CHAPTER 17

Program Evaluation: Principles, Procedures, and Practices

Aurelio José Figueredo, Sally Gayle Olderbak, Gabriel Lee Schlomer, Rafael Antonio Garcia, *and* Pedro Sofio Abril Wolf

Abstract

This chapter provides a review of the current state of the principles, procedures, and practices within program evaluation. We address a few incisive and difficult questions about the current state of the field: (1) What are the kinds of program evaluations? (2) Why do program evaluation results often have so little impact on social policy? (3) Does program evaluation suffer from a counterproductive system of incentives? and (4) What do program evaluators actually do? We compare and contrast the merits and limitations, strengths and weaknesses, and relative progress of the two primary contemporary movements within program evaluation, Quantitative Methods and Qualitative Methods, and we propose an epistemological framework for integrating the two movements as complementary forms of investigation, each contributing to different stages in the scientific process. In the final section, we provide recommendations for systemic institutional reforms addressing identified structural problems within the real-world practice of program evaluation.

Key Words: Program evaluation, formative evaluation, summative evaluation, social policy, moral hazards, perverse incentives, quantitative methods, qualitative methods, context of discovery, context of justification

Introduction

President Barack Obama's 2010 Budget included many statements calling for the evaluation of more U. S. Federal Government programs (Office of Management and Budget, 2009). But what precisely is *meant* by the term *evaluation*? Who should *conduct* these evaluations? Who should *pay* for these evaluations? *How* should these evaluations be conducted?

This chapter provides a review of the principles, procedures, and practices within *program evaluation*. We start by posing and addressing a few incisive and difficult questions about the current state of that field:

1. What are the different kinds of program evaluations?
2. Why do program evaluation results often have so little impact on social policy?
3. Does program evaluation suffer from a counterproductive system of incentives?

We then ask a fourth question regarding the real-world practice of program evaluation: What do program evaluators actually do? In the two sections that follow, we try to answer this question by reviewing the merits and limitations, strengths and weaknesses, and relative progress of the two primary contemporary "movements" within program evaluation and the primary methods of evaluation upon which they rely: Part 1 addresses Quantitative Methods and Part 2 addresses Qualitative Methods. Finally, we propose a framework for the integration of the two movements as complementary forms of investigation in program evaluation, each contributing to different stages in the scientific process. In the final section, we provide recommendations for systemic institutional reforms addressing identified structural problems within the real-world practice of program evaluation.

What Are the Different Kinds of Program Evaluations?

Scriven (1967) introduced the important distinction between *summative* program evaluations as compared with *formative* program evaluations. The goal of a *summative evaluation* is to judge the merits of a fixed, unchanging program as a finished product, relative to potential alternative programs. This judgment should consist of an analysis of the costs and benefits of the program, as compared with other programs targeted at similar objectives, to justify the expenses and opportunity costs society incurs in implementing one particular program as opposed to an alternative program, as well as in contrast to doing nothing at all. Further, a summative evaluation must examine both the intended and the unintended outcomes of the programmatic intervention and not just the specific stated goals, as represented by the originators, administrators, implementers, or advocates of the program (Scriven, 1991). A *formative evaluation*, on the other hand, is an ongoing evaluation of a program that is not fixed but is still in the process of change. The goal of a formative evaluation is to provide feedback to the program managers with the purpose of improving the program regarding what is and what is not working well and not to make a final judgment on the relative merits of the program.

The purely dichotomous and mutually exclusive model defining the differences between summative and formative evaluations has been softened and qualified somewhat over the years. Tharp and Gallimore (1979, 1982), in their research and development (R& D) program for social action, proposed a model of *evaluation succession*, patterned on the analogy of *ecological succession*, wherein an ongoing, long-term evaluation begins as a formative program evaluation and acquires features of a summative program evaluation as the program naturally matures, aided by the continuous feedback from the formative program evaluation process. Similarly, Patton (1996) has proposed a putatively broader view of program evaluation that falls between the summative versus formative dichotomy: (1) knowledge-generating evaluation, evaluations that are designed to increase our conceptual understanding of a particular topic; (2) developmental evaluation, an ongoing evaluation that strives to continuously improve the program; and (3) using the evaluation processes, which involves more intently engaging the stakeholders, and others associated with the evaluation, to think more about the program and ways to improve its efficacy or effectiveness. Patton has argued that the distinction between summative and formative evaluation is decreasing, and there is a movement within the field of program evaluation that applies a more creative use and application of evaluation. What he termed *knowledge-generative evaluation* is a form of evaluation focused not on the instrumental use of evaluation findings (e.g., making decisions based on the results of the evaluation) but, rather, on the conceptual use of evaluation findings (e.g., theory construction).

A *developmental evaluation* (Patton, 1994) is a form of program evaluation that is ongoing and is focused on the development of the program. Evaluators provide constant feedback but not always in the forms of official reports. Developmental evaluation assumes components of the program under evaluation are constantly changing, and so the evaluation is not geared toward eventually requiring a summative program evaluation but, rather, is focused on constantly adapting and evolving the evaluation to fit the evolving program. Patton (1996) proposed that program evaluators should focus not only on reaching the evaluation outcomes, but also on the process of the evaluation itself, in that the evaluation itself can be "participatory and empowering . . . increasing the effectiveness of the program through the evaluation process rather than just the findings" (p. 137).

Stufflebeam (2001) has presented a larger classification of the different kinds of evaluation, consisting of 22 alternative approaches to evaluation that can be classified into four categories. Stufflebeam's first category is called *Pseudoevaluations* and encompasses evaluation approaches that are often motivated by politics, which may lead to misleading or invalid results. Pseudoevaluation approaches include: (1) Public Relations-Inspired Studies and (2) Politically Controlled Studies (for a description of each of the 22 evaluation approaches, please refer to Stufflebeam's [2001] original paper). Stufflebeam's second category is called *Questions-And-Methods-Evaluation Approaches* (Quasi-Evaluation Studies) and encompasses evaluation approaches geared to address a particular question, or apply a particular method, which often result in narrowing the scope of the evaluation. This category includes: (3) Objectives-Based Studies; (4) Accountability, Particularly Payment by Results Section; (5) Objective Testing Program; (6) Outcome Evaluation as Value-Added Assessment; (7) Performance Testing; (8) Experimental Studies; (9) Management Information Systems; (10) Benefit–Cost Analysis Approach; (11) Clarification Hearing; (12) Case Study Evaluations; (13) Criticism and Commentary; (14) Program Theory-Based Evaluation; and (15) Mixed-Methods Studies.

Stufflebeam's (2001) third category, *Improvement/Accountability-Oriented Evaluation Approaches*, is the most similar to the commonly used definition of program evaluation and encompasses approaches that are extensive and expansive in their approach and selection of outcome variables, which use a multitude of qualitative and quantitative methodologies for assessment. These approaches include: (16) Decision/Accountability-Oriented Studies; (17) Consumer-Oriented Studies; and (18) Accreditation/Certification Approach. Stufflebeam's fourth category is called *Social Agenda/Advocacy Approaches* and encompasses evaluation approaches that are geared toward directly benefitting the community in which they are implemented, sometimes so much so that the evaluation may be biased, and are heavily included by the perspective of the stakeholders. These approaches include: (19) Client-Centered Studies (or Responsive Evaluation); (20) Constructivist Evaluation; (21) Deliberative Democratic Evaluation; and (22) Utilization-Focused Evaluation.

These different types of program evaluations are not exhaustive of all the types that exist, but they are the ones that we consider most relevant to the current analysis and ultimate recommendations.

Why Do Program Evaluation Results Often Have So Little Impact on Social Policy?

At the time of writing, the answer to this question is not completely knowable. Until we have more research on this point, we can never completely document the impact that program evaluation has on public policy. Many other commentators on program evaluation (e.g., Weiss, 1999), however, have made the point that program evaluation does not have as much of an impact on *social policy* as we would like it to have. To illustrate this point, we will use two representative case studies: the *Kamehameha Early Education Project (KEEP)*, and the *Drug Abuse Resistance Education (DARE)*. Although the success or failure of a program and the success or failure of a program evaluation are two different things, one is intimately related to the other, because the success or failure of the program evaluation is necessarily considered in reference to the success or failure of the program under evaluation.

The Frustrated Goals of Program Evaluation

When it comes to public policy, the goal of an evaluation should include helping funding agencies, such as governmental entities, decide whether to terminate, cut back, continue, scale up, or disseminate a program depending on success or failure of the program, which would be the main goal of a summative program evaluation. An alternative goal might be to suggest modifications to existing programs in response to data gathered and analyzed during an evaluation, which would be the main goal of a formative program evaluation. Although both goals are the primary purposes of program evaluation, in reality policymakers rarely utilize the evaluation findings for these goals and rarely make decisions based on the results of evaluations. Even an evaluation that was successful in its process can be blatantly ignored and result in a failure in its outcome. We relate this undesirable state of affairs further below with the concept of a *market failure* from economic theory.

According to Weiss (1999) there are four major reasons that program evaluations may not have a direct impact on decisions by policymakers (the "Four I's"). First, when making decisions, a host of competing *interests* present themselves. Because of this competition, the results of different evaluations can be used to the benefit or detriment of the causes of various interested parties. Stakeholders with conflicting interests can put the evaluator between a rock and a hard place. An example of this is when a policymaker receives negative feedback regarding a program. On the one hand, the policymaker is interested in supporting successful programs, but on the other hand, a policymaker who needs to get re-elected might not want to be perceived as "the guy who voted no on drug prevention." Second, the *ideologies* of different stakeholder groups can also be a barrier for the utilization of program evaluation results. These ideologies filter potential solutions and restrict results to which policymakers will listen. This occurs most often when the ideology claims that something is "fundamentally wrong."

For example, an abstinence-only program, designed to prevent teenage pregnancy, may be in competition with a program that works better, but because the program passes out condoms to teenagers, the abstinence-only plan may be funded because of the ideologies of the policymakers or their constituents. Third, the *information* contained in the evaluation report itself can be a barrier. The results of evaluations are not the only source of information and are often not the most salient. Policymakers often have extensive information regarding a potential policy, and the results of the evaluation are competing with every other source of information that can enter the decision-making process. Finally, the *institutional* characteristics of the program itself can become a barrier. The institution is made up of people working within the context of a set structure and a history of behavior. Because of these institutional characteristics, change may be difficult or even considered "off-limits." For example, if an evaluation results in advocating the elimination a particular position, then the results may be overlooked because the individual currently in that position is 6 months from retirement. Please note that we are not making a value judgment regarding the relative merits of such a decision but merely describing the possible situation.

The utilization of the results of an evaluation is the primary objective of an evaluation; however, it is often the case that evaluation results are put aside in favor of other, less optimal actions (Weiss, 1999). This is not a problem novel to program evaluators but a problem that burdens most applied social science. A prime example of this problem is that of the reliability of eyewitness testimony. Since Elizabeth Loftus published her 1979 book, *Eyewitness Testimony*, there has been extensive work done on the reliability of eyewitnesses and the development of false memories. Nevertheless, it took 20 years for the U. S. Department of Justice to institute national standards reflecting the implications of these findings (Wells et al., 2000). Loftus did accomplish what Weiss refers to as "enlightenment" (Weiss, 1980), or the bringing of scientific data into the applied realm of policymaking. Although ideally programs would implement evaluation findings immediately, this simply does not often happen. As stated by Weiss (1999), the volume of information that organizations or policymakers have regarding a particular program is usually too vast to be overthrown by one dissenting evaluation. These problems appear to be inherent in social sciences and program evaluation, and it is unclear how to ameliorate them.

To illustrate how programs and program evaluations can succeed or fail, we use two representative case studies: one notable *success* of the program evaluation process, the *KEEP*, and one notable *failure* of the program evaluation process, *DARE*.

Kamehameha Early Education Project

A classic example of a successful program evaluation described by Tharp and Gallimore (1979) was that of KEEP. Kamehameha Early Evaluation Project was started in 1970 to improve the reading and general education of Hawaiian children. The project worked closely with program evaluators to identify solutions for many of the unique problems faced by young Hawaiian-American children in their education, from kindergarten through third grade, and to discover methods for disseminating these solutions to the other schools in Hawaii. The evaluation took 7 years before significant improvement was seen and involved a multidisciplinary approach, including theoretical perspectives from the fields of psychology, anthropology, education, and linguistics.

Based on their evaluation of KEEP, Tharp and Gallimore (1979) identified four necessary conditions for a successful program evaluation: (1) longevity—evaluations need time to take place, which requires stability in other areas of the program; (2) stability in the values and goals of the program; (3) stability of funding; and (4) the opportunity for the evaluators' recommendations to influence the procedure of the program.

In terms of the "Four I's," the interests of KEEP were clear and stable. The project was interested in improving general education processes. In terms of ideology and information, KEEP members believed that the evaluation process was vital to its success and trusted the objectivity of the evaluators, taking their suggestions to heart. From its inception, the institution had an evaluation system built in. Since continuing evaluations were in process, the program itself had no history of institutional restriction of evaluations.

Drug Abuse Resistance Education

In this notable case, we are not so much highlighting the failure of a specific program evaluation, or of a specific program *per se*, as highlighting the institutional failure of program evaluation as a system, at least as currently structured in our society. In the case of DARE, a series of program evaluations produced results that, in the end, were not acted upon. Rather, what should have been recognized as a failed program lives on to this day. The DARE program was started in 1983, and the goal of the program was to prevent drug use. Although there are different DARE curricula, depending on the targeted age group, the essence of the program is that uniformed police officers deliver a curriculum in safe classroom environments aimed at preventing drug use among the students. As of 2004, DARE has been the most successful school-based prevention program in attracting federal money: The estimated average federal expenditure is three-quarters of a billion dollars per year (West & O'Neal, 2004). Although DARE is successful at infiltrating school districts and attracting tax dollars, research spanning more than two decades has shown that the program is ineffective at best and detrimental at worst. One of the more recent meta-analyses (West & O'Neal, 2004) estimated the average effect size for DARE's effectiveness was extremely low and not even statistically significant ($r = 0.01$; Cohen's $d = 0.02$, 95% confidence interval = –0.04, 0.08).

Early studies pointed to the ineffectiveness of the DARE program (Ennett, Tobler, Ringwalt, & Flewelling, 1994; Clayton, Cattarello, & Johnstone, 1996; Dukes, Ullman, & Stein, 1996). In response to much of this research, the Surgeon General placed the DARE program in the "Does Not Work" category of programs in 2001. In 2003, the U. S. Government Accountability Office (GAO) wrote a letter to congressmen citing a series of empirical studies in the 1990s showing that in some cases DARE is actually iatrogenic, meaning that DARE does more harm than good.

Despite all the evidence, DARE is still heavily funded by tax dollars through the following government agencies: California National Guard, Combined Federal Campaign (CFC), Florida National Guard, St. Petersburg College, Multijurisdictional, Counterdrug Task Force Training, Indiana National Guard, Midwest Counterdrug Training Center/ National Guard, U. S. Department of Defense, U. S. Department of Justice, Bureau of Justice Assistance (BJA), Drug Enforcement Administration, Office of Juvenile Justice and Delinquency Prevention, and the U. S. Department of State.

These are institutional conflicts of interest. As described above, few politicians want to be perceived as "the guy who voted against drug prevention." The failure of DARE stems primarily from these conflicts of interest. In lieu of any better options, the U. S. Federal Government continues to support DARE, simply because to not do so might appear as if they were doing nothing. At the present writing in 2012, DARE has been in effect for 29 years. Attempting to change the infrastructure of a longstanding program like this would be met with a great deal of resistance.

We chose the DARE example specifically because it is a long-running example, as it takes years to make the determination that somewhere something in the system of program evaluation failed. If this chapter were being written in the early 1990s, people in the field of program evaluation might reasonably be predicting that based on the data available, this program should either be substantially modified or discontinued. Rather, close to two decades later and after being blacklisted by the government, it is still a very well-funded program. One may argue that the program evaluators themselves did their job; however, what is the use of program evaluation if policymakers are not following recommendations based on data produced by evaluations? Both the scientific evidence and the anecdotal evidence seem to suggest that programs with evaluations built-in seem to result in better utilization of evaluation results and suggestions. This may partly result from better communication between the evaluator and the stakeholders, but if the evaluator is on a first-name basis (or maybe goes golfing) with the stakeholders, then what happens to his/her ability to remain objective? We will address these important issues in the sections that immediately follow by exploring the extant system of incentives shaping the practice of program evaluation.

What System of Incentives Governs the Practice of Program Evaluation?

Who Are Program Evaluators?

On October 19, 2010, we conducted a survey of the brief descriptions of qualifications and experience of evaluators posted by program evaluators (344 postings in total) under the "Search Resumes" link on the American Evaluation Association (AEA) website (http://www.eval.org/find_an_evaluator/evaluator_search.asp). Program evaluators' skills were evenly split in their levels of quantitative (none: 2.0%; entry: 16.9%; intermediate: 37.5%; advanced: 34.9%; expert: 8.4%; strong: 0.3%) and qualitative evaluation experience (none: 1.5%; entry: 17.2%; intermediate: 41.6%; advanced: 27.3%; expert: 12.2%; strong: 0.3%). Program evaluators also expressed a range of years they were involved with evaluation (<1 year: 12.5%; 1–2 years:

20.1%; 3–5 years: 24.1%; 6–10 years: 19.8%; >10 years: 23.5%).

In general, program evaluators were highly educated, with the highest degree attained being either a masters (58.8%) or a doctorate of some sort (36%), and fewer program evaluators had only an associates (0.3%) or bachelors degree (5.0%). The degree specializations were also widely distributed. Only 12.8% of the program evaluators with posted resumes described their education as including some sort of formal training specifically in evaluation. The most frequently mentioned degree specialization was in some field related to Psychology (25.9%), including social psychology and social work. The next most common specialization was in Education (15.4%), followed by Policy (14.0%), Non-Psychology Social Sciences (12.2%), Public Health or Medicine (11.6%), Business (11.3%), Mathematics or Statistics (5.8%), Communication (2.9%), Science (2.3%), Law or Criminal Justice (2.0%), Management Information Systems and other areas related to Technology (1.5%), Agriculture (1.5%), and Other, such as Music (1.7%).

For Whom Do Program Evaluators Work?

We sampled job advertisements for program evaluators using several Internet search engines: usajobs.gov, jobbing.com, and human resources pages for government agencies such as National Institutes of Health (NIH), the National Institute of Mental Health (NIMH), Centers for Disease Control (CDC), and GAO. Based on this sampling, we determined there are four general types of program evaluation jobs.

Many agencies that deliver or implement social programs organize their own program evaluations, and these account for the first, second, and third types of program evaluation jobs available. The first type of program evaluation job is obtained in response to a call or request for proposals for a given evaluation. The second type of program evaluation job is obtained when the *evaluand* (the program under evaluation) is asked to hire an internal program evaluator to conduct a summative evaluation. The third general type of program evaluation job is obtained when a program evaluator is hired to conduct a formative evaluation; this category could include an employee of the evaluand who serves multiple roles in the organization, such as secretary and data collector.

We refer to the fourth type of program evaluation job as the Professional Government Watchdog. That type of evaluator works for an agency like the GAO. The GAO is an independent agency that answers directly to Congress. The GAO has 3300 workers (http://www.gao.gov/about/workforce/) working in roughly 13 groups: (1) Acquisition and Sourcing Management; (2) Applied Research and Methods; (3) Defense Capabilities and Management; (4) Education, Workforce, and Income Security; (5) Financial Management and Assurance; (6) Financial Markets and Community Investment; (7) Health Care; (8) Homeland Security and Justice; (9) Information Technology; (10) International Affairs and Trade; (11) Natural Resources and Environment; (12) Physical Infrastructure; and (13) Strategic Issues. Each of these groups is tasked with the oversight of a series of smaller agencies that deal with that group's content. For example, the Natural Resources and Environment group oversees the Department of Agriculture, Department of Energy, Department of the Interior, Environmental Protection Agency, Nuclear Regulatory Commission, Army Corps of Engineers, National Science Foundation, National Marine Fisheries Service, and the Patent and Trademark Office.

With the many billions of dollars being spent by the U. S. government on social programs, we sincerely doubt that 3300 workers can possibly process all the program evaluations performed for the entire federal government. Recall that the estimated average federal expenditure for DARE alone is three-quarters of a billion dollars per year and that this program has been supported continuously for 17 years. We believe that such colossal annual expenditures should include enough to pay for a few more of these "watchdogs" or at least justify the additional expense of doing so.

Who Pays the Piper?

The hiring of an internal program evaluator for the purpose of a summative evaluation is a recipe for an ineffective evaluation. There is a danger that the program evaluator can become what Scriven (1976, 1983) has called a *program advocate*. According to Scriven, these program evaluators are not necessarily malicious but, rather, could be biased as a result of the nature of the relationship between the program evaluator, the program funder, and the program management. The internal evaluator is generally employed by, and answers to, the management of the program and not directly to the program funder. In addition, because the program evaluator's job relies on the perceived "success" of the evaluation, there is an incentive to bias the results in favor of the program being evaluated. Scriven has argued that this structure may develop divided loyalties between the program being evaluated and the agency funding the program (Shadish, Cook, & Leviton, 1991). Scriven (1976, 1983) has recommended that summative evaluations are necessary for a society to optimize resource allocation but that we should also periodically re-assign program evaluators to different program locations to prevent individual evaluators from being co-opted into local structures. The risks of co-opting are explained in the next section.

Moral Hazards and Perverse Incentives

As a social institution, the field of program evaluation has professed very high ethical standards. For example, in 1994 The Joint Committee on Standards for Educational Evaluation produced the Second Edition of an entire 222-page volume on professional standards in program evaluation. Not all were what we would typically call *ethical* standards *per se*, but one of the four major categories of professional evaluation standards was called *Propriety Standards* and addressed what most people would refer to as ethical concerns. The other three categories were denoted *Utility Standards, Feasibility Standards,* and *Accuracy Standards*. Although it might be argued that a conscientious program evaluator is ethically obligated to carefully consider the utility, feasibility, and accuracy of the evaluation, it is easy to imagine how an occasional failure in any of these other areas might stem from factors other than an ethical lapse.

So why do we need any protracted consideration of *moral hazards* and *perverse incentives* in a discussion of program evaluation? We should make clear at the outset that we do not believe that most program evaluators are immoral or unethical. It is important to note that in most accepted uses of the term, the expression *moral hazard* makes no assumptions, positive or negative, about the relative moral character of the parties involved, although in some cases the term has unfortunately been used in that pejorative manner. The term *moral hazard* only refers (or *should* only refer) to the structure of perverse incentives that constitute the particular hazard in question (Dembe & Boden, 2000). We wish to explicitly avoid the implication that there are immoral or unethical individuals or agencies out there that intentionally corrupt the system for their own selfish benefit. Unethical actors hardly need moral hazards to corrupt them: They are presumably already immoral and can therefore be readily corrupted, presumably with little provocation. It is the normally moral or ethical people about which we need to worry under the current system of incentives, because this system may actually penalize them for daring to do the right thing for society.

Moral hazards and *perverse incentives* refer to conditions under which the incentive structures in place tend to promote socially undesirable or harmful behavior (e.g., Pauly, 1974). Economic theory refers the socially undesirable or harmful consequences of such behavior as *market failures*, which occur when there is an inefficient allocation of goods and services in a market. Arguably, continued public or private funding of an ineffective or harmful social program therefore constitutes a market failure, where the social program is conceptualized as the product that is being purchased. In economics, one of the well-documented causes of market failures is incomplete or incorrect information on which the participants in the market base their decisions. That is how these concepts may relate to the field of program evaluation.

One potential source of incomplete or incorrect information is referred to in economic theory as that of *information asymmetry*, which occurs in economic transactions where one party has access to either more or better information than the other party. Information asymmetry may thus lead to moral hazard, where one party to the transaction is insulated from the adverse consequences of a decision but has access to more information than another party (specifically, the party that is *not* insulated from the adverse consequences of the decision in question). Thus, moral hazards are produced when the party with *more* information has an incentive to act contrary to the interests of the party with *less* information. Moral hazard arises because one party does not risk the full consequences of its own decisions and presumably acquires the tendency to act less cautiously than otherwise, leaving another party to suffer the consequences of those possibly ill-advised decisions.

Furthermore, a *principal-agent problem* might also exist where one party, called an *agent*, acts on behalf of another party, called the *principal.* Because the principal usually cannot completely monitor the agent, the situation often develops where the agent has access to more information than the principal does. Thus, if the interests of the agent and the principal are not perfectly consistent and mutually aligned with each other, the agent may have an incentive to behave in a manner that is contrary to the interests of the principal. This is the problem of *perverse incentives*, which are incentives that have unintended and undesirable effects ("unintended consequences"), defined as being against the interests of the party providing the incentives (in this case, the principal). A market failure becomes more than a mere mistake and instead becomes the inevitable product of a conflict of interests between the principal and the agent. A conflict of interests may lead the agent to manipulate the information that they provide to the principal. The information asymmetry thus generated will then lead to the kind of market failure referred to as *adverse selection*. Adverse selection is a market failure that occurs when information asymmetries between buyers and sellers lead to suboptimal purchasing decisions on the part of the buyer, such as buying worthless or detrimental goods or services (perhaps like DARE?).

When applying these economic principles to the field of program evaluation, it becomes evident that because program evaluators deal purely in information, and this information might be manipulated—either by them or by the agencies for which they work (or both of them in implicit or explicit collusion)—we have a clear case of *information asymmetry*. This information asymmetry, under *perverse incentives*, may lead to a severe *conflict of interests* between the society or funding agency (the principal) and the program evaluator (the agent). This does not mean that the agent must perforce be corrupted, but the situation does create a moral hazard for the agent, regardless of any individual virtues. If the perverse incentives are acted on (meaning they indeed elicit the execution of impropriety), then it is clearly predicted by economic theory to produce a market failure and specifically adverse selection on the part of the principal.

Getting back to the question of the professional standards actually advocated within program evaluation, how do these lofty ideals compare to the kind of behavior that might be expected under moral hazards and perverse incentives, presuming that program evaluators are subject to the same kind of motivations, fallibilities, and imperfections as the rest of humanity? The Joint Committee on Standards for Educational Evaluation (1994) listed the following six scenarios as examples of conflicts of interest:

> – Evaluators might benefit or lose financially, long term or short term, depending on what evaluation results they report, especially if the evaluators are connected financially to the program being evaluated or to one of its competitors.
>
> – The evaluator's jobs and/or ability to get future evaluation contracts might be influenced by their reporting of either positive or negative findings.
>
> – The evaluator's personal friendships or professional relationships with clients may influence the design, conduct, and results of an evaluation.
>
> – The evaluator's agency might stand to gain or lose, especially if they trained the personnel or developed the materials involved in the program being evaluation.
>
> – A stakeholder or client with a personal financial interest in a program may influence the evaluation process.
>
> – A stakeholder or client with a personal professional interest in promoting the program being evaluated may influence the outcome of an evaluation by providing erroneous surveys or interview responses. (p. 115)

In response to these threats to the integrity of a program evaluation, the applicable Propriety Standard reads: "Conflicts of interest should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results" (The Joint Committee on Standards for Educational Evaluation, 1994, p. 115). Seven specific guidelines are suggested for accomplishing this goal, but many of them appear to put the onus on the individual evaluators and their clients to avoid the problem. For example, the first three guidelines recommend that the evaluator and the client jointly identify in advance possible conflicts of interest, agree in writing to preventive procedures, and seek more balanced outside perspectives on the evaluation. These are all excellent suggestions and should work extremely well in all cases, except where either the evaluator, the client, or both are actually *experiencing* real-world conflicts of interests. Another interesting guideline is: "Make internal evaluators directly responsible to agency heads, thus limiting the influence other agency staff might have on the evaluators" (p. 116). We remain unconvinced that the lower-echelon and often underpaid agency staff have more of a vested interest in the outcome of an evaluation than the typically more highly paid agency head presumably *managing* the program being evaluated.

A similar situation exists with respect to the Propriety Standards for the Disclosure of Findings: "The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation, and any others with expressed legal rights to receive the results" (The

Joint Committee on Standards for Educational Evaluation, 1994, p. 109). This statement implicitly recognizes the problem of *information asymmetry* described above but leaves it up to the "formal parties to an evaluation" to correct the situation. In contrast, we maintain that these are precisely the interested parties that will be most subject to *moral hazards* and *perverse incentives* and are therefore the *least motivated* by the financial, professional, and possibly even political incentives currently in place to act in the broader interests of society as a whole in the untrammelled public dissemination of information.

Besides financial gain or professional advancement, Stufflebeam (2001) has recognized *political* gains and motivations also play a role in the problem of *information asymmetry:*

> The advance organizers for a politically controlled study include implicit or explicit threats faced by the client for a program evaluation and/or objectives for winning political contests. The client's purpose in commissioning such a study is to secure assistance in acquiring, maintaining, or increasing influence, power, and/or money. The questions addressed are those of interest to the client and special groups that share the client's interests and aims. Two main questions are of interest to the client: What is the truth, as best can be determined, surrounding a particular dispute or political situation? What information would be advantageous in a potential conflict situation? . . . Generally, the client wants information that is as technically sound as possible. However, he or she may also want to withhold findings that do not support his or her position. The strength of the approach is that it stresses the need for accurate information. However, because the client might release information selectively to create or sustain an erroneous picture of a program's merit and worth, might distort or misrepresent the findings, might violate a prior agreement to fully release findings, or might violate a "public's right to know" law, this type of study can degenerate into a pseudoevaluation. (p. 10–11)

By way of solutions, Stufflebeam (2001) then offers:

> While it would be unrealistic to recommend that administrators and other evaluation users not obtain and selectively employ information for political gain, evaluators should not lend their names and endorsements to evaluations presented by their clients that misrepresent the full set of relevant findings, that present falsified reports aimed at winning political contests, or that violate applicable laws and/or prior formal agreements on release of findings. (p. 10)

Like most of the guidelines offered by The Joint Committee on Standards for Educational Evaluation (1994) for the Disclosure of Findings, this leaves it to the private *conscience* of the individual administrator or evaluator to not abuse their position of privileged access to the information produced by program evaluation. It also necessarily relies on the individual administrator's or evaluator's self-reflective and self-critical *conscious awareness* of any biases or selective memory for facts that one might bring to the evaluation process, to be intellectually alerted and on guard against them.

To be fair, some of the other suggestions offered in both of these sections of the Propriety Standards are more realistic, but it is left unclear exactly *who* is supposed to be specifically charged with either implementing or enforcing them. If it is again left up to either the evaluator or the client, acting either individually or in concert, it hardly addresses the problems that we have identified. We will take up some of these suggestions later in this chapter and make specific recommendations for systemic institutional reforms as opposed to individual exhortations to virtue.

As should be clear from our description of the nature of the problem, it is impossible under *information asymmetry* to identify specific program evaluations that have been subject to these moral hazards, precisely because they are pervasive and not directly evident (almost by definition) in any individual final product. There is so much evidence for these phenomena from other fields, such as experimental economics, that the problems we are describing should be considered more than unwarranted speculation. This is especially true in light of the fact that some of our best hypothetical examples came directly from the 1994 book cited above on professional evaluation standards, indicating that these problems have been widely recognized for some time. Further, we do not think that we are presenting a particularly pejorative view of program evaluation collectively or of program evaluators individually: we are instead describing how some of the regrettable limitations of human nature, common to all areas of human endeavor, are exacerbated by the way that program evaluations are generally handled at the institutional level. The difficult situation of the honest and well-intentioned program evaluator under the current system of incentives is just a special case of this general human condition, which subjects both individuals and agencies to a variety of moral hazards.

Cui Bono? The Problem of Multiple Stakeholders

In the historic speech, *Pro Roscio Amerino*, given by Marcus Tullius Cicero in 80 bc, he is quoted as having said (Berry, 2000):

> The famous Lucius Cassius, whom the Roman people used to regard as a very honest and wise judge, was in the habit of asking, time and again, "To whose benefit?"

That speech made famous the expression *"cui bono?"* for the next two millennia that followed. In program evaluation, we have a technical definition for the generic answer to that question. *Stakeholders* are defined as the individuals or organizations that are either directly or indirectly affected by the program and its evaluation (Rossi & Freeman, 1993). Although a subtle difference here is that the stakeholders can either gain or lose and do not always stand to benefit, the principle is the same. Much of what has been written about stakeholders in program evaluation is emphatic on the point that the paying client is neither the only, nor necessarily the most important, stakeholder involved. The evaluator is responsible for providing information to a multiplicity of different interest groups. This casts a program evaluator more in the role of a public servant than a private contractor.

For example, The Joint Committee on Standards for Educational Evaluation (1994) addressed the problem of multiple stakeholders under several different and very interesting headings. First, under Utility Standards, they state that *Stakeholder Identification* is necessary so that "[p]ersons involved in or affected by the evaluation should be identified, so that their needs can be addressed" (p. 23). This standard presupposes the rather democratic and egalitarian assumption that the evaluation is being performed to address the needs of all affected and not just those of the paying client.

Second, in the Feasibility Standards, under *Political Viability*, the explain that "[t]he evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation might be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted" (p. 63). This standard instead presupposes that the diverse stakeholder interests have to be explicitly included within the evaluation process because of political expediency, at the very least as a practical matter of being able to effectively carry out the evaluation, given the possible interference by these same special interest groups. The motivation of the client in having to pay to have these interests represented, and of the evaluator in recommending that this be done, might therefore be one of pragmatic or "enlightened" self-interest rather than of purely altruistic and public-spirited goals.

Third, in the Propriety Standards, under *Service Orientation*, they state: "Evaluations should be designed to assist organizations to address and effectively serve the needs of the targeted participants" (p. 83). This standard presupposes that both the client, directly, and the evaluator, indirectly, are engaged in public service for the benefit of these multiple stakeholders. Whether this results from enlightened self-interest on either of their parts, with an eye to the possible undesirable consequences of leaving any stakeholder groups unsatisfied, or to disinterested and philanthropic communitarianism is left unclear.

Fourth, in the Propriety Standards, under Disclosure of Findings, as already quoted above, there is the statement that the full set of evaluation findings should be made accessible to all the persons affected by the evaluation and not just to the client. This standard again presupposes that the evaluation is *intended* and should be *designed* for the ultimate benefit of *all* persons affected. So *all persons affected* are evidently "*cui bono?*"As another ancient aphorism goes, "*vox populi, vox dei*" ("the voice of the people is the voice of god," first attested to have been used by Alcuin of York, who disagreed with the sentiment, in a letter to Charlemagne in 798 ad; Page, 1909, p. 61).

Regardless of the subtle differences in perspective among many of these standards, all of them present us with a very broad view of for whom program evaluators should actually take themselves to working. These standards again reflect very lofty ethical principles. However, we maintain that the proposed mechanisms and guidelines for achieving those goals remain short of adequate to insure success.

What Do Program Evaluators Actually Do? Part I: Training and Competencies

Conceptual Foundations of Professional Training

Recent attempts have been made (King, Stevahn, Ghere, & Minnema, 2001; Stevahn, King, Ghere,

& Minnema, 2005) at formalizing the competencies and subsequent training necessary of program evaluators. These studies have relied on the thoughts and opinions of practicing evaluators in terms of their opinion of the essential competencies of an effective evaluator. In their studies, participants were asked to rate their perceived importance on a variety of skills that an evaluator should presumably have. In this study (King et al., 2001), there was remarkably general agreement among evaluators for competencies that an evaluator should possess. For example, high agreement was observed for characteristics such as the ability to collect, analyze, and interpret data as well as to report the results. In addition, there was almost universal agreement regarding the evaluator's ability to frame the evaluation question as well as understand the evaluation process. These areas of agreement suggest that the essential training that evaluators should have are in the areas of data-collection methods and data-analytic techniques. Surprisingly, however, there was considerable disagreement regarding the ability to do research-oriented activities, drawing a line between conducting evaluation and conducting research. Nonetheless, we believe that training in research-oriented activities is essential to program evaluation because the same techniques such as framing questions, data collection, and data analysis and interpretation are gained through formal training in research methods. This evidently controversial position will be defended further below. Formal training standards are not yet developed for the field of evaluation (Stevahan et al., 2005). However, it does appear that the training necessary to be an effective evaluator includes formal and rigorous training in both research methods and the statistical models that are most appropriate to those methods. Further below, we outline some of the research methodologies and statistical models that are most common within program evaluation.

In addition to purely data-analytic models, however, *logic models* provide program evaluators with an outline, or a roadmap, for achieving the outcome goals of the program and illustrate relationships between resources available, planned activities, and the outcome goals. The selection of outcome variables is important because these are directly relevant to the assessment of the success of the program. An outcome variable refers to the chosen changes that are desired by the program of interest. Outcome variables can be specified at the level of the individual, group, or population and can refer to a change in specific behaviors, practices, or ways of thinking. A generic outline for developing a logic model is presented by the United Way (1996). They define a logic model as including four components. The first component is called *Inputs* and refers to the resources available to program, including financial funds, staff, volunteers, equipment, and any potential restraints, such as licensure. The second component is called *Activities* and refers to any planned services by the program, such as tutoring, counseling, or training. The third component is called *Outputs* and refers to the number of participants reached, activities performed, product or services delivered, and so forth. The fourth component is called *Outcomes* and refers to the benefits produced by those outputs for the participants or community that the program was directed to help. Each component of the logic model can be further divided into initial or intermediate goals, with a long- or short-term timeframe, and can include multiple items within each component.

Table 17.1 displays an example of a logic model. The logic model shown is a tabular representation that we prepared of the *VERB Logic Model* developed for the Youth Media Campaign Longitudinal Survey, 2002–2004 (Center for Disease Control, 2007). This logic model describes the sequence of events envisioned by the program for bringing about behavior change, presenting the expected relations between the campaign inputs, activities, impacts, and outcomes. A PDF of the original figure can be downloaded directly from the CDC website (http://www.cdc.gov/youthcampaign/research/PDF/LogicModel.pdf).

We believe that it is essential for program evaluators to be trained in the development and application of logic models because they can assist immensely in both the design and the analysis phases of the program evaluation. It is also extremely important that the collaborative development of logic models be used as a means of interacting and communicating with the program staff and stakeholders during this process, as an additional way of making sure that their diverse interests and concerns are addressed in the evaluation of the program.

Conceptual Foundations of Methodological and Statistical Training

In response to a previous assertion by Shadish, Cook, and Leviton (1991) that program evaluation was not merely "applied social science," Sechrest and Figueredo (1993) argued that the reason that this was so was:

Table 17.1. Example of a Logic Model: Youth Media Campaign Longitudinal Survey, 2002–2004

Input	Activities	Short-term outcomes	Mid-term outcomes	Long-term outcomes
Consultants Staff Research and evaluation Contractors Community Infrastructure Partnership	Advertising Promotions Web Public relations National and community outreach	Tween and parent awareness of the campaign brand and its messages "Buzz" about the campaign and brand messages	Changes in: Subjective Norms Beliefs Self-efficacy Perceived behavioral control	Tweens engaging in and maintaining physical activity, leading to reducing chronic disease and possibly reducing unhealthy risky behaviors

> Shadish et al. (1991) appeal to the peculiar problems manifest in program evaluation. However, these various problems arise not merely in program evaluation but whenever one tries to apply social science. The problems, then, arise not from the perverse peculiarities of program evaluation but from the manifest failure of much of mainstream social science and the identifiable reasons for that failure. (p. 646–647)

These "identifiable reasons" consisted primarily of various common methodological practices that led to the "chronically inadequate external validity of the results of the dominant experimental research paradigm" (p. 647) that had been inadvisedly adopted by mainstream social science.

According to Sechrest and Figueredo (1993), the limitations of these sterile methodological practices were very quickly recognized by program evaluators, who almost immediately began creating the quasi-experimental methods that were more suitable for real-world research and quickly superseded the older laboratory-based methods, at least within program evaluation:

> Arguably, for quasi-experimentation, the more powerful and sophisticated intellectual engines of causal inference are superior, by now, to those of the experimental tradition. (p. 647)

The proposed distinction between program evaluation and applied social science was therefore more a matter of *practice* than a matter of *principle*. Program evaluation had adopted methodological practices that were appropriate to its content domain, which mainstream social science had not. The strong implication was that the quasi-experimental methodologies developed within program evaluation would very likely be more suitable for applied social science in general than the dominant experimental paradigm.

Similarly, we extend this line of reasoning to argue that program evaluators do not employ a completely unique set of *statistical* methods either. However, because program evaluators *disproportionately* employ a certain subset of *research* methods, which are now in more general use throughout applied psychosocial research, it necessarily follows that they must therefore *disproportionally* employ a certain subset of *statistical* techniques that are appropriate to those particular designs. In the sections below, we therefore concentrate on the statistical techniques that are in most common use in program evaluation, although these data-analytic methods are not unique to program evaluation *per se.*

What Do Program Evaluators Actually Do? Part II: Quantitative Methods

Foundations of Quantitative Methods: Methodological Rigor

Even its many critics acknowledge that the hallmark and main strength of the so-called quantitative approach to program evaluation resides primarily in its methodological rigor, whether it is applied in shoring up the process of measurement or in buttressing the strength of causal inference. In the following sections, we review a sampling of the methods used in quantitative program evaluation to achieve the sought-after methodological rigor, which is the "Holy Grail" of the quantitative enterprise.

Evaluation-Centered Validity

Within program evaluation, and social sciences in general, there are several types of validity that have been identified. Cook and Campbell (1979) formally distinguished between four types of validity more specific to program evaluation: (1) internal validity, (2) external validity, (3) statistical conclusion validity, and (4) construct validity. Internal validity refers to establishing the causal relationship between two variables such as treatment and outcome; external validity refers to supporting the generalization of results beyond a specific study; statistical conclusion validity refers to applying statistical techniques appropriately to a given problem; and construct validity falls within a broader class of validity issues in measurement (e.g. face validity, criterion validity, concurrent validity, etc.) but specifically consists of assessing and understanding program components and outcomes accurately. In the context of a discussion of methods in program evaluation, two forms of validity take primacy: internal and external validity. Each validity type is treated with more detail in the following sections.

INTERNAL VALIDITY

The utility of a given method in program evaluation is generally measured in terms of how internally valid it is believed to be. That is, the effectiveness of a method in its ability to determine the causal relationship between the treatment and outcome is typically considered in the context of threats to internal validity. There are several different types of threat to internal validity, each of which applies to greater and lesser degrees depending on the given method of evaluation. Here we describe a few possible threats to internal validity.

SELECTION BIAS

Selection bias is the greatest threat to internal validity for quasi-experimental designs. Selection bias is generally a problem when comparing experimental and control groups that have not been created by the random assignment of participants. In such quasi-experiments, group membership (e.g., treatment vs. control) may be determined by some unknown or little-known variable that may contribute to systematic differences between the groups and may thus become confounded with the treatment. *History* is another internal validity threat. History refers to any events, not manipulated by the researcher, that occur between the treatment and the posttreatment outcome measurement that might even partially account for that posttreatment outcome. Any events that coincide with the treatment, whether systematically related to the treatment or not, that could produce the treatment effects on the outcome are considered history threats. For example, practice effects in test taking could account for differences pretest and posttreatment if the same type of measure is given at each measurement occasion. *Maturation* is the tendency for changes in an outcome to spontaneously occur over time. For example, consider a program aimed at increasing formal operations in adolescents. Because formal operations tend to increase over time during adolescence, the results of any program designed to promote formal operations during this time period would be confounded with the natural maturational tendency for formal operations to improve with age. Finally, regression to the mean may cause another threat to internal validity. These *regression artifacts* generally occur when participants are selected into treatment groups or programs because they are unusually high or low on certain characteristics. When individuals deviate substantially from the mean, this might in part be attributable to errors of measurement. In such cases, it might be expected that over time, their observed scores will naturally regress back toward the mean, which is more representative of their true scores. In research designs where individuals are selected in this way, programmatic effects are difficult to distinguish from those of regression toward the mean. Several other forms of threats to internal validity are also possible (for examples, *see* Shadish, Cook, & Campbell, 2002; Mark & Cook, 1984; Smith, 2010).

EXTERNAL VALIDITY

External validity refers to the generalizability of findings, or the application of results beyond the given sample in a given setting. The best way to defend against threats of external validity is to conduct randomized experiments on representative samples, where participants are first randomly drawn from the population and then randomly assigned to the treatment and control groups. Because there are no prior characteristics systematically shared by all members of either the control or treatment participants with members of their own corresponding groups, but systematically differing between those groups, it can be extrapolated that the effect of a program is applicable to others beyond the specific sample assessed. This is not to say that the results of a randomized experiment will be applicable to all populations. For example, if a program is specific to adolescence and was only tested on adolescents, then the impact of the treatment may be specific to adolescents. On the contrary, evaluations that involve groups that were nonrandomly assigned face the possibility that the effect of the treatment is specific to the population being sampled and thus becomes ungeneralizable to other populations. For example, if a program is designed to reduce the recidivism rates of violent criminals, but the participants in a particular program are those who committed a specific violent crime, then the estimated impact of that program may be specific to only those individuals who committed that specific crime and not generalizable to other violent offenders.

Randomized Experiments

Randomized experiments are widely believed to offer evaluators the most effective way of assessing the causal influence of a given treatment or program (St. Pierre, 2004). The simplest type of randomized experiment is one in which individuals are randomly assigned to one of at least two groups—typically a treatment and control group. By virtue of random assignment, each group is approximately equivalent in their characteristics and thus threats to internal validity as a result of selection bias are, by definition, ruled out. Thus, the only systematic difference between the groups is implementation of the treatment (or program participation), so that any systematic differences between groups can be safely attributed to receiving or not receiving the treatment. It is the goal of the evaluator to assess this degree of difference to determine the effectiveness of the treatment or program (Heckman & Smith, 1995; Boruch, 1997).

Although randomized experiments might provide the best method for establishing the causal influence of a treatment or program, they are not without their problems. For example, it may simply be undesirable or unfeasible to randomly assign participants to different groups. Randomized experiments may be undesirable if results are needed quickly. In some cases, implementation of the treatment may take several months or even years to complete, precluding timely assessment of the treatment's effectiveness. In addition, it is not feasible to randomly assign participant characteristics. That is, questions involving race or sex, for example, cannot be randomly assigned, and, therefore, use of a randomized experiment to answer questions that center on these characteristics is impossible. Although experimental methods are useful for eliminating these confounds by distributing participant characteristics evenly across groups, when research questions center on these prior participant characteristics, experimental methods are not feasible methods to apply to this kind of problem. In addition, there are ethical considerations that must be taken into account before randomly assigning individuals to groups. For example, it would be unethical to assign participants to a cigarette smoking condition or other condition that may cause harm. Furthermore, it is ethically questionable to withhold effective treatment from some individuals and administer treatment to others, such as in cancer treatment or education programs (*see* Cook, Cook, & Mark, 1977; Shadish et al., 2002). Randomized experiments may also suffer other forms of selection bias insensitive to randomization. For example, selective attrition from treatments may create nonequivalent groups if some individuals are systematically more likely to drop out than others (Smith, 2010). Randomized experiments may also suffer from a number of other drawbacks. For a more technical discussion of the relationship between randomized experiments and causal inference, *see* Cook, Scriven, Coryn, and Evergreen (2010).

Quasi-Experiments

Quasi-experiments are identical to randomized experiments with the exception of one element: randomization. In quasi-experimental designs, participants are not randomly assigned to different groups, and thus the groups are considered non-equivalent. However, during data analysis, a program evaluator may attempt to construct equivalent groups through matching. Matching involves creating control and treatment groups that are similar in their characteristics, such as age, race, and sex. Attempts to create equivalent groups through matching may result in undermatching, where groups may be similar in one characteristic (such as race) but nonequivalent in others (such as socioeconomic status). In such situations, a program evaluator may make use of statistical techniques that control for undermatching (Smith, 2010) or decide to only focus on matching those characteristics that could moderate the effects of the treatment.

Much debate surrounds the validity of using randomized experiments versus quasi-experiments in establishing causality (*see*, for example, Cook et. al. 2010). Our goal in this section is not to evaluate the tenability of asserting causality within quasi-experimental designs (interested readers are referred to Cook & Campbell, 1979) but, rather, to describe some of the more common methods that fall under the rubric of quasi-experiments and how they relate to program evaluation.

ONE-GROUP, POSTTEST-ONLY DESIGN

Also called the one-shot case study (Campbell, 1957), the one-group, posttest-only design provides the evaluator with information only about treatment participants and only after the treatment has been administered. It contains neither a pretest nor a control group, and thus conclusions about program impact are generally ambiguous. This design can be diagrammed:

$$\text{NR X O}_1$$

The NR refers to the nonrandom participation in this group. The X refers to the treatment, which from left to right indicates that it temporally precedes the outcome (O), and the subscript 1 indicates that the outcome was measured at time-point 1. Although simple in its formulation, this design has a number of drawbacks that may make it undesirable. For example, this design is vulnerable to several threats to internal validity, particularly history threats (Kirk, 2009; Shadish et al., 2002). Because there is no other group with which to make comparisons, it is unknown if the treatment is directly associated with the outcome or if other events that coincide with treatment implementation confound treatment effects.

Despite these limitations, there is one circumstance in which this design might be appropriate. As discussed by Kirk (2009), the one-group, posttest-only design may be useful when sufficient knowledge about the expected value of the dependent variable in the absence of the treatment is available. For example, consider high school students who have taken a course of calculus and recently completed an exam. To assess the impact of the calculus course, one would have to determine the average expected grade on the exam had the students not taken the course and compare it to the scores they actually received (Shadish et al., 2002). In this situation, the expected exam grade for students had they not taken the course would likely be very low compared to the student's actual grades. Thus, this technique is only likely useful when the size of the effect (taking the class) is relatively large and distinct from alternative possibilities (such has history threat).

POSTTEST-ONLY, NONEQUIVALENT GROUPS DESIGN

This design is similar to the one-group, posttest-only design in that only posttest measures are available; however, in this design, a comparison group is available. Unlike a randomized experiment with participants randomly assigned to a treatment and a control group, in this design participant group membership is not randomized. This design can be diagrammed:

$$\frac{\text{NR X O}_1}{\text{NR X O}_1}$$

Interpretation of this diagram is similar to that of the previous one; however, in this diagram, the dashed line indicates that the participants in each of these groups are different individuals. It is important to note that the individuals in these two groups represent nonequivalent groups and may be systematically different from each other in some uncontrolled extraneous characteristics. This design is a significant improvement over the one-group, posttest-only design in that a comparison group that has not experienced the treatment can be compared on the dependent variable of interest. The principal drawback, however, is that this method may suffer from selection bias if the control and treatment groups differ from each other in a systematic way this is not related to the treatment (Melvin & Cook, 1984). For example, participants selected into a treatment based on their need for the treatment may differ on characteristics other than treatment need from those not selected into the treatment.

Evaluators may implement this method when pretest information is not available, such as when a treatment starts before the evaluator has been consulted. In addition, an evaluator may choose to use this method if pretest measurements have the potential to influence posttest outcomes (Willson & Putnam, 1982). For example, consider a program designed to increase spelling ability in middle childhood. At pretest and posttest, children are given a list of words to spell. Program effectiveness would then be assessed via estimating the improvement in spelling by comparing their spelling performance before and after the program. However, if the same set of words were given to children at posttest that where administered in the pretest, then the effect of the program might be confounded with a practice effect.

Although it is possible that pretest measures may influence posttest outcomes, such situations are likely to be relatively rare. In addition, the costs of not including a pretest may significantly outweigh the potential benefits (*see* Shadish et al., 2002).

ONE-GROUP, PRETEST–POSTTEST DESIGN

In the pretest–posttest design, participants are assessed before the treatment and assessed again after the treatment has been administered. However, there is no control group comparison. The form of this design is:

$$\text{NR } O_1 X \ O_2.$$

This design provides a baseline with which to compare the same participants before and after treatment. Change in the outcome between pretest and posttest is commonly attributed to the treatment. This attribution, however, may be misinformed as the design is vulnerable to threats to internal validity. For example, history threats may occur if uncontrolled extraneous events coincide with treatment implementation. In addition, maturation threats may also occur if the outcome of interest is related with time. Finally, if the outcome measure was unusually high or low at pretest, then the change detected by the posttest may not be the result of the treatment but, rather, of regression toward the mean (Melvin & Cook, 1984).

Program evaluators might use this method when it is not feasible to administer a program only to one set of individuals and not to another. For example, this method would be useful if a program has been administered to all students in a given school, where there cannot be a comparative control group.

PRETEST AND POSTTEST, NONEQUIVALENT GROUPS DESIGN

The pretest and posttest nonequivalent groups design is probably the most common to program evaluators (Shadish et al., 2002). This design combines the previous two designs by not only including pretest and posttest measures but also a control group at pretest and posttest. This design can be diagrammed:

$$\frac{\text{NR } O_1 X \ O_2}{\text{NR } O_1 O_2.}$$

The advantage of this design is that threats to internal validity can more easily be ruled out (Mark & Cook, 1984). When threats to internal validity are plausible, they can be more directly assessed in this design. Further, in the context of this design, statistical techniques are available to help account for potential biases (Kenny, 1975). Indeed, several authors make recommendations that data should be analyzed in a variety of ways to determine the proper effect size of the treatment and evaluate the potential for selection bias that might be introduced as a result of nonrandom groups (*see* Cook & Campbell, 1979; Reichardt, 1979; Bryk, 1980).

In summary, the pretest and posttest, nonequivalent groups design, although not without its flaws, is a relatively effective technique for assessing treatment impact. An inherent strength of this design is that with the exception of selection bias as a result of nonrandom groups, no single general threat to internal validity can be assigned. Rather, threats to internal validity are likely to be specific to the given problem under evaluation.

INTERRUPTED TIME SERIES DESIGN

The interrupted time series design is essentially an extension of the pretest and posttest, nonequivalent groups design, although it not strictly necessary for one to include a control group. Ideally, this design consists of repeated measures of some outcome prior to treatment, implementation of the treatment, and then repeated measures of the outcome after treatment. The general form of this design can be diagrammed:

$$\frac{\text{NR } O_1 O_2 O_3 O_4 O_5 X \ O_6 O_7 O_8 O_9 O_{10}}{\text{NR } O_1 O_2 O_3 O_4 O_5 O_6 O_7 O_8 O_9 O_{10.}}$$

In this diagram, the first line of Os refers to the treatment group, which can be identified by the X among the Os. The second line of Os refers to the control condition, as indicated by the lack of an X. The dashed line between the two conditions indicates participants are different between the two groups, and the NR indicates that individuals and nonrandomly distributed between the groups.

Interrupted time series design is considered by many to be the most powerful quasi-experimental design to examine the longitudinal effects of treatments (Wagner et al., 2002). Several pieces of information can be gained about the impact of a treatment. The first is a change in the level of the outcome (as indicated by a change in the intercept of the regression line) after the treatment. This simply means that change in mean levels of the outcome as a result of the treatment can be assessed. The second is change in the temporal trajectory of the outcome (as indicated by a change in the slope of the regression line). Because of the longitudinal nature of the data, the temporal trajectories of the outcome can be assessed both pre- and posttreatment, and any change in the trajectories can be estimated. Other effects can be assessed as well, such as any changes in the variances of the outcomes after treatment, whether the effect of the treatment is continuous or discontinuous and if the effect of the treatment is immediate or delayed (*see* Shadish et al., 2002). Thus, several different aspects of treatment implementation can be assessed with this design.

In addition to its utility, the interrupted time series design (with a control group) is robust against many forms of internal validity threat. For example, with a control group added to the model, history is no longer a threat because any external event that might have co-occurred with the treatment should have affected both groups, presumably equally. In addition, systematic pretest differences between the treatment and control groups can be more accurately assessed because there are several pretest measures. Overall, the interrupted time series design with a nonequivalent control group is a very powerful design (Mark & Cook, 1984).

A barrier to this design includes the fact that several measurements are needed both before and after treatment. This may be impossible if the evaluator was not consulted until after the treatment was implemented. In addition, some evaluators may have to rely on the availability of existing data that they did not collect or historical records. These limitations may place constraints on the questions that can be asked by the evaluator.

REGRESSION DISCONTINUITY DESIGN

First introduced to the evaluation community by Thistlethwaite and Campbell (1960), the regression-discontinuity design (RDD) provides a powerful and unbiased method for estimating treatment effects that rivals that of a randomized experiment (*see* Huitema, 1980). The RDD contains both a treatment and a control group. Unlike other quasi-experimental designs, however, the determination of group membership is perfectly known. That is, in the RDD, participants are assigned to either a treatment or control group based on a particular cutoff (*see also* Trochim, 1984, for a discussion of so-called fuzzy regression discontinuity designs). The RDD takes the following form:

$$O_A C\ X\ O_2$$

$$O_A C\ O_{2.}$$

O_A refers to the pretest measure for which the criterion for group assignment is determined, C refers to the cutoff score for group membership, X refers to the treatment, and O_2 refers to the measured outcome. As an example, consider the case where elementary school students are assigned to a program aimed at increasing reading comprehension. Assignment to the program versus no program is determined by a particular cutoff score on a pretest measure of reading comprehension. In this case, group membership (control vs. treatment) is not randomly assigned; however, the principle or decision rule for assignment is perfectly known (e.g., the cut-off score). By directly modeling the known determinant of group membership, the evaluator is able to completely account for the selection process that determined group membership.

The primary threat to the internal validity of the RDD is history, although the tenability of this factor as a threat is often questionable. More importantly, the analyses of RDDs are by nature complex, and correctly identifying the functional forms of the regression parameters (linear, quadratic, etc.) can have a considerable impact on determining the effectiveness of a program (*see* Reichardt, 2009, for a review).

Measurement and Measurement Issues in Program Evaluation

In the context of program evaluation, three types of measures should be considered: (1) input measures, (2) process measures, and (3) outcome measures (Hollister & Hill, 1995). Input measures consist of more general measures about the program and the participants in them, such as the number of individuals in a given program or the ethnic composition of program participants. Process measures center on the delivery of the program, such as a measure of teaching effectiveness in a program designed to improve reading comprehension in schoolchildren. Outcome measures are those measures that focus on the ultimate result of the program, such as a measure of reading comprehension at the conclusion of the program. Regardless of the type of measurement being applied, it is imperative that program evaluators utilize measures that are consistent with the goals of the evaluation. For example, in an evaluation of the performance of health-care systems around the world, the World Health Organization (WHO) published a report (World Health Organization, 2000) that estimated how well the different health-care systems of different countries were functioning. As a part of this process, the authors of the report sought to make recommendations based on empirical evidence rather than WHO ideology. However, their measure of overall health system functioning was based, in part, on an Internet-based questionnaire of 1000 respondents, half of whom were WHO employees. In this case, the measure used to assess health system functioning was inconsistent with the goals of the evaluation, and this problem did not go unnoticed (*see* Williams, 2001). Evaluators should consider carefully what the goals of a given program are and choose measures that are appropriate toward the goals of the program.

An important part of choosing measures appropriate to the goals of a program is choosing measures that are psychometrically sound. At minimum, measures should be chosen that have been demonstrated in past research to have adequate internal consistency. In addition, if the evaluator intends to administer a test multiple times, then the chosen measure should have good test–retest reliability. Similarly, if the evaluator chooses a measure that is scored by human raters, then the measure should show good inter-rater reliability. In addition to these basic characteristics of reliability, measures should also have good validity, in that they actually measure the constructs that they are intended to measure. Published measures are more likely to already possess these qualities and thus may be less problematical when choosing among possible measures.

It may be the case, however, that either an evaluator is unable to locate an appropriate measure or no appropriate measures currently exist. In this case, evaluators may consider developing their own scales of measurement as part of the process of program evaluation. Smith (2010) has provided a nice tutorial on constructing a survey-based scale for program evaluation. Rather than restate these points, however, we discuss some of the issues that an evaluator may face when constructing new measures in the process of program evaluation. Probably the most important point is that there is no way, *a priori*, to know that the measure being constructed is valid, in that it measures what it intended to measure. Presumably the measure will be high in face validity, but this does not necessarily translate into construct validity. Along these lines, if an evaluator intends to create their own measure of a given construct in the context of an evaluation, then the measure should be properly vetted regarding its utility in assessing program components prior to making any very strong conclusions.

One way to validate a new measure is to add additional measures in the program evaluation to show convergent and divergent validity. In addition, wherever possible, it would be ideal if pilot data on the constructed measure could be obtained from some of the program participants to help evaluate the psychometric properties of the measure prior to its administration to the larger sample that will constitute the formal program evaluation.

Another problem that program evaluators may face is that of "re-inventing the wheel," when creating a measure from scratch. When constructing a measure, program evaluators are advised to research the construct that they intend to measure so that useful test items can be developed. One way to avoid re-inventing the wheel may be to either borrow items for other validated scales or to modify an existing scale to suit the needs of the program and evaluation, while properly citing the original sources. Collaboration with academic institutions can help facilitate this process by providing resources to which an evaluator may not already have access.

Statistical Techniques in Program Evaluation

Program evaluators may employ a wide variety of techniques to analyze the results of their evaluation. These techniques range from "simple" correlations, t-tests, and analyses of variance (ANOVAs) to more intensive techniques such as multilevel modeling, structural equation modeling, and latent growth curve modeling. It is often the case that the research method chosen for the evaluation dictates the statistical technique used to analyze the resultant data. For experimental designs and quasi-experimental designs, various forms of ANOVA, multiple regression, and non-parametric statistics may suffice. However, for longitudinal designs, there may be more options for the program evaluator in terms of how to analyze the data. In this section, we discuss some of the analytical techniques that might be employed when analyzing longitudinal data and, more specifically, the kind of longitudinal data derived from an interrupted time series design. For example, we discuss the relative advantages and disadvantages of repeated measures analysis of variance (RM-ANOVA), multilevel modeling, and latent growth curve modeling. For a more systematic review of some of the more basic statistical techniques in program evaluation, readers are referred to Newcomer and Wirtz (2004).

To discuss the properties of each of these techniques, consider a hypothetical longitudinal study on alcohol use among adolescents. Data on alcohol consumption were collected starting when the adolescents were in sixth grade and continued through the twelfth grade. As a part of the larger longitudinal study, a group of adolescents were enrolled in a program aimed at reducing alcohol consumption during adolescence. The task of the evaluator is to determine the effectiveness of the program in reducing alcohol use across adolescence.

One way to analyze such data would be to use RM-ANOVA. In this analysis, the evaluator would have several measures of alcohol consumption across time and another binary variable that coded whether a particular adolescent received the program. When modeling this data, the repeated measures of alcohol consumption would be treated as a repeated measure, whereas the binary program variable would be treated as a fixed factor. The results of this analysis would indicate the functional form of the alcohol consumption trend over time as well as if the trend differed between the two groups (program vs. no program). The advantage of the repeated measures technique is that the full form of the alcohol consumption trajectory can be modeled, and increases and decreases in alcohol consumption can easily be graphically displayed (e.g., in SPSS). In addition, the shape of the trajectory (e.g., linear, quadratic, cubic, etc.) of alcohol consumption can be tested empirically through significance testing. The primary disadvantage of RM-ANOVA in this case is that the test of the difference between the two groups is limited to the shape of the overall trajectory and cannot be extended to specific periods of time. For example, prior to the treatment, we would expect that the two groups should not differ in their alcohol consumption trajectories; only after the treatment do we expect differences. Rather than specifically testing the difference in trajectories following the treatment, a test is being conducted about the overall shape of the curves. In addition, this technique cannot test the assumption that the two groups are equal in their alcohol consumption trajectories prior to the treatment, a necessary precondition needed to make inferences about the effectiveness of the program. To test these assumptions, we need to move to multilevel modeling (MLM).

Multilevel modeling is a statistical technique designed for use with data that violate the assumption of independence (*see* Kenny, Kashy, & Cook, 2006). The assumption of independence states that after controlling for an independent variable, the residual variance between variables should be independent. Longitudinal data (as well as dyadic data) tend to violate this assumption. The major advantage of MLM is that the structure of these residual covariances can be directly specified (*see* Singer, 1998, for examples). In addition, and more specifically in reference to the current program evaluation example, the growth function of longitudinal data can be more directly specified in a number of flexible ways (*see*, for example, Singer & Willett, 2003, p. 138). One interesting technique that has seen little utilization in the evaluation field is what has been called a piecewise growth model (*see* Seltzer, Frank, & Bryk, 1994, for an example). In this model, rather than specifying a single linear or curvilinear slope, two slopes with a single intercept are modeled. The initial slope models change up to a specific point, whereas the subsequent slope models change after a specific point. Perhaps by now, the utility of this method has been discovered as it applies to time series analysis in that trajectories of change can be modeled before and after the implementation of a treatment, intervention, or program. In terms of the present example, change in alcohol consumption can be a model for the entire sample before and after the program implementation. Importantly, different slopes can be estimated for the two different groups (program vs. no program) and empirically tested for differences in the slopes. For example, consider a model that specified a linear growth trajectory for the initial slope (prior to the program) and another linear growth trajectory for the subsequent slope (after the program). In a piecewise growth model, significance testing (as well as the estimation of effect sizes) can be performed separately for both the initial slope and subsequent slope. Further, by adding the fixed effect of program participation (program vs. no program), initial and subsequent slopes for the different groups can be modeled and the differences between the initial and subsequent slopes for the two groups can be tested. With piecewise growth modeling, the evaluator can test the assumption that the initial slopes between the two groups are, in fact, the same as well as test the hypotheses that following the program the growth trajectories of the two groups differ systematically, with the intended effect being that the program group shows a less positive or even negative slope over time (increased alcohol consumption among adolescents being presumed undesirable).

Although this method is very useful for interrupted time series design, it is not without its drawbacks. Perhaps one drawback is the complexity of model building; however, this drawback is quickly ameliorated with some research on the topic and perhaps some collaboration. Another drawback to this technique is that the change in subsequent slope may be driven primarily by a large change in behavior immediately following the program and does not necessarily indicate a lasting change over time. Other modeling techniques can be used to explore such variations in behavioral change over time. The interested reader can refer to Singer and Willett (2003).

Structural equation modeling can also be used to model longitudinal data through the use of latent growth curve models. For technical details on how to specify a latent growth curve model, the interested reader can refer to Duncan, Duncan, and Stryker (2006). The primary advantage of using latent growth curve modeling over MLM is that latent variables can be used (indeed, piecewise growth models can be estimated in a latent growth model framework as well; *see* Muthén & Muthén, 2009, p. 105). In addition, more complex models such as multilevel latent growth curve models can be implemented. Such models also account for the interdependence of longitudinal data but are also useful when data are nested—for example, when there is longitudinal data on alcohol consumption in several different schools. These models can become increasingly complex, and it is recommended that evaluators without prior knowledge of this statistical technique seek the advice and possible collaboration with experts on this topic.

What Do Program Evaluators Actually Do? Part III: Qualitative Methods

Foundations of Qualitative Methods: Credibility and Quality

The two principal pillars on which qualitative program evaluation rests are *credibility* and *quality*. These two concepts lie at the heart of all qualitative research, regardless of any more specific philosophical or ideological subscriptions (Patton, 1999). Although these concepts are not considered to be purely independent of each other in the literature, for the sake of clarity of explanation, we will treat them as such unless otherwise specified.

CREDIBILITY

When performing a literature search on the credibility concept within the qualitative paradigms, the emphasis seems to be primarily with the researcher and only secondarily on the research itself. The points most notably brought to light are those of researcher *competence* and *trustworthiness*.

COMPETENCE

Competence is the key to establishing the credibility of a researcher. If a researcher is deemed as incompetent, then the credibility and quality of the entire study immediately comes into question. One of the biggest issues lies with training of qualitative researchers in methods. In a classic example of the unreliability of eyewitness testimonies, Katzer, Cook, and Crouch (1978) point out what can happen when sufficient training does not occur. Ignorance is not bliss, at least in science. Giving any researcher tools without the knowledge to use them is simply bad policy. Subsequent to their initial training, the next most important consideration with respect to competence is the question of their scientific "track record." If an evaluator has demonstrated being able to perform high-quality research many times, then it can be assumed that the researcher is competent.

TRUSTWORTHINESS

Something else to note when considering the credibility of an evaluator is trustworthiness. There is little doubt that the researcher's history must be taken into account (Patton, 1999). Without knowing where the researcher is "coming from," in terms of possible ideological commitments, the reports made by a given evaluator may appear objective but might actually be skewed by personal biases. This is especially a problem with more phenomenological methods of qualitative program evaluation, such as interpretive and social constructionist. As Denzin (1989) and many others have pointed out, pure neutrality or impartiality is rare. This means that not being completely forthright about any personal biases should be a "red flag" regarding the trustworthiness (or lack thereof) of the evaluator.

JUDGING CREDIBILITY

There are those that argue that credibility and trustworthiness are not traits that an evaluator can achieve themselves, but rather that it has to be established by the stakeholders, presumably democratically and all providing equal input (Atkinson, Heath, & Chenail, 1991). This notion seems to be akin to that of external validity. This is also fundamentally different from another school of thought that claims to be able to increase "truth value" via external auditing (Lincoln & Guba, 1985). Like external validation, Atkinson would argue that evaluators are not in a position to be able to judge their own work and that separate entities should be responsible for such judging. According to this perspective, stakeholders need to evaluate the evaluators. If we continue down that road, then the evaluators of the evaluators might need to be evaluated, and they will need to be evaluated, and so on and so forth. As the *Sixth Satire*, written by First Century Roman poet Decimus Iunius Iuvenalis, asks: *"quis custodiet ipsos custodes?"* ("who shall watch the watchers?"; Ramsay, 1918) The way around this infinite regress is to develop some sort of standard by which comparisons between the researcher and the standard can be made.

Evaluators can only be as credible as the credibility of the system that brought them to their current positions. Recall that there is a diverse array of backgrounds among program evaluators and a broad armamentarium of research methods and statistical models available from which they can select, as well as the fact that there are currently no formal training standards in program evaluation (Stevahan et al., 2005). Until a standard of training is in place, there is no *objective* way to assess the credibility of a researcher, and evaluators are forced to rely on highly subjective measures of credibility, fraught with biases and emotional reactions.

Quality

The other key concern in qualitative program evaluation is quality. Quality concerns echo those voiced regarding questions of reliability and validity in quantitative research, although the framing of these concepts is done within the philosophical framework of the research paradigm (Golafshani, 2003). Patton, as the "go-to guy" for how to do qualitative program evaluations, has applied quantitative principles to qualitative program evaluation throughout his works (Patton, 1999, 1997, 1990), although they seem to fall short in application. His primary emphases are on rigor in testing and interpretation.

RIGOROUS TESTING

Apart from being thorough in the use of any single *qualitative method*, there appears to be a single key issue with respect to testing rigor, and this is called *triangulation*.

Campbell discussed the concept of methodological triangulation (Campbell, 1953, 1956; Campbell & Fiske, 1959). Triangulation is the use of multiple methods, each having their own unique biases, to measure a particular phenomenon. This multiple convergence allows for the systematic variance ascribable to the "trait" being measured by multiple indicators to be partitioned from the systematic variance associated with each "method" and from the unsystematic variance attributable to the inevitable and random "error" of measurement, regardless of the method used. Within the context of qualitative program evaluation, this can consist either of mixing quantitative and qualitative methods or of mixing qualitative methods. Patton (1999) outspokenly supported the use of either form of triangulation, because each method of measurement has its own advantages and disadvantages.

Other contributors to this the literature have claimed that the "jury is still out" concerning the advantages of triangulation (Barbour, 1998) and that clearer definitions are needed to determine triangulation's applicability to qualitative methods. Barbour's claim seems unsupported because there is a clear misinterpretation of Patton's work. Patton advocates a convergence of evidence. Because the nature of qualitative data is not as precise as the nature of quantitative data, traditional hypothesis testing is virtually impossible. Barbour is under the impression that Patton is referring to *perfectly* congruent results. This is obviously not possible because, as stated above, there will always be different divergences between different measures based on which method of measurement is used. Patton is advocating the use of multiple and mixed methods to produce consistent results. One example of how to execute triangulation within the qualitative paradigm focused on three different educational techniques (Oliver-Hoyo & Allen, 2006). For cooperative grouping, hands-on activities, and graphical skills, these authors used interviews, reflective journal entries, surveys, and field notes. The authors found that the exclusive use of surveys would have led to different conclusions, because the results of the surveys alone indicated that there was either no change or a negative change, whereas the other methods instead indicated that there was a positive change with the use of these educational techniques. This demonstrates the importance of using triangulation. When results diverge, meaning that they show opposing trends using different methods, the accuracy of the findings falls into question.

Lincoln and Guba (1985) have also discussed the importance of triangulation but have emphasized its importance in increasing the rigor and trustworthiness of research with respect to the interpretation stage. This is ultimately because all methods will restrict what inferences can be made from a qualitative study.

RIGOROUS INTERPRETATION

As with quantitative program evaluation, qualitative methods require rigorous interpretation at two levels: the microscale, which is the sample, and the macroscale, which is the population for quantitative researchers and is most often the social or global implication for qualitative researchers.

Looking at qualitative data is reminiscent of exploratory methods in quantitative research but without the significance tests. Grounded Theory is one such analytic method. The job of the researcher is to systematically consider all of the data and to extract theory from the data (Strauss & Corbin, 1990). The only exception made is for theory extension when going with a preconceived theory is acceptable.

Repeatedly throughout the literature (e.g., Patton, 1999; Atkinson, Heath, & Chenail, 1991; Lincoln & Guba, 1985), the evaluator is emphasized as the key instrument in analysis of data. Although statistics can be helpful, they are seen as restricting and override any "insight" from the researcher. Analysis necessarily depends on the "astute pattern recognition" abilities of the investigating researcher (Patton, 1999). What Leech and Onwuegbuzie (2007) have called "data analysis triangulation" is essentially an extension of the triangulation concept described by Patton (1999) as applied to data analytics. The idea is that by analyzing data with different techniques, convergence can be determined, making the findings more credible or trustworthy.

Because a large part of qualitative inquiry is subjective and dependent on a researcher's creativity, Patton (1999) has advocated reporting all relevant data and making explicit all thought processes, thus avoiding the problem of interpretive bias. This may allow anyone that reads the evaluation report to determine whether the results and suggestions were sufficiently grounded. Shek et al. (2005) have outlined the necessary steps that must occur to demonstrate that the researcher is not simply forcing their opinions into their research.

Qualitative Methods in Program Evaluation

The most common methods in qualitative program evaluation are straightforward and fall into one of two broad categories: first-party or third-party methods (done from the perspective of the evaluands, which are the programs being evaluated). These methods are also used by more quantitative fields of inquiry, although they are not usually framed as part of the research process.

FIRST-PARTY METHODS

When an evaluator directly asks questions to the entities being evaluated, the evaluator is utilizing a first-party method. Included in this method are techniques such as interviews (whether of individuals or focus groups), surveys, open-ended questionnaires, and document analyses.

Interviews, surveys, and open-ended questionnaires are similar in nature. In interviews, the researcher begins with a set of potential questions, and depending on the way in which the individuals within the entity respond, the questions will move in a particular direction. The key here is that the questioning is fluid, open, and not a forced choice. In the case of surveys and open-ended questionnaires, fixed questions are presented to the individual, but the potential answers are left as open as possible, such as in short-answer responding. Like with interviews, if it can be helped, the questioning is open and not a forced choice (*see* Leech & Onwuegbuzie, 2007; Oliver-Hoyo & Allen, 2006; Pugach, 2001; Patton, 1999).

Although document analysis is given its own category in the literature (Pugach, 2001; Patton, 1999), it seems more appropriate to include the document analysis technique along with other first-party methods. Document analysis will usually be conducted on prior interviews, transcribed statements, or other official reports. It involves doing "archival digging" to gather data for the evaluation. Pulling out key "success" or "failure" stories are pivotal to performing these kinds of analyses and utilized as often as possible for illustrative purposes.

The unifying theme of these three techniques is that the information comes from within the entity being evaluated.

THIRD-PARTY METHODS

The other primary type of methodology used in qualitative research is third-party methods. The two major third-party methods are naturalistic observations and case studies. These methods are more phenomenological in nature and require rigorous training on the part of the researcher for proper execution. These methods are intimately tied with the Competence section above.

Naturalistic observation has been used by biological and behavioral scientists for many years and involves observation of behavior within its natural context. This method involves observing some target (whether that is a human or nonhuman animal) performing a behavior in its natural setting. This is most often accomplished reviewing video recordings or recording the target in person while not interacting with the target. There are, however, many cases of researchers interacting with the target and then "going native" or becoming a member of the group they initially sought to study (Patton, 1999). Some of the most prominent natural scientists have utilized this method (e.g., Charles Darwin, Jane Goodall, and Isaac Newton). According to Patton (1999), there are well-documented problems with this method, including phenomena like researcher presence effects, "going native," researcher biases, and concerns regarding researcher training. Despite the inherent risks and problems with naturalistic observation, it has been, and will likely continue to be, a staple method within scientific inquiry.

Case studies can be special cases of a naturalistic observation or can be a special kind of "artificial" observation. Case studies provide extensive detail about a few individuals (Banfield & Cayago-Gicain, 2006; Patton, 1999) and can simply be used to demonstrate a point (as in Abma, 2000). Case studies usually take a substantial amount of time to gather appropriate amounts of idiographic data. This method utilizes any records the researcher can get their hands on, regarding the individual being studied (self-report questionnaires, interviews, medical records, performance reviews, financial records, etc.). As with naturalistic observation, case study researchers must undergo much training before they can be deemed "capable" of drawing conclusions based on a single individual. The problems with case studies are all of those in naturalistic observation but with the addition of a greater probability of a sampling error. Because case studies are so intensive, they are often also very expensive. The salience and exhaustion of a few cases makes it difficult to notice larger, nominal trends in the data (Banfield & Cayago-Gicain, 2006). This could also put a disproportionate emphasis on the "tails" of the distribution, although that may be precisely what the researcher wants to accomplish (*see* next section).

Critiques/Criticisms of Quantitative Methods

One of the major critiques of *quantitative methods* by those in qualitative evaluation is that of credibility. Relevance of findings using quantitative evaluation to what is "important" or what is "the essence" of the question, according to those using qualitative evaluation methods, is rather poor (*see* discussion in Reichardt & Rallis, 1994a, 1994b). Recall that according to Atkinson (1991), the relevance of findings, and whether they are appropriate, cannot be determined by the evaluator. The stakeholders are the only ones that can determine relevance. Although there are those in qualitative program evaluation that think almost everything is caused by factors like "social class" and "disparity in power," Atkinson would argue that the evaluator is not able to determine what is or is not relevant to the reality experienced by the stakeholders.

Another criticism is that quantitative research tends to focus simply on the majority, neglecting the individuals in the outer ends of the normal distribution. This is a valid critique for those quantitative researchers who tend to "drop" their outliers for better model fits. Banfield and Cayago-Gicain (2006) have pointed out that qualitative research allows for more detail on a smaller sample. This allows for more context surrounding individuals to be presented. With additional knowledge from the "atypical" (tails of the distribution) cases, theory can be extracted that fits all of the data best and not just the "typical" person.

Beyond the Qualitative/Quantitative Debate

Debate about the superiority of qualitative versus quantitative methodology has a long history in program evaluation. Prior to the 1970s, randomized experiments were considered the gold standard in impact assessment. More and more, however, as evaluators realized the limitations of randomized experiments, quasi-experiments became more acceptable (Madey, 1982). It was also not until the early 1970s that qualitative methods became more acceptable; however, epistemological differences between the two camps prevailed in perpetuating the debate, even leading to distrust and slander between followers of the different perspectives (Kidder & Fine, 1987). In an effort to ebb the tide of the qualitative–quantitative debate, some evaluators have long called for integration between the two approaches. By recognizing that methods typically associated with qualitative and quantitative paradigms are not inextricably linked to these paradigms (Reichardt & Cook, 1979), an evaluator has greater flexibility with which to choose specific methods that are simply the most appropriate for a given evaluation question (Howe, 1988). Further, others have pointed out that because the qualitative and quantitative approaches are not entirely incompatible (e.g., Reichardt & Rallis, 1994a, 1994b), common ground can be found between the two methods when addressing evaluation questions.

An evaluator thus may choose to use quantitative or qualitative methods alone or may choose to use both methods in what is known as a mixed methods design. A mixed methods approach to evaluation has been advocated on the basis that the two methods: (1) provide cross-validation (triangulation) of results and (2) complement each other, where the relative weakness of one method becomes the relative strength of the other. For example, despite the purported epistemological differences between the two paradigms, the different approaches to evaluation often lead to the same answers (Sale, Lohfeld, & Brazil, 2002). Thus, combining both methods into the same evaluation can result in converging lines of evidence. Further, each method can be used to complement the other. For example, the use of qualitative data collection techniques can help in the development or choice of measurement instruments, as the personal interaction with individual participants may pave the way for collecting more sensitive data (Madey, 1982).

Despite the promise of integrating qualitative and quantitative methods through a mixed method approach, Sale et al. (2002) challenged the notation that qualitative and quantitative methods are separable from their respective paradigms, contrary to the position advocated by Reichardt and Cook (1979). Indeed, these authors have suggested that because the two approaches deal with fundamentally different perspectives, the use of both methods to triangulate or complement each other is invalid. Rather, mixed methods should be used in accordance with one another only with the recognition of the different questions that they address. In this view, it should be recognized that qualitative and quantitative methods do address different questions, but at the same time they can show considerable overlap. Thus, mixed methods designs provide a more complete picture of the evaluation space by providing all three components: cross-validation, complimentarity, and unique contributions from each.

Despite the utility in principle of integrating both qualitative and quantitative methods in evaluation and the more recent developments in mixed methodology (*see* Greene & Caracelli, 1997), the overwhelming majority of published articles in practice employ either qualitative or quantitative methods to the exclusion of the other. Perhaps one reason for the persistence of the single methodology approach is the lack of training in both approaches in evaluation training programs. For example, the AEA website (http://www.eval.org) lists 51 academic programs that have an evaluation focus or evaluation option. In a review of each of these programs, we found that none of the evaluation programs had a mixed methods focus. Moreover, when programs did have a focus, it was on quantitative methods. Further, within these programs quantitative methods and qualitative methods were generally taught in separate classes, and there was no evidence of any class in any program that was focused specifically on mixed methods designs. Indeed, Johnson and Onwuegbuzie (2004) have noted that " . . . graduate students who graduate from educational institutions with an aspiration to gain employment in the world of academia or research are left with the impression that they have to pledge allegiance to one research school of thought or the other" (p. 14). Given the seeming utility of a mixed methods approach, it is unfortunate that more programs do not offer specific training in these techniques.

Competing Paradigms or Possible Integration?

In summary, the quantitative and qualitative approaches to program evaluation have been widely represented as incommensurable Kuhnian paradigms (e.g., Guba & Lincoln, 1989). On the other hand, it has been suggested that perhaps the road to reconciliation lies with Reichenbach's (1938) important distinction between the *context of discovery* versus the *context of justification* in scientific research. Sechrest and Figueredo (1993) paraphrased their respective definitions:

> In the context of discovery, free reign is given to speculative mental construction, creative thought, and subjective interpretation. In the context of justification, unfettered speculation is superseded by severe testing of formerly favored hypotheses, observance of a strict code of scientific objectivity, and the merciless exposure of one's theories to the gravest possible risk of falsification. (p. 654)

Based on that philosophical perspective, Sechrest and Figueredo (1993) recommended the following methodological resolution of the quantitative/qualitative debate:

> We believe that some proponents of qualitative methods have incorrectly framed the issue as an absolute either/or dichotomy. Many of the limitations that they attribute to quantitative methods have been discoursed upon extensively in the past. The distinction made previously, however, was not between quantitative and qualitative, but between exploratory and confirmatory research. This distinction is perhaps more useful because it represents the divergent properties of two complementary and sequential stages of the scientific process, rather than two alternative procedures . . . Perhaps a compromise is possible in light of the realization that although rigorous theory testing is admittedly sterile and nonproductive without adequate theory development, creative theory

> construction is ultimately pointless without scientific verification. (p. 654)

We also endorse that view. However, in case Sechrest and Figueredo (1993) were not completely clear the first time, we will restate this position here a little more emphatically. We believe that qualitative methods are most useful in exploratory research, meaning early in the evaluation process, the so-called context of discovery, in that they are more flexible and open and permit the researcher to follow intuitive leads and discover previously unknown and unimagined facts that were quite simply not predicted by existing theory. Qualitative methods are therefore a useful tool for *theory construction*. However, the potentially controversial part of this otherwise conciliatory position is that it is our considered opinion that qualitative methods are inadequate for confirmatory research, the so-called context of justification, in that they do not and cannot even in principle be designed to rigorously subject our theories to critical risk of falsification, as by comparison to alternative theories (Chamberlin, 1897; Platt, 1964; Popper, 1959; Lakatos, 1970, 1978). For that purpose, quantitative methods necessarily excel because of their greater methodological rigor and because they are equipped to do just that. Quantitative methods are therefore a more useful tool for *theory testing*. This does not make quantitative evaluation in any way superior to qualitative evaluation, in that exploration and confirmation are both part of the necessary cycle of scientific research.

It is virtually *routine* in many other fields, such as in the science of ethology, to make detailed observations regarding the natural history of any species before generating testable hypotheses that predict their probable behavior. In cross-cultural research, it is standard practice to do the basic ethnographical exploration of any new society under study prior to making any comparative behavioral predictions. These might be better models for program evaluation to follow than constructing the situation as an adversarial one between supposedly incommensurable paradigms.

Conclusions and Recommendations for the Future

As a possible solution to some of the structural problems, moral hazards, and perverse incentives in the practice of program evaluation that we have reviewed, Scriven (1976, 1991) long ago suggested that the program funders should pay for summative evaluations and pay the summative evaluators *directly.* We completely agree with this because we believe that the summative program evaluators must *not* have to answer to the evaluands and that the results of the evaluation should not be "filtered" through them.

For example, in the Propriety Standards for Conflicts of Interest, The Joint Committee on Standards for Educational Evaluation (1994) has issued the following guideline: "Wherever possible, obtain the evaluation contract from the funding agency directly, rather than through the funded program or project" (p. 116). Our only problem with this guideline is that the individual evaluator is called on to implement this solution. Should an ethical evaluator then decline contracts offered by the funded program or project? This is not a realistic solution to the problem. As a self-governing society, we should simply *not accept* summative evaluations in which the funded programs or projects (evaluands) have contracted their own program evaluators. This is a simple matter of protecting the public interest by making the necessary institutional adjustments to address a widely recognized moral hazard.

Similarly, in the Propriety Standards for Disclosure of Findings, The Joint Committee on Standards for Educational Evaluation (1994) has issued various guidelines for evaluators to negotiate in advance with clients for complete, unbiased, and detailed disclosure of all evaluation findings to all directly and indirectly affected parties. The problem is that there is currently no incentive in place for an individual evaluator to do so and possibly jeopardize the award of an evaluation contract by demanding conditions of such unrestricted dissemination of information to which almost no client on this planet is very likely to agree.

On the other hand, we recommend that the evaluands *should* pay for formative evaluations and pay the formative evaluators *directly.* This is because we believe that formative evaluators should provide continuous feedback to the evaluands and not publish those results externally before the program is fully mature (e.g., Tharp & Gallimore, 1979). That way, the formative evaluator can gain the complete trust and cooperation of the program administrators and the program staff. Stufflebeam (2001) writes:

> Clients sometimes can legitimately commission covert studies and keep the findings private, while meeting relevant laws and adhering to an appropriate advance agreement with the evaluator. This can be the case in the United States for private organizations not governed by public disclosure laws. Furthermore,

> an evaluator, under legal contractual agreements, can plan, conduct, and report an evaluation for private purposes, while not disclosing the findings to any outside party. The key to keeping client-controlled studies in legitimate territory is to reach appropriate, legally defensible, advance, written agreements and to adhere to the contractual provisions concerning release of the study's findings. Such studies also have to conform to applicable laws on release of information. (p. 15)

In summary, *summative* evaluations should generally be *external*, whereas *formative* evaluations should generally be *internal*. Only strict adherence to these guidelines will provide the correct incentive system for all the parties concerned, including the general public, which winds up paying for all this. The problem essentially boils down to one of intellectual property. Who actually owns the data generated by a program evaluation? In a free market society, the crude but simple answer to this question is typically "whoever is paying for it!" In almost no case is it the program evaluator, who is typically beholden to one party or another for employment. We should therefore arrange for the owner of that intellectual property to be in every case the party whose interests are best aligned with those of the society as a whole. In the case of a formative evaluation, that party is the program-*providing* agency (the evaluand) seeking to improve its services with a minimum of outside interference, whereas in the case of a summative evaluation, that party is the program-*funding* agency charged with deciding whether any particular program is worth society's continuing investment and support.

Many informative and insightful comparisons and contrasts have been made on the relative merits and limitations of internal and external evaluators (e.g., Braskamp, Brandenburg, & Ory, 1987; Love, 1991; Mathison, 1994; Meyers, 1981; Newman & Brown, 1996; Owen & Rogers, 1999; Patton, 1997; Tang, Cowling, Koumijian, Roeseler, Lloyd, & Rogers, 2002; Weiss, 1998). Although all of those considerations are too many to list here, internal evaluators are generally valued for their greater availability and lower cost as well as for their greater contextual knowledge of the particular organization and ability to obtain a greater degree of commitment from stakeholders to the ultimate recommendations of the evaluation, based on the perceived legitimacy obtained through their direct experience in the program. We believe that these various strengths of internal evaluators are ideally suited to the needs of *formative* evaluation; however, some of these same characteristics might compromise their credibility in the context of a *summative* evaluation. In contrast, external evaluators are generally valued for their greater technical expertise as well as for their greater independence and objectivity, including greater accountability to the public interest and ability to criticize the organization being evaluated—hence their greater ability to potentially position themselves as mediators or arbiters between the stakeholders. We believe that these various strengths of external evaluators are ideally suited to the needs of *summative* evaluation; however, some of these same characteristics might compromise their effectiveness in the context of a *formative* evaluation.

A related point is that *qualitative* methods are arguably superior for conducting the kind of *exploratory* research often needed in a *formative* evaluation, whereas *quantitative* methods are arguably superior for conducting the *confirmatory* research often needed in a *summative* evaluation. By transitive inference with our immediately prior recommendation, we would envision *qualitative* methods being of greater use to *internal* evaluators and *quantitative* methods being of greater use to *external* evaluators, if each method is being applied to what they excel at achieving, within their contingently optimal contexts. With these conclusions, we make our final recommendation that the qualitative/quantitative debate be officially *ended*, with the recognition that both kinds of research each have their proper and necessary place in the cycle of scientific research and, by logical implication, that of program evaluation. Each side must abandon the claims that their preferred methods can do it all and, in the spirit of the great evaluation methodologist and socio-cultural evolutionary theorist Donald Thomas Campbell, to recognize that all our methods are *fallible* (Campbell & Fiske, 1959) and that only through exploiting their mutual *complementarities* can we put all of the interlocking fish scales of omniscience back together (Campbell, 1969).

References

Abma, T. A. (2000). Stakeholder conflict: A case study. *Evaluation and Program Planning, 23,* 199–210.

Atkinson, B., Heath, A., & Chenail, R. (1991). Qualitative research and the legitimization of knowledge. *Journal of Marital and Family Therapy, 17*(2), 175–180.

Barbour, R. S. (1998). Mixing qualitative methods: Quality assurance or qualitative quagmire? *Qualitative Health Research, 8*(3), 352–361.

Banfield, G., & Cayago-Gicain, M. S. (2006). Qualitative approaches to educational evaluation: A regional conference-workshop. *International Education Journal, 7*(4), 510–513.

Berry, D. H. (2000) *Cicero Defense Speeches*, trans. New York: Oxford University Press.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Braskamp, L.A., Brandenburg, D.C. & Ory, J.C. (1987). Lessons about clients' expectations. In J Nowakowski (Ed.), *The client perspective on evaluation: New Directions For Program Evaluation*, 36, 63–74. San Francisco, CA: Jossey-Bass.

Bryk, A. S. (1980). Analyzing data from premeasure/postmeasure designs. In S. Anderson, A. Auquier, W. Vandaele, & H. I. Weisburg (Eds.), *Statistical methods for comparative studies* (pp. 235–260). Hoboken, NJ: John Wiley & Sons.

Campbell, D. T. (1953). *A study of leadership among submarine officers.* Columbus, OH: The Ohio State University, Personnel Research Board.

Campbell, D. T. (1956). *Leadership and its effects upon the group.* Columbus, OH: Bureau of Business Research, The Ohio State University.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Campbell, D. T. (1969). Ethnocentrism of disciplines and the fish-scale model of omniscience. In M. Sherif and C.W. Sherif, (Eds.), *Interdisciplinary Relationships in the Social Sciences,* (pp. 328–348). Chicago IL: Aldine.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56*(2), 81–105.

Center for Disease Control (2007). *Youth media campaign, VERB logic model*. Retrieved May 18, 2012, from http://www.cdc.gov/youthcampaign/research/logic.htm

Chamberlin, T.C. (1897). The method of multiple working hypotheses. *Journal of Geology, 5*, 837–848.

Clayton, R. R., Cattarello, A. M., & Johnstone B. M. (1996). The effectiveness of Drug Abuse Resistance Education (Project DARE): 5-year follow-up results. *Preventive Medicine 25*(3), 307–318.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* Chicago, IL: Rand-McNally.

Cook, T. D., Cook, F. L., & Mark, M. M. (1977). Randomized and quasi-experimental designs in evaluation research: An introduction. In L. Rutman (Ed.), *Evaluation research methods: A basic guide* (pp. 101–140). Beverly Hills, CA: Sage.

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*, 105–117.

Dembe, A. E., & Boden, L. I. (2000). Moral hazard: A question of morality? *New Solutions 2000, 10*(3), 257–279.

Denzin, N. K. (1989). *Interpretive interactionism*. Newbury Park, CA: Sage.

Dukes, R. L., Stein, J. A., & Ullman, J. B. (1996). Long-term impact of Drug Abuse Resistance Education (D.A.R.E.). *Evaluation Review, 21*(4), 483–500.

Dukes, R. L., Ullman, J. B., & Stein, J. A. (1996). Three-year follow-up of Drug Abuse Resistance Education (D.A.R.E.). *Evaluation Review, 20*(1), 49–66.

Duncan, T. E., Duncan, S. C., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd Ed.). Mahwah, NJ: Laurence Erlbaum.

Ennett, S. T., Tobler, M. S., Ringwalt, C. T., & Flewelling, R. L. (1994). How effective is Drug Abuse Resistance Education? A meta-analysis of Project DARE outcomes evaluations. *American Journal of Public Health, 84*(9), 1394–1401.

General Accountability Office (2003). Youth Illicit Drug Use Prevention (Report No. GAO-03-172R). Marjorie KE: Author.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, 8(4), 597–606.

Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms* (pp. 5–17).(New Directions for Evaluation, No. 74). San Francisco: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: Sage.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives, 9*, 85–110.

Hollister, R. G. & Hill, J (1995). Problems in the evaluation of community-wide initiatives. In Connell, J. P., Kubish, A. C., Schorr, L. B., & Weiss, C. H. (Eds.), *New approaches to evaluating community intiatives: Concepts, methods, and contexts* (pp. 127–172). Washington, DC: Aspen Institute.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*, 10–16.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York, NY: John Wiley & Sons.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed-methods research: A research paradigm whose time has come. *Educational Researcher, 33*, 14–26.

Katzer, J., Cook, K. and Crouch, W. (1978). *Evaluating information: A guide for users of social science research*. Reading, MA: Addison-Wesley.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. *Psychological Bulletin, 82*, 345–362.

Kenny, D. A., Kashy, D. A. & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford Press.

Kidder, L. H., & Fine, M. (1987). Qualitative and quantitative methods: When stories converge. In M. M. Mark & R. L. Shotland (Eds.), *Multiple methods in program evaluation* (pp. 57–75). Indianapolis, IN: Jossey-Bass.

King, J., A., Stevahn, L., Ghere, G. & Minnema, J. (2001). Toward a taxonomy of essential evaluator competencies. *American Journal of Evaluation, 22*, 229–247.

Kirk, R. E. (2009). Experimental Design. In R.E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 23–45). Thousand Oaks, CA: Sage.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I., & Musgrave, A., (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Lakatos, I. (1978). *The methodology of scientific research programs.* Cambridge, UK: Cambridge University Press.

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Psychology Quarterly, 22*(4), 557–584.

Loftus, E.F. (1979). Eyewitness Testimony, Cambridge, MA: Harvard University Press.

Love, A.J. (1991). *Internal evaluation: Building organizations from within*. Newbury Park, CA: Sage.

Madey, D. L. (1982). Some benefits of integration qualitative and quantitative methods in program evaluation. *Educational Evaluation and Policy Analysis, 4*, 223–236.

Mark, M. M., & Cook, T. D. (1984). Design of randomized experiments and quasi-experiments. In L. Rutman (Ed.), *Evaluation research methods: A basic guide* (pp. 65–120). Beverly Hills, CA: Sage.

Mathison, S. (1994). Rethinking the evaluator role: partnerships between organizations and evaluators. *Evaluation and Program Planning*, 17(3), 299–304.

Meyers, W. R. (1981). *The Evaluation Enterprise: A Realistic Appraisal of Evaluation Careers, Methods, and Applications*. San Francisco, CA: Jossey-Bass.

Muthén, L. K. & Muthén, B. O. (1998–2009). *Mplus user's guide. Statistical analysis with latent variables.* Los Angeles, CA: Muthén & Muthén.

Newcomer, K. E. & Wirtz, P. W. (2004). Using statistics in evaluation. In Wholey, J. S., Hatry, H. P. & Newcomer, R. E. (Eds.), *Handbook of practical program evaluation* (pp. 439–478). San Francisco, CA: John Wiley & Sons.

Newman, D. L. & Brown, R. D. (1996). *Applied ethics for program evaluation*, San Francisco, CA: Sage.

Office of Management and Budget. (2009). *A new era of responsibility: Renewing America's promise*. Retrieved May 18, 2012, from http://www.gpoaccess.gov/usbudget/fy10/pdf/fy10-newera.pdf

Oliver-Hoyo, M., & Allen, D. (2006). The use of triangulation methods in qualitative educational research. *Journal of College Science Teaching*, 35, 42–47.

Owen, J. M., & Rogers, P. J. (1999). *Program Evaluation: Forms and Approaches* (2nd ed.), St Leonards, NSW: Allen & Unwin.

Page, R. B. (1909). The Letters of Alcuin. New York: The Forest Press.

Patton, M. Q. (1990) *Qualitative evaluation and research methods.* Thousand Oaks, CA: Sage.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice. 15*(3), 311–319.

Patton, M. Q. (1996). A world larger than formative and summative. *Evaluation Practice, 17*(2), 131–144.

Patton, M.Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 35:5 Part II,* 1189–1208.

Pauly, M. V. (1974). Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection. *Quarterly Journal of Economics, 88,* 44–62.

Platt, J. R. (1964). Strong inference. *Science,* 146, 347–353.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Pugach, M. C. (2001). The stories we choose to tell: Fulfilling the promise of qualitative research for special education. *The Council for Exceptional Children, 67*(4), 439–453.

Ramsay, G. G. (1918). *Juvenal and Persius.* trans. New York: Putnam.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent groups design. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–206). Chicago, IL: Rand-McNally.

Reichardt, C. S. (2009). Quasi-experimental design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 46–71). Thousand Oaks, CA: Sage.

Reichardt, C. S., & Cook, T. D. (1979). Beyond qualitative versus quantitative methods. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 7–32). Beverly Hills, CA: Sage.

Reichardt, C. S., & Rallis, S. F. (1994b). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. *New Directions for Program Evaluation, 61*, 85–91.

Reichardt, C.S., & Rallis, S. F. (1994a). The relationship between the qualitative and quantitative research traditions. *New Directions for Program Evaluation, 61,* 5–11.

Reichenbach, H. (1938). *Experience and prediction.* Chicago: University of Chicago Press.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality & Quantity, 36*, 43–53.

Scriven, M. (1967). The methodology of evaluation. In Gredler, M. E., (Ed.), *Program Evaluation* (p. 16). Englewood Cliffs, New Jersey: Prentice Hall, 1996.

Scriven, M. (1976). Evaluation bias and its control. In C. C. Abt (Ed.) *The Evaluation of Social Programs,* (pp. 217–224). Beverly Hills, CA: Sage.

Scriven, M. (1983). Evaluation idiologies. In G.F. Madaus, M. Scriven & D.L. Stufflebeam (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 229–260). Boston: Kluwer-Nijhoff.

Scriven, M. (1991). Pros and cons about goal-free evaluation. *Evaluation Practice*, 12(1), 55–76.

Sechrest, L., & Figueredo, A. J. (1993). Program evaluation. *Annual Review of Psychology,* 44, 645–674.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. Education *Evaluation and Policy Analysis, 16*, 41–49.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (2001). *Foundations of program evaluations: Theories of practice.* Newberry Park, CA: Sage.

Shek, D. T. L., Tang, V. M. Y., & Han, X. Y. (2005). Evaluation of evaluation studies using qualitative research methods in the social work literature (1990–2003): Evidence that constitutes a wake-up call. *Research on Social Work Practice, 15,* 180–194.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smith, M. J. (2010). Handbook of program evaluation for social work and health professionals. New York: Oxford University Press.

St. Pierre, R. G. (2004). Using randomized experiments. In J.S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed., pp. 150–175). San Fransisco, CA: John Wiley & Sons.

Stevahn, L., King, J. A., Ghere, G. & Minnema, J. (2005). Establishing essential compentencies for program evaluators. *American Journal of Evaluation, 26,* 43–59.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

Stufflebeam,D. L. (2001). Evaluation models. *New Directions for Evaluation*, 89, 7–98.

Tang, H., Cowling, D.W., Koumijian, K., Roeseler, A., Lloyd, J., & Rogers, T. (2002). Building local program evaluation capacity toward a comprehensive evaluation. In R. Mohan, D.J. Bernstein, & M.D. Whitsett (Eds.), *Responding to sponsors and Stakeholders in Complex Evaluation Environments* (pp. 39–56). New Directions for Evaluation, No. 95. San Francisco, CA: Jossey-Bass.

Tharp, R., & Gallimore, R. (1979). The ecology of program research and development: A model of evaluation succession. In L. B. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation Studies Review Annual* (Vol. 4, pp. 39–60). Beverly Hills, CA: Sage.

Tharp, R., & Gallimore, R. (1982). Inquiry process in program development. *Journal of Community Psychology, 10*(2), 103–118.

The Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards* (2nd ed.). Thousand Oaks, CA: Sage.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *The Journal of Educational Psychology, 51*, 309–317.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Newbury Park, CA: Sage.

United Way. (1996). *Guide for logic models and measurements*. Retrieved May 18, 2012, from http://www.yourunitedway.org/media/GuideforLogModelandMeas.ppt

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics, 27*, 299–309.

Weiss, C.H. (1980) Knowledge creep and decision accretion. *Knowledge: Creation, Diffusion, Utilisation 1*(3): 381–404.

Weiss, C. H. (1998). *Evaluation: Methods for Studying Programs and Policies*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Weiss, C. J. (1999) The interface between evaluation and public policy. *Evaluation, 5*(4), 468–486.

Wells, G.L., Malpass, R.S., Lindsay, R.C.L., Fisher, R.P., Turtle, J.W., & Fulero, S.M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*(6), 581–598.

West, S.L., & O'Neal, K.K. (2004) Project D.A.R.E. outcome effectiveness revisited. *American Journal of Public Health, 94*(6) 1027–1029.

Williams, A. (2001). Science or marketing at WHO? Commentary on 'World Health 2000'. *Health Economics, 10*, 93–100.

Willson, E. B., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*, 249–258.

World Health Organization (2000). *The World Health Report 2000 – Health Systems: Improving Performance*. World Health Organization: Geneva, Switzerland.