority: An Experimental View*. 1st ed. New York: Harper & Row.
- Miller, Warren E., and J. Merrill Shanks. 1996. *The New American Voter*. Cambridge, MA: London: Harvard University Press.
- Nie, Norman H., Jane Junn, and Kenneth Stehlik-Barry. 1996. *Education and Democratic Citizenship in America*. Chicago: University of Chicago Press.
- Peffley, Mark, and Jon Hurwitz. 2007. "Persuasion and Resistance: Race and the Death Penalty in America." *American Journal of Political Science* 51: 996-1012.
- Rosenstone, Steven J., and John Mark Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Macmillan.
- Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38: 827-74.
- Sanbonmatsu, Lisa, Jeffery R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. 2006. "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." *Journal of Human Resources* XLI: 649-91.
- Schweinhart, L. J., Helen V. Barnes, and David P. Weikart. 1993. *Significant Benefits: The High-Scope Perry Preschool Study Through Age 27*. Monographs of the High/Scope Educational Research Foundation, Vol. 10. Ypsilanti, MI: High/Scope Press.

- Sondheimer, Rachel Milstein and Donald P. Green. 2010. "Using Experiments to Estimate the Effects of Education on Voter Turnout." *American Journal of Political Science* 54: 174-89.
- Sovey, Allison J., and Donald P. Green. 2009. "Instrumental Variables Estimation: A Reader's Guide." Working Paper.
- Stock, James H., and Mark W. Watson. 2007. *Introduction to econometrics*. 2nd ed. Boston: Pearson Addison Wesley.
- Tenn, Steven. 2007. "The Effect of Education on Voter Turnout." *Political Analysis* 15: 446-64.
- Verba, Sidney, and Norman H. Nie. 1972. *Participation in America: Political Democracy and Social Equality*. New York: Harper & Row.
- Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.
- Wooldridge, Jeffrey M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. Mason, Ohio: South-Western Cengage Learning.

Table 34-1. Classification of Target Population in Downstream Analysis of Educational Intervention¹

Group Number	Type	Graduates from high school if Assigned to Treatment?	Graduates from High School if Assigned to Control?	Votes if Graduates from High School? (y_{it})	Votes if Does Not Graduate from High School? (y_{i0})	Share of the Population
1	Never-takers	No	No	No	No	p_1
2		No	No	Yes	No	p_2
3		No	No	No	Yes	$p_3^{a b}$
4		No	No	Yes	Yes	$p_4^{a b}$
5	Compliers	Yes	No	No	No	p_5
6		Yes	No	Yes	No	p_6^a
7		Yes	No	No	Yes	p_7^b
8		Yes	No	Yes	Yes	$p_8^{a b}$
9	Always-takers	Yes	Yes	No	No	p_9
10		Yes	Yes	Yes	No	$p_{10}^{a b}$
11		Yes	Yes	No	Yes	p_{11}
12		Yes	Yes	Yes	Yes	$p_{12}^{a b}$
13	Defiers	No	Yes	No	No	p_{13}
14		No	Yes	Yes	No	p_{14}^b
15		No	Yes	No	Yes	p_{15}^a
16		No	Yes	Yes	Yes	$p_{16}^{a b}$

¹ This table is adapted from Sovey and Green (2009)

^a This share of the population votes if assigned to the treatment group.

^b This share of the population votes if assigned to the control group.

ⁱ The views expressed in this paper are solely those of the author and do not represent the views of the United States Military Academy, the Department of the Army, and/or the Department of Defense.

ⁱⁱ Imbens and Angrist (1994, 467-475) provide formal analysis of the distinction between LATE and average treatment effects.

35. Mediation Analysis Is Harder than It Looks

John G. Bullock and Shang E. Haⁱ

Mediators are variables that transmit causal effects from treatments to outcomes. Those who undertake mediation analysis seek to answer “how” questions about causation: how does this treatment affect that outcome? Typically, we desire answers of the form “the treatment affects a causally intermediate variable, which in turn affects the outcome.” Identifying these causally intermediate variables is the challenge of mediation analysis.

Conjectures about political mediation effects are as old as the study of politics. But codification of procedures by which to test hypotheses about mediation is a relatively new development. The most common procedures are now ubiquitous in psychology (Quiñones-Vidal et al. 2004) and increasingly popular in the other social sciences, not least political science.

Unfortunately, the most common procedures are not very good. They call for indirect effects—the portions of treatment effects that are transmitted through mediators—to be estimated via multi-equation regression frameworks. These procedures do not require experimental manipulation of mediators; instead, they encourage the study of mediation with data from unmanipulated mediators (see MacKinnon et al. 2002, 86; Spencer, Zanna, and Fong 2005). The procedures are therefore prone to producing biased estimates of mediation effects. Warnings about this problem have been issued for decades by statisticians, psychologists, and political scientists.

Recognizing that nonexperimental methods of mediation analysis are likely to be biased, social scientists are slowly turning to methods that involve experimental manipulation of mediators. This is a step in the right direction. But experimental mediation analysis is difficult – more difficult than it may seem – because experiment-based inferences about indirect effects are subject to important but little-recognized limitations. The point of this chapter is to explain the bias to which nonexperimental methods are prone and to describe experimental methods that

hold out more promise of generating credible inferences about mediation. But it is also to describe the limits of experimental mediation analysis.

This chapter proceeds as follows. We begin by characterizing the role that mediation analysis plays in political science. We then describe conventional methods of mediation analysis and the bias to which they are prone. We proceed by describing experimental methods that can reliably produce accurate estimates of mediation effects. The experimental approach has several important limitations, and we end the section by explaining how these limitations imply both best practices and an agenda for future research. We consider objections to our argument in the next section, including the common objection that manipulation of mediators is often infeasible. Our last section reviews and concludes.

1. Mediation Analysis in Political Science

The questions that animate political scientists can be classified epistemologically. Some are purely descriptive. Others – the ones to which experiments are especially well-suited – are about treatment effects. (“Does *X* affect *Y*? How much? Under what conditions?”) But questions about mediation belong to a different category. When social scientists seek information about “processes” or demand to know about the “mechanisms” through which treatments have effects, they are asking about mediation. Indeed, when social scientists speak about “explanation” and “theory,” mediation is usually what they have in mind.

Social scientists often try to buttress their claims about mediation with data. They use a variety of methods to do so, but nearly all are based on crosstabulations or multi-equation regression frameworks. In this chapter, we focus on one such method: the one proposed by Baron and Kenny (1986). We focus on it because it is simple, by far the most common method, and similar to almost all other methods in use today. It originated in social psychology, where its influence is now hard to overstate.ⁱⁱ And within political science, it is most prominent among articles that have explicitly psychological aims. For example, Brader, Valentino, and Suhay (2008) use the procedure to examine the ways in which emotions mediate the effects of news about immigration on willingness to send a message about the issue to members of Congress.

Fowler and Dawes (2008, 586-8) use it to test hypotheses about mediators of the connection between genes and turnout. And several political scientists have used it to understand the mechanisms that underpin priming and framing effects in political contexts (e.g., Nelson 2004; Malhotra and Krosnick 2007).

To some, the increasing use of the Baron-Kenny method in political science seems a good thing: it promises to bring about “valuable theoretical advances” and is just what we need to “push the study of voting up a notch or two in sophistication and conceptual payoffs” (Malhotra and Krosnick 2007, 250, 276). But increasing use of the Baron-Kenny method is not a good thing. Like related methods that do not require manipulation of mediators, it is biased, and in turn it leads researchers to biased conclusions about mediation.

2. Nonexperimental Mediation Analyses Are Prone to Bias

Like many related procedures, the method proposed by Baron and Kenny (1986) is based on three models:

$$M = \alpha_1 + aX + e_1, \tag{1}$$

$$Y = \alpha_2 + cX + e_2, \text{ and} \tag{2}$$

$$Y = \alpha_3 + dX + bM + e_3, \tag{3}$$

Where Y is the outcome of interest, X is a treatment, M is a potential mediator of the treatment, and α_1 , α_2 , and α_3 are intercepts. For simplicity, we assume that X and M are binary variables coded either 0 or 1. The unobservable disturbances e_1 , e_2 , and e_3 are mean-zero error terms that represent the cumulative effect of omitted variables. It is not difficult to extend this framework to include multiple mediators and other covariates, and our criticisms apply with equal force to models that include such variables. For notational clarity and comparability to previous articles about mediation analysis, we limit our discussion to the three-variable regression framework.

For simplicity, we assume throughout this chapter that X is randomly assigned such that it is independent of the disturbances: $e_1, e_2, e_3 \perp\!\!\!\perp X$. As we shall see, randomization of X alone does

not ensure unbiased estimation of the effects of mediators. Consequently, we refer to designs in which only X is randomized as *nonexperimental* for the purpose of mediation analysis, reserving *experimental* for studies in which both X and M are randomized.

The coefficients of interest are a , b , c , and d . The total effect of X on Y is c . To see how c is typically decomposed into “direct” and “indirect” effects, substitute Equation 1 into Equation 3, yielding

$$Y = \alpha_3 + X(d + ab) + (\alpha_1 + e_1)b + e_3.$$

The direct effect of X is d . The indirect or “mediated” effect is ab (or, equivalently, $c - d$).ⁱⁱⁱ

Baron and Kenny do not say how the coefficients in these equations are to be estimated; in practice, OLS is almost universally used. But the OLS estimator of b in Equation 3 is biased:

$$E[\hat{b}] = b + \frac{\text{cov}(e_1, e_3)}{\text{var}(e_1)}.$$

The OLS estimator of d is biased, too:

$$E[\hat{d}] = d - a \cdot \frac{\text{cov}(e_1, e_3)}{\text{var}(e_1)}.$$

(A proof is given in Bullock, Green, and Ha 2008, 39-40.) OLS estimators of direct and indirect effects will therefore be biased as well.

In expectation, the OLS estimators of b and d produce accurate estimates only if $\text{cov}(e_1, e_3) = 0$.^{iv} But this condition is unlikely to hold unless both X and M are randomly assigned. The problem is straightforward: if an unobserved variable affects both M and Y , it will cause e_1 and e_3 to covary. And even if no unobserved variable affects both M and Y , these disturbances are likely to covary if M is merely correlated with an unobserved variable that affects Y —for example, another mediator. This “multiple-mediator problem” is a serious threat to social-science mediation analysis because most of the effects that interest social scientists are likely to have multiple correlated mediators. Indeed, we find it difficult to think of any political effects that do

not fit this description.^v

The standard temptation in nonexperimental analysis is to combat this problem by controlling for potential mediators other than M . But it is normally impossible to measure all possible mediators. Indeed, it may be impossible to merely think of all possible mediators. And controlling for some potential mediators but not all of them is no guarantee of better estimates; to the contrary, it may make estimates worse (Clarke 2009). Fighting endogeneity in nonexperimental mediation analysis by adding control variables is a method with no clear stopping rule or way to detect bias—a shaky foundation on which to build beliefs about mediation.

Political scientists who use the Baron-Kenny method and related methods often want to test hypotheses about several potential mediators rather than one. In these cases, the most common approach is “one-at-a-time” estimation, whereby Equation 3 is estimated separately for each mediator. This practice makes biased inferences about mediation even more likely. The researcher, who already faces the spectre of bias due to the omission of variables over which she has no control, compounds the problem by intentionally omitting variables that are likely to be important confounds. Nonexperimental mediation analysis is problematic enough, but one-at-a-time testing of mediators stands out as an especially bad practice.

The Baron-Kenny method and related methods are often applied to experiments in which the treatment has been randomized but the mediator has not, and there seems to be a widespread belief that such experiments are sufficient to ensure unbiased estimates of direct and indirect effects. But randomization of the treatment is not enough to protect researchers from biased estimates. It can ensure that X bears no systematic relationship to e_1 , e_2 , or e_3 , but it says nothing about whether M is systematically related to those variables, and thus nothing about whether $\text{cov}(e_1, e_3) = 0$.^{vi}

Stepping back from mediation analysis to the more general problem of estimating causal effects, note that estimators tend to be biased when one controls for variables that are affected by

the treatment. One does this whenever one controls for M in a regression of Y on X , which the Baron-Kenny method requires. This “post-treatment bias” has been discussed in statistics and political science (e.g., Rosenbaum 1984, 188-94; King and Zeng 2006, 146-48), but its relevance to mediation analysis has gone largely unnoticed. At root, it is one instance of an even more general rule: estimators of the parameters of regression equations are likely to be unbiased only if the predictors in those equations are independent of the disturbances. And in most cases, the only way to ensure that M is independent of the disturbances is to randomly assign its values. By contrast, “the benefits of randomization are generally destroyed by including post-treatment variables” that have not been manipulated (Gelman and Hill 2007, 192).

Within the last decade, statisticians and political scientists have advanced several different methods of mediation analysis that do not call for manipulation of mediators. These methods improve on Baron and Kenny (1986), but they do not overcome the problem of endogeneity in nonexperimental mediation analysis. For example, Frangakis and Rubin (2002) propose “principal stratification,” which entails dividing subjects into groups on the basis of their potential outcomes for mediators. Causal effects are then estimated separately for each of these “principal strata.” The problem is that some potential outcomes for each subject are necessarily unobserved, and those who use principal stratification must infer the values of these potential outcomes on the basis of covariates. In practice, “this reduces to making the same kinds of assumptions as are made in typical observational studies when ignorability is assumed” (Gelman and Hill 2007, 193).

In a different vein, Imai, Keele, and Yamamoto (2010) show that indirect effects can be identified even when the mediator is not randomized – provided that we stipulate the size of $\text{cov}(e_1, e_3)$. This is helpful: if we are willing to make assumptions about the covariance of unobservables, we may be able to place bounds on the likely size of the indirect effect. But in no sense is this method a substitute for experimental manipulation of the mediator. Instead, it requires us to make strong assumptions about the properties of unobservable disturbances, just as other methods do when they are applied to nonexperimental data. Moreover, Imai et al. (2010,

43) note that even if we are willing to stipulate the value of $\text{cov}(e_1, e_3)$, the method that they propose permits causal interpretation of indirect-effect estimates only if all unobserved confounders are causally prior to the treatment. This rules out use of their method to estimate indirect effects whenever the mediator of interest is directly affected by both the treatment and another mediator. This point is crucial because many effects that interest political scientists seem likely to be transmitted by multiple mediators that affect each other.

None of these warnings imply that all nonexperimental mediation research is equally suspect. All else equal, research in which only a treatment is randomized is preferable to research in which no variables are randomized; treatment-only randomization does not make accurate mediation inference likely, but it does clarify the assumptions required for accurate inference. And in general, nonexperimental research is better when its authors attempt to justify the assumption that their proposed mediator is uncorrelated with other variables, including unobserved variables, that may also be mediators. This sort of argument can be made poorly or well. But even the best arguments of this type typically warrant far less confidence than arguments about unconfoundedness that follow directly from manipulation of both the treatment and the mediator.

This discussion should make clear that the solution to bias in nonexperimental mediation analyses is unlikely to be another nonexperimental mediation analysis. The problem is that factors affecting the mediator and the outcome are likely to covary. We are not likely to solve this problem by controlling for more variables, measuring them more accurately, or applying newer methods to nonexperimental data. To calculate unbiased estimates of mediation effects, we should look to experiments.

3. Experimental Methods of Mediation Analysis

The simplest experimental design that permits accurate estimation of indirect effects entails direct manipulation of treatments and mediators. We have described such a design elsewhere (Bullock et al. 2008), but in many cases, limited understanding of mediators precludes direct manipulation. For example, although we can assign subjects to conditions in which their

feelings of efficacy are likely to be heightened or diminished, we do not know how to gain direct experimental control over efficacy. That is, we do not know how to assign specific levels of efficacy to different subjects. The same is true of party identification, emotions, cultural norms, modes of information processing, and other likely mediators of political processes. These variables and others are beyond direct experimental control.

But even when mediators are beyond direct experimental control, we can often manipulate them indirectly. The key in such cases is to create an instrument for M , the endogenous mediator. To be a valid instrument for M , a variable must be correlated with M but uncorrelated with e_3 . Many variables are likely to satisfy the first condition: whatever M is, it is usually not hard to think of a variable that is correlated with it, and once we have measured this new variable, estimating the correlation is trivial. But satisfying the second condition is more difficult. Because e_3 is unobservable, we can never directly test whether it is uncorrelated with the potential instrument. Worse, almost every variable that is correlated with M is likely to be correlated with other factors that affect Y , and thus likely to be correlated with e_3 .^{vii}

Fortunately, a familiar class of variables meets both conditions: assignment-to-treatment variables. Use of these instrumental variables is especially common in analyses of field experiments, where compliance with the treatment is likely to be partial. For example, Gerber and Green (2000) use a field experiment to study various means of increasing voter turnout. They cannot directly manipulate the treatments of interest: they cannot compel their subjects to read mail, answer phone calls, or speak to face-to-face canvassers. Instead, they use random assignments to these treatments as instruments for the treatments themselves. Doing so permits them to recover accurate estimates of treatment effects even though the treatments are beyond direct experimental control (for elaboration of this point, see Angrist, Imbens, and Rubin 1996; Gerber's chapter in this volume).

Although the instrumental variables approach is increasingly used to estimate average treatment effects, it has not yet been used in political science to study mediation. We think that it should be. It has already been used multiple times to study mediation in social psychology, and its use in that discipline suggests how it might be used in ours. For example, Zanna and Cooper (1974) hypothesize that attitude-behavior conflict produces feelings of unpleasant tension (“aversive arousal”), which in turn produces attitude change. They cannot directly manipulate levels of tension, so they use an instrument to affect it indirectly: subjects swallow a pill and are randomly assigned to hear that it will make them tense, make them relax, or have no effect. In a related vein, Bolger and Amarel (2007) hypothesize that the effect of social support on the stress levels of recipients is mediated by efficacy: support reduces recipients’ stress by raising their feelings of efficacy. Bolger and Amarel cannot directly assign different levels of efficacy to different participants in their experiment. Instead, they randomly assign subjects to receive personal messages that are designed to promote or diminish their feelings of efficacy. In this way, they indirectly manipulate efficacy.

To see how such instruments might be created and used in political science, consider research on issue framing. A controversial hypothesis is that framing an issue in a particular way changes attitudes by increasing the accessibility of particular thoughts about the issue, i.e., the ease with which particular thoughts come to mind (see Iyengar and Kinder 1987, esp. ch. 7; Nelson, Clawson, and Oxley 1997; Miller and Krosnick 2000). Political scientists do not know how to directly manipulate the accessibility of particular thoughts. But they do know how to indirectly manipulate accessibility by priming people in different ways (e.g., Burdein, Lodge, and Taber 2006, esp. 363-64; see also Lodge and Taber’s chapter in this volume). Experimental analysis of the hypothesis is therefore possible. Following Equation 3, consider the model:

$$attitudes = \alpha_3 + d(framing) + b(accessibility) + e_3.$$

In this model, *framing* indicates whether subjects were assigned to a control condition (*framing* =

0) or an issue frame ($framing = 1$); *accessibility* is reaction times in milliseconds in a task designed to gauge the accessibility of particular thoughts about the issue; and e_3 is a disturbance representing the cumulative effect of other variables. Crucially, *accessibility* is not randomly assigned. It is likely to be affected by framing and to be correlated with unobserved variables represented by e_3 : age, intelligence, and political predispositions, among others.

The OLS estimator of b , the effect of accessibility, is therefore likely to be biased. (The OLS estimator of d , the direct effect of the framing manipulation, is also likely to be biased.) But suppose that in addition to the framing manipulation and the measurement of accessibility, some subjects are randomly assigned to a condition in which relevant considerations are primed. This priming manipulation may make certain thoughts about the issue more accessible. In this case, accessibility remains nonexperimental, but the priming intervention generates an instrumental variable that we can use to consistently estimate b . If we also estimate a – for example, by conducting a second experiment in which only framing is manipulated – our estimator of ab , the extent to which priming mediates framing, will also be consistent.

The most common objection to experimental mediation approaches is that they often cannot be used because mediators often cannot be manipulated. We take up this objection below, but for the moment, we stress that researchers need not seek complete experimental control over mediators. They need only seek some randomization-based purchase on mediators. Consider, for example, one of the best-known and least tractable variables in political behavior research: party identification. The history of party-ID studies suggests that it should be difficult to manipulate. It is one of the most stable individual-level influences on votes and attitudes, and no one knows how to assign different levels of party ID to different subjects. But party ID can be changed by experiments, and such experiments are the key to understanding its mediating power. For example, Brader and Tucker (2008) use survey experiments to show that party cues can change Russians' party IDs. And Gerber, Huber, and Washington (2008) use a field experiment to show

that registering with a party can produce long-term changes in party ID. The most promising path to secure inferences about party ID as a mediator is to conduct studies in which interventions like these are coupled with manipulations of policy preferences, candidate evaluations, or other treatments. And in general, the most promising path to secure inferences about mediation is to design studies that include experimental manipulations of both treatments and mediators.

4. Three Limitations of Experimental Mediation Analysis

Despite its promise, the experimental approach has limitations that merit more attention than they typically receive. It requires researchers to devise experimental manipulations that affect one mediator without affecting others. Even if researchers succeed, their estimates of indirect effects will typically apply only to a subset of the experimental sample. Finally, if causal effects are not identical for all members of a sample, even a well-designed experiment may lead to inaccurate inferences about indirect effects. We discuss these limitations at length in other work (Bullock, Green, and Ha 2010; Green, Ha, and Bullock 2010); here, we offer a brief overview of each.^{viii}

An experimental intervention is useful for mediation analysis if it affects one mediator without affecting others. If the intervention instead affects more than one mediator, it violates the exclusion restriction – in terms of Equation 3, it is correlated with e_3 – and is not a valid instrument. In this case, the instrumental-variables estimator of the indirect effect will be biased. For example, issue frames may affect attitudes by changing the accessibility of relevant considerations but also by changing the subjective relevance of certain values to the issue at hand (Nelson et al. 1997). In this case, an experimental intervention can identify the mediating role of accessibility only if it primes relevant considerations without affecting the subjective relevance of different values. And by the same token, an experimental intervention will identify the mediating role of value weighting only if it affects the subjective relevance of different values without changing the accessibility of considerations. The general challenge for experimental researchers, then, is to devise manipulations that affect one mediator without affecting others.^{ix}

Even if researchers isolate particular mediators, they must confront another dilemma:

some subjects never take a treatment even if they are assigned to take it, and a treatment effect cannot be meaningfully estimated for such people. Consequently, the experimental approach to mediation analysis produces estimates of the average treatment effect not for all subjects but only for “compliers” who can be induced by random assignment to take it (Imbens and Angrist 1994). For example, if some subjects are assigned to watch a presidential campaign advertisement while others are assigned to a no-advertisement control group, the average effect of the ad can be identified not for all subjects but only for (1) treatment-group subjects who are induced by random assignment to watch the ad and (2) control-group subjects who would have been induced to watch the ad if they had been assigned to the treatment group. One may assume that the average indirect effect is the same for these subjects as for others, but this is an assumption, not an experimental result. Strictly speaking, estimates of the average indirect effect apply only to a subset of the sample. We can usually learn something about the characteristics of this subset (Angrist and Pischke 2009, 166-72), but we can never know exactly which subjects belong to it.

An unintuitive consequence follows: even if we use experiments to manipulate both a treatment and a mediator, we may not be able to estimate an average indirect effect for our experimental sample or any subset of it. To see why, recall that the indirect effect of X on Y in Equations 1 through 3 is ab . By manipulating X , we can recover \hat{a} , an estimate of the average effect of X on M among those whose value of X can be affected by the X -manipulation. And by manipulating M , we can recover \hat{b} , an estimate of the average effect of M on Y among those whose value of M can be affected by the M -manipulation. If these two populations are the same, $\hat{a}\hat{b}$ is a sensible estimate of the local average treatment effect. But if these two populations differ – if one set of subjects is affected by the manipulation of X but a different set is affected by the manipulation of M – $\hat{a}\hat{b}$ is the causal effect of X on M for one group of people times the causal effect of M on Y for another group of people. This product has no causal interpretation. It is just an unclear mixture of causal effects for different groups.^x

A related problem is that experiments cannot lead to accurate estimates of indirect effects when the effects of X on M are not the same for all subjects or when the effects of M on Y are not

the same for all subjects. When we are not studying mediation, the assumption of unvarying effects does little harm: if the effect of randomly manipulated X on Y varies across subjects, and we regress Y on X , the coefficient on X simply indicates the average effect of X . But if the effects of X and M vary across subjects, it will typically be difficult to estimate an average indirect effect (Glynn 2009). To see why, consider an experimental sample in which there are two groups of subjects. In the first group, the effect of X on M is positive, and the effect of M on Y is also positive. In the second group, the effect of X on M is negative, and the effect of M on Y is also negative. In this case, the indirect effect of X is positive for every subject in the sample: to slightly adapt the notation of Equations 1 and 3, $a_i b_i$ is positive for every subject. But \hat{a} , the estimate of the average effect of X on M , may be positive, negative, or zero. And \hat{b} , the estimate of the average effect of M on Y , may be positive, negative, or zero. As a result, the estimate of the average indirect effect, $\hat{a}\hat{b}$, may be zero or negative – even though the true indirect effect is positive for every subject.

Such problems may arise whenever different people are affected in different ways by X and M . For example, Cohen (2003) wants to understand how reference-group cues (X) affect attitudes toward social policy (Y). In his experiments, politically conservative subjects receive information about a generous welfare policy; some of these subjects are told that the policy is endorsed by the Republican Party, while others receive no endorsement information. Cohen's findings are consistent with cues (endorsements) promoting systematic thinking (M) about the policy information, and with systematic thinking in turn promoting positive attitudes toward the policy (Cohen 2003, esp. 817).^{xi} On the other hand, Petty and Wegener (1998, 345) and others suggest that reference-group cues inhibit systematic thinking about information, and that such thinking promotes the influence of policy details – which might be expected to lead, in this case, to negative attitudes toward the welfare policy among the conservative subjects. For present purposes, there is no need to favor either of these theories or to attempt a reconciliation. We need only note that they suggest a case in which causal effects may be heterogeneous, and in which mediation analysis is therefore difficult. Let some subjects in an experiment be “Cohens”: for

these people, exposure to reference-group cues heightens systematic thinking (a_i is positive) and systematic thinking makes attitudes toward a generous welfare policy more favorable (b_i is positive). But other subjects are “Petties”: for them, exposure to reference-group cues limits systematic thinking (a_i is negative) and systematic thinking makes attitudes toward a generous welfare policy less favorable (b_i is negative). Here again, the indirect effect is positive for every subject, because $a_i b_i > 0$ for all i . But if the experimental sample includes both Cohens and Petties, \hat{a} and \hat{b} may each be positive, negative, or zero. Conventional estimates of the average indirect effect— $\hat{a}\hat{b}$ and related quantities—may therefore be zero or even negative.

Moreover, causal effects need not differ so sharply across members of a sample to make mediation analysis problematic. Conventional estimates of indirect effects will be biased if a and b merely covary among subjects within a sample. For example, if a subset of subjects is more sensitive than the rest of the sample to changes in X and to changes in M , estimates of indirect effects will be biased. This problem cannot be traced to a deficiency in the methods that are often used to calculate indirect effects: it is fundamental, not a matter of statistical technique (Robins 2003; Glynn 2009).

5. An Agenda for Mediation Analysis

These limitations of experimental mediation analysis—it requires experimenters to isolate particular mediators, produces estimates that apply only to an unknown subset of subjects, and cannot produce meaningful inferences about mediation when causal effects covary within a sample—are daunting. Experiments are often seen as simplifying causal inference, but taken together, these limitations imply that strong inferences about mediation are likely to be difficult even when researchers use fully experimental methods of mediation analysis. Still, none of our cautions imply that experiments are useless for mediation analysis. Nor do they imply that experimental mediation analysis is no better than the nonexperimental alternative. Unlike nonexperimental methods, experiments offer – albeit under limited circumstances – a systematic way to identify mediation effects. And the limitations that we have described here are helpful

inasmuch as they delineate an agenda for future mediation analysis.

First, researchers who do not manipulate mediators should try to explain why the mediators are independent of the disturbances in their regression equations – after all, the accuracy of their estimates hinges on this assumption. In practice, justifying this assumption entails describing unmeasured mediators that may link X to Y and explaining why these mediators do not covary with the measured mediators. Such efforts are rarely undertaken, but without them, it is hard to hold out hope that nonexperimental mediation analysis will generate credible findings about mediation.

Second, researchers who experimentally manipulate mediators should explain why they believe that their manipulations are isolating individual mediators. This entails describing the causal paths by which X may affect Y and explaining why each experimental manipulation affects only one of these paths. The list of alternative causal paths may be extensive, and multiple experiments may be needed to demonstrate that a given intervention tends not to affect the alternative paths in question.

Third, researchers can improve the state of mediation analysis simply by manipulating treatments and then measuring the effects of their manipulations on many different outcomes. To see how this can improve mediation analysis, consider studies of the effects of campaign contact on voter turnout. In addition to assessing whether a particular kind of contact increases turnout, one might also survey participants to determine whether this kind of contact affects interest in politics, feelings of civic responsibility, knowledge about where and how to vote, and other potential mediators. In a survey or laboratory experiment, this extra step need not entail a new survey: relevant questions can instead be added to the post-test questionnaire. Because this kind of study does not include manipulations of both treatments and mediators, it cannot reliably identify mediation effects. But if some variables seem to be unaffected by the treatment, one may begin to argue that they do not explain why the treatment works.

Fourth, researchers should know that if the effects of X and M vary from subject to subject within a sample, it may be impossible to estimate the average indirect effect for the entire sample.

To determine whether this is a problem, one can examine the effects of X and M among different types of subjects. If the effects differ little from group to group (e.g., from men to women, whites to nonwhites, the wealthy to the poor), we can be relatively confident that causal heterogeneity is not affecting our analysis.^{xii} On the other hand, if there are large between-group differences in the effects of X or M , mediation estimates made for an entire sample may be inaccurate even if X and M have been experimentally manipulated. In this case, researchers should aim to make multiple inferences for relatively homogeneous subgroups rather than single inferences about the size of an indirect effect for an entire sample.

6. Defenses of Conventional Practice

In different ways, statisticians (Rosenbaum 1984; Rubin 2004; Gelman and Hill 2007, 188-94), social psychologists (James 1980; Judd and Kenny 1981, 801), and political scientists (King and Zeng 2006, 146-48; Glynn 2009) have all warned that methods like the one proposed by Baron and Kenny (1986) are likely to produce meaningless or inaccurate conclusions when applied to observational data. Why have their arguments not taken hold? Some of the reasons are mundane: the arguments are typically made in passing; their relevance to mediation analysis is not always clear; there are few such arguments in any one discipline, and scholars rarely read outside of their own disciplines. But these are not the only reasons. Another part of the answer lies with three defenses of nonexperimental mediation analysis, which can also be framed as criticisms of the experimental approach.

The first and most common defense is that many mediators cannot be manipulated and that insistence on experimental mediation analysis therefore threatens to limit the production of knowledge (e.g., James 2008; Kenny 2008). Manipulation of mediators is indeed difficult in some cases, but we think that this objection falls short on several counts (Bullock et al. 2008, 28-9). First, it follows from a misunderstanding of the argument. No one maintains that unmanipulable variables should not be studied or that causal inferences should be drawn only from experiments. The issue lies instead with the accuracy of nonexperimental inferences and the degree of confidence that we should place in them. In the absence of natural experiments,

dramatic effects, or precise theory about data-generating processes – that is, in almost all situations that social scientists examine – nonexperimental studies are likely to produce biased estimates of indirect effects and to justify only weak inferences. Moreover, the objection is unduly pessimistic, likely because it springs from a failure to see that many variables that cannot be directly manipulated can be indirectly manipulated. Perhaps some mediators defy even indirect manipulation, but in light of increasing experimental creativity throughout the discipline – exemplified by several other chapters in this volume – we see more cause for optimism than for despair.

A second objection is that the problem of bias in mediation analysis is both well understood and unavoidable. The solution, according to those who make this objection, is not to embrace experimentation but to “build better models” (e.g., James 1980). The first part of this objection is implausible: those who analyze mediation may claim to be aware of the threat of bias, but they typically act as though they are not. Potential mediators other than the one being tested are almost never discussed in conventional analyses, even though their omission makes bias likely. When several mediators are hypothesized, it is common to see each one analyzed in a separate set of regressions rather than collectively, which further increases the probability of bias.

This makes the second part of the objection – that the way to secure inferences about mediation is to “build better models” – infeasible. In the absence of experimental benchmarks (e.g., LaLonde 1986), it is difficult to know what makes a model better. Merely adding more controls to a nonexperimental mediation analysis is no guarantee of better estimates, common practice to the contrary. It may well make estimates worse (Clarke 2009).

A more interesting argument is that social scientists are not really interested in point estimation of causal effects (Spencer et al. 2005, 846). They report precise point estimates in their tables, but their real concern is statistical significance, i.e., bounding effects away from zero. And for this purpose, the argument goes, conventional methods of mediation analysis do a pretty good job. The premise of this argument is correct: many social scientists care more about bounding effects away from zero than they do about learning the size of effects. But this indifference to the

size of effects is regrettable. Our stock of accumulated knowledge speaks much more to the existence of effects than to their size, and this makes it difficult to know which effects are important. And even if the emphasis on bounding results away from zero were appropriate, there would not be reason to think that conventional mediation analysis does a good job of helping us to learn about bounds. As Imai et al. (2010, 43) note, even a well-developed framework for sensitivity analysis cannot produce meaningful information about mediation when important omitted variables are causally subsequent to the treatment.

7. Conclusion

Experiments have taught us much about treatment effects in politics, but our ability to explain these effects remains limited. Even when we are confident that a particular variable mediates a treatment effect, we are usually unable to speak about its importance in either an absolute sense or relative to other mediators. Given this state of affairs, it is not surprising that many political scientists want to devote more attention to mediation.

But conventional mediation analysis, which draws inferences about mediation from unmanipulated mediators, is a step backward. These analyses are biased, and their widespread use threatens to generate a store of misleading inferences about causal processes in politics. The situation would be better if we could hazard guesses about the size and direction of the biases. But we can rarely take even this small step with confidence, because conventional mediation analyses rarely discuss mediators other than those that have been measured. Instead, conventional analyses are typically conducted as though they were fully experimental, with no consideration of threats to inference.

A second, worse problem is the impression conveyed by the use and advocacy of these methods: the impression that mediation analysis is easy, or at least no more difficult than running a few regressions. In reality, secure inferences about mediation typically require experimental manipulation of both treatments and mediators. But experimental inference about mediation, too, is beset by limitations. It requires researchers to craft interventions that affect one mediator without affecting others. If researchers succeed in this, their inferences will typically apply only

to an unknown subset of subjects in their sample. And if the effects of the treatment and the mediator are not the same for every subject in the sample, even well-designed experiments may be unable to yield meaningful estimates of average mediation effects for the entire sample. In the most difficult cases, it may be impossible to learn about mediation without making strong and untestable assumptions about the relationships among observed and unobserved variables.

The proper conclusion is not that mediation analysis is hopeless but that it is difficult. Experiments with theoretically refined treatments can help by pointing to mediators that merit further study. Experiments in which mediators are manipulated are even more promising. And analysis of distinct groups of subjects can strengthen mediation analysis by showing us whether it is possible to estimate average indirect effects for general populations or whether we must instead tailor our mediation analyses to specific groups. But because of the threats to inference that we have described, any single experiment is likely to justify only the most tentative inferences about mediation. Understanding the processes that mediate even a single treatment effect will typically require a research program comprising multiple experiments—experiments that address the challenges described here.

It is worthwhile to draw a lesson from other social sciences, where manipulation of mediators is rare but mediation analysis is ubiquitous. In these disciplines, promulgation of nonexperimental procedures has given rise to a glut of casual inferences about mediation that warrant little confidence. Even the scholar who has arguably done most to promote nonexperimental mediation analysis now laments that social scientists often “do not realize that they are conducting causal analyses” and fail to justify the assumptions that underpin those analyses (Kenny 2008, 356). It would be a shame if political scientists went the same route. We can stay on track by remembering that inference about mediation is difficult—much more difficult than conventional practice suggests.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. “Identification of Causal

- Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91: 444-55.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton, NJ: Princeton University Press.
- Angrist, Joshua D., Victor Lavy, and Analia Schlosser. In press. “Multiple Experiments for the Causal Link Between the Quantity and Quality of Children.” *Journal of Labor Economics*.
- Baron, Reuben M., and David A. Kenny. 1986. “The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.” *Journal of Personality and Social Psychology* 51: 1173-82.
- Bartels, Larry M. 1991. “Instrumental and ‘Quasi-Instrumental’ Variables.” *American Journal of Political Science* 35: 777-800.
- Bolger, Niall, and David Amarel. 2007. “Effects of Social Support Visibility on Adjustment to Stress: Experimental Evidence.” *Journal of Personality and Social Psychology* 92: 458-75.
- Bound, John, David A. Jaeger, and Regina M. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak.” *Journal of the American Statistical Association* 90: 443-50.
- Brader, Ted A., and Joshua A. Tucker. 2008. “Reflective and Unreflective Partisans? Experimental Evidence on the Links between Information, Opinion, and Party Identification.” New York University. Manuscript.
- Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. “What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat.” *American Journal of Political Science* 52: 959-78.
- Bullock, John G., Donald P. Green, and Shang E. Ha. 2008. “Experimental Approaches to Mediation: A New Guide for Assessing Causal Pathways.” Unpublished Manuscript, Yale University.
- Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. “Yes, But What’s the Mechanism? (Don’t Expect an Easy Answer).” *Journal of Personality and Social Psychology* 98: 550-8.
- Burdein, Inna, Milton Lodge, and Charles Taber. 2006. “Experiments on the Automaticity of Political Beliefs and Attitudes.” *Political Psychology* 27: 359-71.
- Campbell, Angus, Philip E. Converse, Warren Miller, and Donald Stokes. 1960. *The American Voter*. Chicago: University of Chicago Press.
- Clarke, Kevin A. 2009. “Return of the Phantom Menace: Omitted Variable Bias in Political

- Research.” *Conflict Management and Peace Science* 26: 46-66.
- Cohen, Geoffrey L. 2003. “Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs.” *Journal of Personality and Social Psychology* 85: 808-22.
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: HarperCollins.
- Fowler, James H., and Christopher T. Dawes. 2008. “Two Genes Predict Voter Turnout.” *Journal of Politics* 70: 579-94.
- Frangakis, Constantine E., and Donald B. Rubin. 2002. “Principal Stratification in Causal Inference.” *Biometrics* 58: 21-9.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
- Gerber, Alan S., and Donald P. Green. 2000. “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment.” *American Political Science Review* 94: 653-62.
- Gerber, Alan S., Gregory A. Huber, and Ebonya Washington. 2008. “Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment.” Unpublished manuscript, Yale University.
- Glynn, Adam N. 2009. “The Product and Difference Fallacies for Indirect Effects.” Unpublished manuscript, Harvard University.
- Green, Donald P., Shang E. Ha, and John G. Bullock. 2010. “Enough Already about ‘Black Box’ Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose.” *The Annals of the American Academy of Political and Social Sciences* 628: 200-8.
- Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects.” *Statistical Science* 25: 51-71.
- Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. In press. “Unpacking the Black Box: Learning about Causal Mechanisms from Experimental and Observational Studies.” Unpublished manuscript, Princeton University.
- Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62: 467-75.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News that Matters: Television and American Opinion*. Chicago: University of Chicago Press.
- James, Lawrence R. 1980. “The Unmeasured Variables Problem in Path Analysis.” *Journal of Applied Psychology* 65: 415-21.
- James, Lawrence R. 2008. “On the Path to Mediation.” *Organizational Research Methods* 11:

359-63.

- Judd, Charles M., and David A. Kenny. 1981. "Process Analysis: Estimating Mediation in Treatment Evaluations." *Evaluation Review* 5: 602-19.
- Kenny, David A. 2008. "Reflections on Mediation." *Organizational Research Methods* 11: 353-8.
- King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14: 131-59.
- LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76: 604-20.
- MacKinnon, David P., Chondra M. Lockwood, Jeanne M. Hoffman, Stephen G. West, and Virgil Sheets. 2002. "A Comparison of Methods to Test Mediation and Other Intervening Variable Effects." *Psychological Methods* 7: 83-104.
- Malhotra, Neil, and Jon A. Krosnick. 2007. "Retrospective and Prospective Performance Assessments during the 2004 Election Campaign: Tests of Mediation and News Media Priming." *Political Behavior* 29: 249-78.
- Miller, Joanne M., and Jon A. Krosnick. 2000. "News Media Impact on the Ingredients of Presidential Evaluations." *American Journal of Political Science* 44: 295-309.
- Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference*. New York: Cambridge University Press.
- Nelson, Thomas E. 2004. "Policy Goals, Public Rhetoric, and Political Attitudes." *Journal of Politics* 66: 581-605.
- Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91: 567-94.
- Pearl, Judea. 2010. "The Mediation Formula: A Guide to the Assessment of Causal Pathways in Non-Linear Models." Unpublished manuscript. Retrieved from <http://ftp.cs.ucla.edu/~kaoru/r363.pdf>.
- Petty, Richard E., and Duane T. Wegener. 1998. "Attitude Change: Multiple Roles for Persuasion Variables." In *The Handbook of Social Psychology*, eds. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey. 4th ed. New York: McGraw-Hill.
- Quiñones-Vidal, E., J. J. López-García, M. Peñaranda-Ortega, and F. Tortosa-Gil. 2004. "The Nature of Social and Personality Psychology as Reflected in *JPSP*, 1965-2000." *Journal of Personality and Social Psychology* 86: 435-52.
- Robins, James M. 2003. "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In *Highly Structured Stochastic Systems*, eds. Peter J. Green, Nils Lid

- Hjort, and Sylvia Richardson. New York: Oxford University Press.
- Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable that Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147: 656-66.
- Rubin, Donald B. 2004. "Direct and Indirect Causal Effects via Potential Outcomes." *Scandinavian Journal of Statistics* 31: 161-70.
- Spencer, Steven J., Mark P. Zanna, and Geoffrey T. Fong. 2005. "Establishing a Causal Chain: Why Experiments Are Often More Effective than Mediational Analyses in Examining Psychological Processes." *Journal of Personality and Social Psychology* 89: 845-51.
- Zanna, Mark P., and Joel Cooper. 1974. "Dissonance and the Pill: An Attribution Approach to Studying the Arousal Properties of Dissonance." *Journal of Personality and Social Psychology* 29: 703-9.

Afterword

36. Campbell's Ghost

Donald R. Kinder

Congratulations, intrepid reader! You have come a very long way. And after 34 chapters, many hundreds of pages of text, and cascades of footnotes, what more is there to say on the subject? Exactly. My remarks will be brief.

The conference that led to this volume began with remarks by Jamie Druckman, who reminded us how right and proper it was that Northwestern University host our gathering. Northwestern was Donald Campbell's home for more than three decades, and no one was more important than Campbell in bringing the philosophy, logic, and practice of experimental methods to the social sciences. In graduate school I was trained in experimental methods by Barry Collins, a student of Campbell's. For more than thirty years, first at Yale and later at Michigan, I have been teaching a graduate seminar on research design that draws heavily on Campbell's work. Those who survived this experience have taught similar courses elsewhere. Alterations and additions there have been along the way, of course, but the core ideas and the heart of the syllabus go back to Campbell. Throughout the conference, Campbell was never far from my mind, and he is back again now.¹

This is presumptuous of me to say, but I will go ahead and say it anyway: I think Campbell would have been thrilled with both the conference proceedings and the handbook you are doing your best to keep balanced in your hands. Surely he would have been delighted by the rapid rise to prominence of experimental methods in political science; by the sheer range and ingenuity in the application of experimental methods to

political questions; and by the display of sophistication in the statistical analyses of experimental data. For anyone with an interest in what an experimental political science might mean, the handbook is brimming with exemplary illustrations and excellent advice.

It was not always so. Thirty years ago, when Shanto Iyengar and I began to cook up our research on television news (published eventually in 1987 as *News that Matters*), we chose experiments as our principal method of inquiry. Trained as a social psychologist, experimentation is what I knew. Trained as a political scientist, Iyengar was simply ahead of his time. As a team, we gravitated more or less naturally to experimentation – but at the time, experiments were far from commonplace. When we began, experiments were seen by much of the political science establishment as exotic or irrelevant; experimentation was a subject not fit for serious discussion. *Experiments?* Experiments were what went on over in the chemistry building or in psychology laboratories – they had nothing to do with how political scientists conducted their business. The science of politics, according to the standard assumption of the time, could not be an experimental one.

Things change. Experiments are no longer eccentric (see Figure 36-1).ⁱⁱ Today, experimental results are published regularly, cited widely, and discussed in increasingly sophisticated ways (Druckman et al. 2006). The conference served to illustrate this progress and to mark the occasion. Those of us who have pushed for experimentation in political science are no longer a beleaguered insurgency; we have arrived. From time to time, the proceedings in Evanston understandably took on an air of festive celebration - in as much as political scientists are capable of such a thing.

I grew up in a small town in the Midwest. Celebration does not come easily to me. Predictably, I worried about us celebrating too much. I remembered Campbell's warning: to not expect too much from experiments. Those of us determined to march under the experimental banner, Campbell cautioned, should be prepared for a "poverty" of results (Campbell and Stanley 1966):

If, as seems likely, the ecology of our science is one in which there are available many more wrong responses than correct ones, we may anticipate that most experiments will be disappointing. . . . We must instill in our students the expectation of tedium and disappointment and the duty of thorough persistence, by now so well achieved in the biological and physical sciences (3).

Experimentation is a powerful method of testing – but it is no magic elixir. An experiment is no substitute for a good idea. An experiment cannot compensate for muddy conceptualization. It cannot do the work that must be done by measurement. Even the best experiment cannot answer a question poorly posed. We shouldn't ask too much of our experiments.

There is a larger and more important point here, I think. In our enthusiasm for experimental methods, we may be in danger of overlooking an important epistemological premise, one that is central to Campbell's work. Some years ago, thrown together in that best of all possible worlds, the Center for Advanced Study in the Behavioral Sciences, Tom Palfrey and I collaborated on an edited volume on experimentation for political science (Kinder and Palfrey 1993). We identified and reprinted a set of excellent applications of experimental methods to core problems in politics, and we wrote an introductory essay making the case for including experimentation within the methodological repertoire of modern political science.

The essay drew heavily on Campbell. This is perhaps a polite way to put it. For my part, I tried my best to *channel* Campbell. Palfrey and I argued that all methods are fallible. None can provide a royal road to truth. What to do in the face of this inescapable predicament? Campbell says: pursue research questions from a variety of methodological angles, each one fallible, but fallible in different ways. Dependable knowledge is grounded in no single method, but rather in convergent results across complementary methods. Hypotheses prove their mettle by surviving confrontation with a series of equally prudent but distinct tests.

Depending on a single method is risky business, epistemologically speaking. It is also a narrowing business. Relying exclusively on one method constricts the range of questions that seem worth pursuing. Which questions are interesting and which are not is seen through the filter of what one is able to do. Methodological preoccupations inevitably and insidiously shape substantive agendas. This is the “law of the instrument” at work (Kaplan 1964):

Give a small boy a hammer, and he will find that everything he encounters needs hammering. It comes as no particular surprise to discover that a scientist formulates problems in a way which requires for their solution just those techniques in which he himself is especially skilled (28).

It was on these grounds that Palfrey and I argued for expanding the methodological repertoire of political science to include experimentation. We did not envision the day when experiments would dominate the scientific study of politics. Such a goal seemed to us not only unrealistic but undesirable. Our goal was diversification. We believed that “a political science based on a variety of empirical methods,

experimentation prominent among them, is both within our reach and well worth reaching for ...” (Kinder and Palfrey 1993, 1, italics in the original).

I find it surprising, and not a little ironic, that I now find myself obliged to repeat that argument, but in reverse. Experimentation is a powerful tool. Indeed, for testing causal propositions, there is nothing quite like a well-designed experiment. The analysis of nonexperimental data to test causal claims, as Green and Gerber (2002) put it in their fine essay on field experiments, is informative only to the extent that “nature performs a convenient randomization on our behalf or that statistical correctives enable us to approximate the characteristics of an actual experiment” (808). Experiments have additional virtues: they enable the analytic decomposition of complex forces into component parts; they turn up “stubborn facts” that provoke theoretical invention; they apply to various levels of aggregation equally; and they accelerate interdisciplinary conversations (Kinder and Palfrey 1993 develop each of these points). I believe all this. Experiments are exceedingly useful – but they are not infallible.

Like all methods, experiments are accompanied by shortcomings as well as strengths. These include the inevitable ambiguity that surrounds the meaning of experimental treatments; the unsuitability of experimental methods to some of our core questions; and the hazards that inescapably attend generalizing from experimental results. All methods are fallible.

Going forward, this means taking Campbell’s insistence on multiple methods seriously. Going forward, it means that we should be aware not only of what experiments can tell us, but also what they cannot. Going forward, it means carrying out programs of research that incorporate methods in addition to experimentation. I think this is harder

than it might seem. It requires not just extra effort, but a special kind of thinking: moving back and forth between ever deeper conceptual analysis of the question under investigation, on the one hand, and judicious exploitation of particular methodological complementarities, on the other. Perhaps an example or two of what I think we should be up to will help clarify the point I am straining to make.

Given multiple sources of cultural difference in society, when and why does one difference become the basis for political competition and conflict rather than another? Posner's (2004, 2005) answer to this important and often overlooked question hinges on the degree to which cultural groups are useful vehicles for political coalition building. To test his theory, Posner takes clever advantage of a natural experiment, the drawing of the border between the African nations of Zambia and Malawi by the British South African Company in 1891 – a border drawn for purely administration purposes, with no attention to the geography of the indigenous peoples living nearby. The differences that Posner observes associated with living on one side of the border as against the other support his account, but – and this is the point I want to stress here – the results of the natural experiment are much more convincing when fortified by convergent results emanating from Posner's analysis of campaigns and election returns.

A second example, equally excellent, comes from Winter's (2008) research on "dangerous frames." Winter is interested in the intersection between elite rhetoric and public opinion. He shows that politicians, interest groups, and parties routinely frame issues in ways that prime audiences to respond not only to the issue at hand but also to the way the frames resonate with deeply held beliefs about race and gender. Winter establishes this point with well-designed, subtle experiments *and* with sharp analysis of

carefully chosen national survey data. Again, Winter is much more convincing (in my book) on how everyday ideas about gender and race become implicated in the public's views on matters of policy for having provided two kinds of empirical demonstrations, not just one.

* * * * *

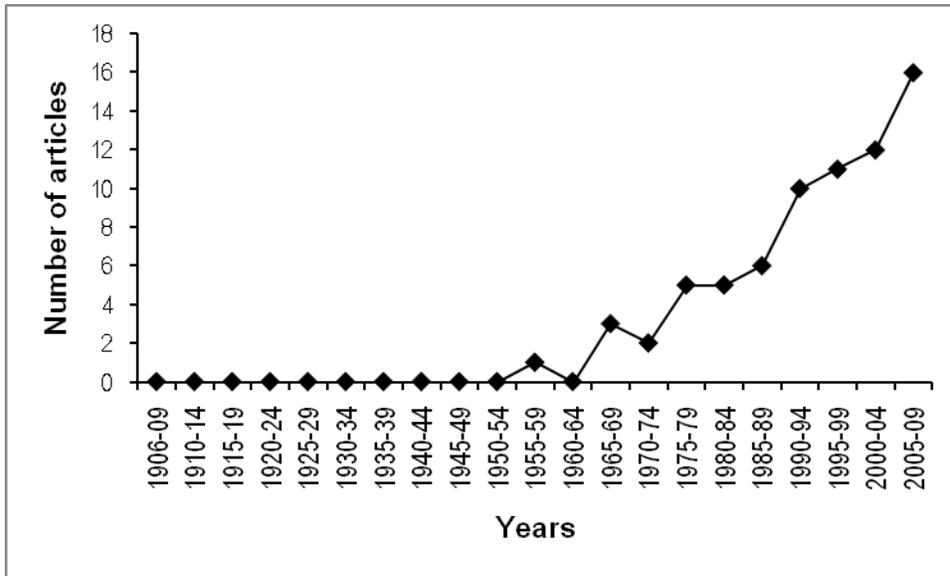
“Politics is an observational, not an experimental science.” So said A. Lawrence Lowell in his presidential address to the American Political Science Association in 1909.ⁱⁱⁱ One hundred years later, this is no longer so. Now we would say – we would be required to say – “Politics is an observational science *and* an experimental science.” No doubt Campbell would be pleased.

References

- Campbell, Donald T. 1957. “Factors Relevant to the Validity of Experiments in Social Settings.” *Psychological Bulletin* 54: 297-312.
- Campbell, Donald T. 1966. “Pattern Matching as an Essential in Distal Knowing.” In *The Psychology of Egon Brunswik*, ed. Kenneth R. Hammond. New York: Holt, Rinehart and Winston.
- Campbell, Donald T. 1969. “Reforms as Experiments.” *American Psychologist* 24: 409-429.
- Campbell, Donald T. 1970. “Natural Selection as an Epistemological Model.” In *A Handbook of Methods in Cultural Anthropology*, eds. Raoul Naroll, and Ronald Cohen. New York: Natural History Press.
- Campbell, Donald T. 1974. “Evolutionary Epistemology.” In *The Philosophy of Karl Popper*, ed. Paul Arthur Schilpp. La Salle, IL: Open Court Press.
- Campbell, Donald T. 1975. “‘Degrees of Freedom’ and the Case Study.” *Comparative Political Studies* 8: 178-93.
- Campbell, Donald T., and Donald W. Fiske. 1959. “Convergent and Discriminant Validation by the Mutitrait-Multimethod Matrix.” *Psychological Bulletin* 56: 81-105.

- Campbell, Donald T., and H. L. Ross. 1968. "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis." *Law and Society Review* 3: 33-53.
- Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis for Field Settings*. Chicago: Rand McNally.
- Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100: 627-35.
- Green, Donald P., and Alan S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: State of the Discipline*, eds. Ira Katznelson, and Helen V. Milner. New York: Norton.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News that Matters: Television and American Opinion*. Chicago: University of Chicago Press.
- Kaplan, Abraham. 1964. *The Conduct of Inquiry*. San Francisco: Chandler Press.
- Kinder, Donald R., and Thomas R. Palfrey. 1993. "On Behalf of an Experimental Political Science." In *Experimental Foundations of Political Science*, eds. Donald R. Kinder, and Thomas R. Palfrey. Ann Arbor: University of Michigan Press.
- Posner, Daniel N. 2004. "The Political Salience of Cultural Differences: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98: 529-45.
- Posner, Daniel N. 2005. *Institutions and Ethnic Politics in Africa*. New York: Cambridge University Press.
- Riecken, Henry W., and Robert F. Boruch. 1974. *Social Experimentation: A Method for Planning and Evaluating Social Innovations*. New York: Academic Press.
- Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee B. Sechrest. 1966. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.
- Winter, Nicholas J. G. 2008. *Dangerous Frames. How Ideas about Race and Gender Shape Public Opinion*. Chicago: University of Chicago Press.

Figure 36-1. Number of Articles Featuring Experiments Published in *The American Political Science Review*, 1906-2009



ⁱ A previous version of this paper was presented at the Experimentation in Political Science Conference at Northwestern University, May 28-29, 2009. We thank Jamie Druckman, Alan Jacobs, Scott Matthews, Nora Ng, David Nickerson, Judea Pearl, Dustin Tingley, and Lynn Vavreck for comments. Particular thanks to Donald Green, our co-author on several related papers, for many helpful discussions.

ⁱⁱ Quiñones-Vidal et al. (2004) show that the article is already the best-cited in the history of the *Journal of Personality and Social Psychology*. Our own search turns up more than 19,000 citations (<http://scholar.google.com/scholar?cites=13824678930131800739>). Analogous searches suggest that Downs (1957) has been cited fewer than 14,000 times and that *The American Voter* (Campbell et al. 1960) has been cited fewer than 4,000 times.

ⁱⁱⁱ This discussion of direct and indirect effects elides a subtle but important assumption: the effect of M on Y is the same regardless of the value of X . This additivity or “no-interaction” assumption is implied in linear models, e.g., Equation 3. See Robins (2003, 76-77) for a detailed consideration.

^{iv} Imai, Keele, and Yamamoto (in press, 3) show that the indirect effect ab is identified under the assumption of sequential ignorability, i.e., independence of X from the potential outcomes of M and Y , and independence of M from the potential outcomes of Y . This is a stronger identifying assumption than $cov(e_1, e_3) = 0$ (Imai, Keele, and Yamamoto in press, 10), but it has the virtue of being grounded in a potential-outcomes framework.

^v An occasional defense of the Baron-Kenny method is that the method itself is unbiased: the problem lies in its application to nonexperimental data, and it would vanish if the method were applied to studies in which both X and M are randomized. This is incorrect. In fact, when both X and M are randomized, the Baron-Kenny method calls for researchers to conclude that M does not mediate X even when M strongly mediates X . For details, see Bullock et al. (2008, 10-11).

^{vi} This warning is absent from Baron and Kenny (1986), but it appears clearly in one of that article’s predecessors, which notes that what would come to be known as the Baron-Kenny procedure is “likely to yield biased estimates of causal parameters . . . even when a randomized experimental research design has been used” (Judd and Kenny 1981, 801, emphasis in original).

^{vii} See Angrist et al. (1996) for a thorough discussion of the conditions that a variable must satisfy to be an instrument for another variable.

^{viii} We do not take up two other limitations. One is the unreliability of instrumental-variable approaches to mediation in nonlinear models (Pearl 2010). The other is the “weak instruments” problem: when instruments are weakly correlated with the endogenous variables, IV estimators have large standard errors and even slight violations of the exclusion restriction ($\text{cov}[Z, e_3] = 0$ where Z is the instrument for the endogenous mediator) may cause the estimator to have a large asymptotic bias (Bartels 1991; Bound, Jaeger, and Baker 1995). This is a large concern in econometric studies, where instruments are often weak and exclusion-restriction violations likely. But instruments that are specifically created by random assignment to affect endogenous mediators are likely to meet the exclusion restriction and unlikely to be “weak” by econometric standards.

^{ix} Econometric convention permits the use of multiple instruments to simultaneously identify the effects of a single endogenous variable. But estimators based on multiple instruments have no clear causal interpretation in a potential-outcomes framework; they are instead difficult-to-interpret mixtures of local average treatment effects (Morgan and Winship 2007, 212). This is why we recommend that experimenters create interventions that isolate individual mediators.

^x The same problem holds if we express the indirect effect as $c - d$ rather than ab .

^{xi} This is only one aspect of Cohen (2003). So far as mediation is concerned, Cohen’s main suggestion is that reference-group cues affect policy attitudes not by changing the extent to which people think systematically about policy information but by otherwise changing perceptions of the policies under consideration.

^{xii} This is exactly the approach that Angrist, Lavy, and Schlosser (in press) take in their study of family size and the long-term welfare of children.

ⁱ Campbell wrote on an astonishing range of important subjects and left a mark on half a dozen disciplines. Within the treasure trove that was Campbell’s contribution to methodological excellence are seminal papers on the logic of experimentation (Campbell 1957; Campbell and Stanley 1966); measurement (Campbell and Fiske 1959; Webb et al. 1966); types of validity (Cook and Campbell 1979, chapter 1); the design and analysis of quasi-experiments (Campbell and Ross 1968; Cook and Campbell 1979); case studies in comparative politics (Campbell 1975); the experimenting society (Campbell 1969; Riecken and Boruch 1974); and the philosophy of science (Campbell 1966, 1970, 1974). Campbell died in 1996 at the age of 79.

ⁱⁱ Figure 36-1 updates a figure presented by Druckman and colleagues in their essay on experimentation for the 100th anniversary issue of the *American Political Science Review* (Druckman et al. 2006).

ⁱⁱⁱ Lowell’s speech quoted by Druckman et al. (2006, 627).