An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs

Vivian C. Wong
Thomas D. Cook
W. Steven Barnett
Kwanghee Jung

Abstract

Since 1980, the number of state pre-kindergarten (pre-K) programs has more than doubled, with 38 states enrolling more than one million children in 2006 alone. This study evaluates how five state pre-K programs affected children's receptive vocabulary, math, and print awareness skills. Taking advantage of states' strict enrollment policies determined by a child's date of birth, a regression-discontinuity design was used to estimate effects in Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia. For receptive vocabulary, only New Jersey and Oklahoma yielded significant standardized impacts, though two of the three other coefficients were in a direction indicating positive effects. For math, all the coefficients were positive but only Michigan and New Jersey yielded reliable results. The largest impacts were for print awareness, where all five coefficients were positive and four were reliable in Michigan, New Jersey, South Carolina, and West Virginia. The five states were not randomly selected and, on average, have higher quality program standards than non-studied states, precluding formal extrapolation to the nation at large. However, our sample of states differed in many other ways, permitting the conclusion that state pre-K programs can have positive effects on children's cognitive skills, though the magnitude of these effects varies by state and outcome. © *2008 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

We take it as axiomatic that higher levels of human capital increase a nation's wealth and add vibrancy to its political and cultural life. Perhaps the best-validated determinant of human capital is the amount of engaged time that learners spend on age-appropriate cognitive tasks. This concept has spawned the advocacy of educational policies as diverse as increasing the time devoted to cognitive learning during each school day, adding to the length of the school day and to the number of school days per year, and increasing the fraction of students graduating from high school, attending college, or starting school at an even younger age (thus, before kindergarten). The evidence is overwhelming that preschool programs can work to increase performance in the early school grades (Campbell & Ramey, 1994; McCarton et al., 1997; Reynolds & Temple, 1995; Weikart, Bond, & McNeil, 1978). Evidence also suggests that these programs can positively affect later high school graduation rates, labor force participation, stable household formation, and criminal behavior (Barnett, 1995; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Currie, 1995; Ludwig & Miller, 2005; Schweinhart et al., 2005; for reviews, see Barnett, 1995;

Journal of Policy Analysis and Management, Vol. 27, No. 1, 122–154 (2008)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20310

Currie, 2001; Gormley, 2007; and Heckman & Masterov, 2005). However, studies demonstrating long-term effects tend to be small and local in scope, as are the studies showing only short-term effects; and two of the studies claiming long-term benefits left program implementation up to the program developer. So while it is well warranted that preschool programs can work, the main policy question is whether they do in fact work when they are implemented at scale. That is, when the program's reach is state-wide or federal, when its management is in the hands of education bureaucrats, and when its daily classroom implementation depends on local officials and teachers whose knowledge and motivation may not match those of the program developer's own staff. In the language of medical research, the evidence for the efficacy of preschool programs is stronger than the evidence for their effectiveness (Flay, 1986). In this paper, we use the term "preschool" to refer to all supervised early childhood education (ECE) programs, including private center-based programs and model early intervention programs; we use "state pre-kindergarten" (state pre-K) to identify state-funded or state-supervised programs often administered by local school districts; and we use Head Start to describe the largest single federally funded pre-K program.

Evaluations of state pre-K programs are rare and limited in scope, restricted to one city, Tulsa (Gormley, Gayer, Phillips, & Dawson, 2005; Gormley & Phillips, 2005), and to one state, Georgia (Henry et al., 2003). The paucity of relevant evidence reveals an important lacuna since the recent growth in pre-K enrollments has been at the state and not federal level. The number of state pre-K programs is now 38 and has more than doubled since 1980; the rate of enrollment is now steeper across the state programs than Head Start; and the total enrollment of 4-year-olds in state programs is over one million and exceeds the number of 4-year-olds in Head Start (Barnett, Hustedt, Hawkinson, & Robin, 2006). The main purpose of this paper is to estimate the effectiveness of a sample of state pre-K programs, given the large and growing importance of such programs on the national early education policy scene. The need is to know whether they are promoting the school readiness of children. Past reviews of the efficacy of pre-K efforts have emphasized school readiness in terms of academic achievement (Barnett, 1995; Currie, 2001; Gormley, 2007; Heckman, Stixrud, & Urzua, 2006) and so the present study assesses three academic achievement outcomes—Peabody Picture Vocabulary Test (PPVT) scores, print awareness scores, and math scores. However, increasing academic achievement is not the sole rationale for pre-K programs, and various social, behavioral, and health goals are also important. But we lack the data to examine them here and their links to human capital are not as clear cut as increases in early math, reading, and vocabulary.

States differ considerably in pre-K program attributes that might plausibly affect children's cognitive preparation. Some states provide for one year of education prior to kindergarten, and others two. Some pre-K programs are administered exclusively through state departments of education, while others also involve cooperation from state human services departments. Most states fund their services through a mix of funds from local school district, state, and federal sources, the last including Title I, Individuals with Disabilities Education Act, TANF, and even Head Start. However, the precise mix varies by state. There is broad consensus across states about the cognitive, social, behavioral, and mental health attributes of school readiness. But consensus is much lower about the priority of each of these domains and about the priority pre-K deserves relative to other state goals in education or elsewhere. As a result, the average state spending per enrolled child was about $3,482 in 2005–2006, but some states spent more than double this while others spent less than half of it (Barnett et al., 2006). States also vary in the mix of services they support. Some provide full-day services all year round for five days a week,

while others offer half-day services for only the academic year; some states provide voluntary universal access to pre-K services (UPK), while others target at-risk groups only; and some states require teachers to have a B.A. with specialized training in early childhood education, while others have no such requirements. This program variation suggests the need to examine whether states also vary in program effectiveness and to explore whether any state variation in outcomes is related to the quality of a state's pre-K offerings. This is the second focus of the paper.

Public policy is always concerned with alternatives and how well they perform relative to each other. It is possible to conceive of state pre-K programs and Head Start as competitors for national early education resources (Besharov & Higney, 2007b). So if it were shown that states with higher quality standards (such as higher teacher education requirements) outperformed Head Start, one could then argue either for raising Head Start quality standards to the level found in the most effective states (Barnett, 2007) or for increasing states' authority over the $7 billion in Head Start funds and, at the extreme, even advocating that these funds be sent to the states as block grants.[1] There is a well designed Head Start study (Puma, Bell, Cook, Heid, & Lopez, 2005) in which sites were first randomly selected from the national registry and applicant children were then randomly assigned to program or control status within each site. Unbiased, national estimates resulted from this two-stage procedure, and while commentators can, should, and do disagree about what these Head Start impact estimates mean for policy (Barnett, 2007; Besharov & Higney, 2007a; Currie, 2007), there is little quibble about the validity of the estimates themselves. But since no corresponding national estimates exist for state pre-K programs, policy analysts who want to compare the effectiveness of state programs and Head Start cannot do so. For many reasons that will be apparent later, we are suspicious of efforts to contrast effect sizes for different programs when different evaluation methods have been used for each program (Cook, 2006; Ludwig & Phillips, 2007). But we readily acknowledge that comparisons will be made between competing programs even when we prefer not to do so. So a minor purpose of this paper is to comment on what careful state-level estimates of effectiveness do and do not imply for making policy choices between state programs and Head Start.

Barnett (1995), Gilliam and Zigler (2001), and Gormley (2007) have all noted that evaluations of ECE programs tend to suffer from serious methodological limitations, of which selection bias is central because program participation might be endogenously related to some unobserved variable correlated with children's outcomes. Differential attrition, small sample sizes, and measurement problems are also common. Fortunately, each of the states examined in this study allocated children to pre-K based on their date of birth, thus permitting this date to be used as a cutoff score in a regression-discontinuity design (RDD). It has been shown, both theoretically (Goldberger, 1972a, 1972b) and empirically (Aiken, West, Schwalm, Carroll, & Hsuing, 1998; Black, Galdo, & Smith, 2005; Buddelmeyer & Skoufias, 2003; for review, see Cook & Wong, in press), that RDD can yield unbiased treatment effect estimates when the functional form of the relationship between the assignment and outcome variables is completely modeled (Goldberger, 1972a), and when misallocation about the cutoff is small or well modeled (Trochim, 1984). Realization of the virtues of RDD has been long delayed and uses of it have only recently entered into the public policy literature (Cook, in press). So an incidental purpose

[1] In 2003, the Bush Administration proposed to alter the Head Start grant-based program by converting it to a state block grant program, and the House of Representatives narrowly approved a measure to block-grant Head Start in as many as eight states by a vote of 217 to 216. The legislation was not enacted into law because the U.S. Senate did not vote on it before the congressional session ended. Most recently, Georgia Congressman Tom Price proposed a pilot project for eight states to take over their local Head Start programs—the same provision that helped stall the 2007 Head Start reauthorization bill.

of this study is to model for readers how an RDD should be carried out to ensure that its assumptions are met.

To sum up, this study has two major goals and two incidental purposes. The major goals are: (1) To estimate the effects of state-level pre-K programs; and (2) to relate the variability in these effects to the nature of the program implemented there. Its incidental purposes are: (1) To discuss the implications of any effect size differences between state and Head Start estimates of effectiveness; and (2) to model how an RDD analysis should be carried out. To address these purposes, we use pre-K data from five states—Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia. These states were not selected at random from the 38 states that currently have pre-K programs, and so they cannot support formal generalization to the nation at large. However, Table 1 shows that the five states vary in some attributes that are presumed to indicate the quality of pre-K services (for example, duration, funding level, eligibility requirements). The five-state sample is also much larger and more varied than is currently available in the literature evaluating state pre-K efforts. Moreover, within each state an attempt was made to randomly select both pre-K sites and children within sites. The selection of states is not optimal, but it is clearly superior to past samples based on a single city (Gormley et al., 2005; Gormley & Phillips, 2005) or state (Henry et al., 2003).

PRIOR STUDIES ON STATE PRE-KINDERGARTEN PROGRAMS

A meta-analysis of 13 state pre-K studies from 1977 to 1998 (Gilliam & Zigler, 2001) found that none used random assignment and three had no comparison groups. Of the other ten, none used adequate matching procedures and only half of them had pretest measures on the same scale as the outcome. The net result is quasi-experiments of the type that Cook and Campbell (1979) labeled "generally uninterpretable" causally. Nonetheless, Gilliam and Zigler concluded that state pre-K programs had positive impacts on children's cognitive development, school attendance, achievement, and retention.

More recently, two evaluations have examined how voluntary universal state pre-K programs affect children's school readiness skills. The Georgia Early Childhood Study (Henry et al., 2003) compared learning outcomes for probability samples from state pre-K, Head Start, and private preschool programs. Pretests were administered in the fall of the school year, thus helping to identify selection differences. Significant differences were observed between the three types of program. On average, the Georgia Head Start children had the lowest cognitive scores at pretest and lived in the most disadvantaged households; the private preschool children had the highest scores at pretest and lived in the most advantaged households. So the authors used instrumental variable (IV) and statistical matching techniques to try to control for selection. Their main conclusion was that, after a year of intervention, children in the state and private programs did not differ on any of the five language and cognitive skill outcomes, whereas the Head Start children performed less well than the other children on three of the tests. The problem here is to know whether all of the group selection differences were accounted for by use of the pretest.

The second study took place in Tulsa, Oklahoma, and is technically superior to the Georgia study from the standpoint of internal but not external validity (Gormley et al., 2005; Gormley & Phillips, 2005). An RDD was used that took advantage of a strict enrollment policy in Tulsa based on children's birthdays. Children with birthdays after a certain date were allowed to enroll in state pre-K while those with earlier birthdays were required to wait another year—a deterministic assignment process that enables complete modeling of the selection process into treatment and

Table 1. Key Components of State pre-K Programs in Our Sample (2004–2005 School Year).

State	Year Established	Average Amount State Spent on pre-K per Child	Number of 4-Year-Olds Served	% of 4 Year Olds Served	Teacher/ Child Ratio	Maximum Class Size	Duration	Teacher Education	Comprehensive Curriculum Standard
Michigan	1985	$5,031	24,729	19%	1:08	18	Half-day	BA degree for teachers in public schools	No
New Jersey Abbott	1998 standards raised in 2002	$10,361	21,286	79% of Abbott children*	2:15	15	Full day	BA degree with training in early education	Yes
Oklahoma	1990 universal in 1998	$6,167	30,180	65%	1:10	20	Varied	BA degree with training in early education	Yes
South Carolina	1984	$3,219	17,821	32%	1:10	20	Half-day	BA degree with training in early education	No
West Virginia	1983 universal by 2010	$6,829	6,541	33%	1:10	20	Varied	BA or AA degree with training in early education	Yes

* New Jersey's Abbott districts include about 1/4 of the state's children; statewide enrollment in Abbott and non-Abbott state pre-K was 25% at age 4.

control groups. Achievement test scores for 1,567 city children entering state pre-K were then compared with scores from 1,461 kindergarteners who had just completed pre-K. The analysts concluded that pre-K participation increased Woodcock-Johnson means for Letter-Word identification, Spelling, and Applied Problems and that minority students benefited from the program as much as others. The main concern here is with the generalization of results. They are limited to Tulsa, and it is not clear whether the services offered are representative of Oklahoma, let alone the United States. Tulsa is, after all, the largest and most urban school district in Oklahoma, and some evidence suggests that its state pre-K program is of exceptionally high quality. Phillips, Gormley, and Lowenstein (2007) compared the level of instructional and emotional support from state pre-K teachers in Tulsa, and the amount of time spent on pre-literacy and math activities there, to the levels obtained in a multi-state study of pre-K classrooms (Early et al., 2005). The Tulsa classrooms scored significantly higher on all four dimensions of instructional support and on one of four dimensions of emotional support. Also, more time was spent there on reading and literacy, on math, and on science, leading the authors to characterize Tulsa as a national example of high quality pre-K programs. The implication is that Tulsa should not be seen as representative of the quality of state pre-K programming in the nation at large. Since neither Tulsa not Georgia offers much traction for broad generalization, it is important to examine the effects of state pre-K with a broader sample and a methodology capable of warranting strong causal inference.

METHODOLOGY

The Sampling Design

The sampling plan was designed and implemented by researchers at the National Institute for Early Education Research (NIEER). It has three levels: states, classrooms within states, and children within classrooms. Officials in states with a pre-K program were solicited to participate in the study and five agreed to do so, offering their support and cooperation for the study. Table 1 provides state-by-state summaries of the number of 4-year-olds enrolled, percentage of 4-year-olds in the state served, duration of pre-K program, and requirements for maximum class size, teacher education, and comprehensive curriculum standards. The table shows that the sampled states are quite diverse with respect to program duration, funding levels, and eligibility requirements for enrollment. But they also have high quality standards in terms of paying teachers on public school scales and requiring pre-K teachers to have at least B.A. degrees—and they tended to be more mature. New Jersey is the exception here, since its program was created in 1998 and its standards were substantially raised in 2002, thus becoming the highest cost program in the nation and of special interest for this feature alone. This is clearly an opportunistic sample of states with a volunteer bias and truncated at the lower end of the national distribution. But the sample is nonetheless quite heterogeneous in the kinds of pre-K offerings available from state to state.[2]

In four of the five states, pre-K classrooms were first randomly selected from a list of the total number of state-funded pre-K classrooms. Then the same number of kindergarten classrooms was sampled within the districts from which the pre-K classrooms had been selected. Children were then randomly selected within classrooms. For the fifth state, New Jersey, a stratified random sample of classrooms was selected within the state's largest pre-K program. Stratification was based on factors like district enrollment, geographic location, urban versus rural setting, and the

[2] Since these data were collected in 2004, three additional states, Arkansas, California, and New Mexico, have agreed to participate and are in varying stages of doing so.

percentage of bilingual students. Only pre-K programs targeted at 4-year-olds were sampled. Compliance was not perfect, with some districts, schools, and classrooms refusing to participate. In Michigan, the Detroit school district granted permission too late in the year to be included in this study; and in West Virginia, 41 percent of the students initially selected opted not to participate. Where refusals were substantial, more classrooms and students were added to a state's sample, though not always at random. As a result, the child samples do not perfectly represent all students enrolled in a state's pre-K programs, even though they are more representative than in prior pre-K studies. Below, we provide an overview of the five state pre-K programs evaluated in this study.

The Michigan School Readiness Program (MSRP)

Targeting only at-risk 4-year-olds, at least half or more of the children at each site either had to meet an income eligibility criterion and also exhibit one other risk factor from a list of 25, or they had to exhibit more than one of these 25 risk factors. Pre-K programs took place in public schools, Head Start programs, and private care centers, and each site was open for at least half the school day and for at least 30 weeks per year. In the 2004–2005 school year, Michigan spent $84 million on MSRP, or about $3,366 per student, though this is only the state's contribution and does not include funding from local and federal sources (Barnett, Hustedt, Robin, & Schulman, 2005).[3] From K-12 spending patterns in Michigan, we estimate that *total* expenditure per child was approximately $5,000.

New Jersey's Abbott Preschool Program

As a result of a 1998 state Supreme Court ruling, the New Jersey Abbott Program provides voluntary pre-K for 3- and 4-year-olds in school districts where at least 40 percent of children qualified for subsidized lunch at the time of the ruling. The Abbott program is one of three state-funded pre-K initiatives, and the state Supreme Court ruling resulted in the implementation of much higher quality standards in the program beginning in 2002. In addition to requirements for maximum class size and teacher education, the court order included a provision for coaches to help teachers improve their classroom practice. The other two New Jersey pre-K programs have lower funding levels and fewer numbers of students. Thus, the Abbott program served 19 percent of the state's 4-year-olds in 2005, while the other two pre-K programs served 7 percent. Our results apply only to the Abbott program. To supplement the state Department of Education's funding for 6 hours of the school day during the 180 day school year, the Human Services Department provides additional funding for wraparound child care services for up to 10 hours a day, 5 days a week, all year round. In the 2004–2005 school year, New Jersey spent $400 million on its Abbott program, or about $10,361 per student (Barnett et al., 2005). This is one of the few state pre-K programs funded entirely by the state.

Oklahoma's Early Childhood Four-Year-Old Program

In 1980, Oklahoma began providing pre-K services for 4-year-olds on a pilot basis. Ten years later, the program was broadened to include all 4-year-olds eligible for Head Start. But in 1998, Oklahoma became only the second state to offer free voluntary pre-K to all 4-year-olds.[4] Participation in pre-K increased steadily over the

[3] As Barnett et al. (2005) write in their annual report on state pre-kindergarten programs, there are numerous limitations to identifying all pre-K funding sources at the local, state, and federal levels.

last decade and since 2002, Oklahoma enrolled a greater percentage of its 4-year-olds than any other state. Most children were served in public schools, though districts could also collaborate with private childcare or Head Start centers to provide services. In the 2004-2005 school year, the state spent over $100 million on pre-K education, approximately $3,500 per child, though the state school formula relies on local schools' support for a portion of their funding. Expenditure per child from all sources was estimated to exceed $6,100 per child (Barnett et al., 2005, 2006).

South Carolina's Early Childhood Programs

South Carolina's state pre-K initiative comprises two programs, the Half-Day Child Development Program (4K), and the First Steps to School Readiness initiative. Funds from First Steps are used to supplement 4K, such as by adding new pre-K classes or serving additional children in half-day classes. Although eligibility for the state pre-K program is determined at the district level, it is based on a list of risk factors identified by the state. Poverty is one such factor. Most children were served in the public school system, though some services were provided in Head Start centers or private child care centers through public-private partnerships. Programs operated for about 2.5 hours per day, 5 days per week for the academic year. About 15 percent of programs used additional district, state, and federal funds to provide full day pre-K. In the 2004–2005 school year, the South Carolina state legislature spent about $24 million on early childhood education, or about $1,400 per child (Barnett et al., 2005). Even with the expected local contributions, the funding level in South Carolina is one of the lowest in the country at an estimated $3,219 per child (Barnett et al., 2006).

West Virginia Early Childhood Education Program

The West Virginia state pre-K program began in 1983 when a revision in the school board code allowed local districts to create pre-K programs. Currently, the state is in the process of expanding access, with the goal of providing voluntary universal pre-K to all 4-year-olds. Eligibility for 4-year-olds is determined at the local level, with some counties enrolling students on a first come/first serve basis or by lottery. Children are served in a variety of settings, including public schools, Head Start centers, and child care and private preschool centers. Pre-K programs last for the academic year, but the hours of operation vary by site. Typical programs operate for nine months a year, two full days per week, or four full days with Fridays reserved for home visits and planning. In the 2004–2005 school year, West Virginia spent $34.5 million on state pre-K education, or $4,323 per child (Barnett et al., 2005), with total funding from all state and local sources amounting to at least $6,829 per child enrolled (Barnett et al., 2006).

A major implication of these descriptions is that states differ in programmatic ways that might plausibly affect achievement (whole- or half-day programs; targeted versus universal eligibility). Another is that they differ in methodological features that can affect conclusions, as in the level of noncompliance with random selection

[4] Georgia was the first state to enact legislation that offered voluntary universal pre-K to 4-year-olds, but enrollment figures suggest that in practice, Oklahoma was the first state to offer voluntary universal pre-K to all. Funding for the Georgia program was limited by monies that could be made available through the state lottery system, while Oklahoma funded any 4-year-old that school districts could enroll. Thus, from 2004 to 2006, Georgia enrollment rates of 4-year-olds remained stagnant at 55, 55, and 51 percent (respectively), while Oklahoma's enrollment rate grew steadily, from 64 percent in 2004 to 68 percent in 2005 and 70 percent in 2006 (Barnett et al., 2004, 2005, 2006).

and differences in state sample sizes and hence, statistical power. State samples ranged from 2,072 students in New Jersey to 720 students in West Virginia. With only five states, policy differences are bound to be partly confounded with methods.

DATA COLLECTION PROCEDURE

In each state, we worked with a local research partner to train child assessors on issues related to testing children in school environments, confidentiality, protocol, and professional etiquette, as well as training specific to the assessment measures and sampling procedures. Assessors were trained on each measure and then shadow scored in practice measures. Site coordinators were responsible for assuring adequate reliability throughout the study. A liaison at each site gathered information on the children's pre-K status, usually from existing school records but occasionally from parent reports, and was reimbursed $5 per child for obtaining the information.

Children were tested in early fall of the 2004–2005 school year. On all measures, children were tested in English or Spanish, depending on their strongest language, which was ascertained from the classroom teacher. A very small number of children who did not speak either English or Spanish well enough to be tested were not included in the sample. Assessments were conducted one-on-one in the child's school, and assessments were scheduled to avoid meal, nap, and outdoor playtimes. Testing sessions lasted 20–40 minutes.

Individualized assessments were selected to measure the contributions of the pre-K programs to children's learning, with emphasis on skills important for early school success. Criteria for selection of measures included: (1) availability of equivalent tasks in Spanish and English; (2) reliability and validity, particularly pre-literacy skills that are good predictors of later reading ability; and (3) appropriateness for children ages 3 to 5. Although it would have been highly desirable to have measures of social and emotional development, most such instruments have teachers rate children relative to their age (school year) cohort. This approach is incompatible with the RDD approach. Each measure is discussed in detail below.

MEASURES OF SCHOOL READINESS

Children's receptive vocabulary was measured by the Peabody Picture Vocabulary Test, 3rd Edition (PPVT-3) (Dunn & Dunn, 1997). The PPVT-III is a 204-item test in standard English administered by having children point to one of four pictures shown when given a word to identify. The PPVT-III directly measures vocabulary size and the rank order of item difficulties is highly correlated with the frequency with which words are used. This test is also used as a quick indicator of general cognitive ability, and it correlates reasonably well with other measures of linguistic and cognitive development related to school success. Children tested in Spanish were given the Test de Vocabulario en Imagenes Peabody (TVIP) (Dunn, Padilla, Lugo, & Dunn, 1986). The TVIP uses 125 translated items from the PPVT to assess receptive vocabulary acquisition of Spanish-speaking and bilingual students.

The PPVT has been used for many years (over several versions) and substantial information is available on its technical properties. Reliability is good as judged by either split-half reliabilities or test-retest reliabilities. The test is adaptive in that the assessor establishes a floor which the child is assumed to know all the answers and a ceiling above which the child is assumed to know none of the answers. This is important for avoiding floor and ceiling problems (Rock & Stenner, 2005).

Children's early mathematical skills were measured with the Woodcock-Johnson Tests of Achievement, 3rd Edition (Woodcock, McGrew & Mather, 2001) Subtest 10 Applied Problems. Spanish-speakers were given the Bateria Woodcock-Munoz pruebas de Aprovechamiento-Revisado (Woodcock & Munoz, 1990) Prueba 25, Problemas Aplicados. The manuals report good reliability for the Woodcock-Johnson achievement subtests, and they have been widely and successfully used in studies of the effects of preschool programs including Head Start.

Print Awareness abilities were measured using the print awareness subtest of the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPP) (Lonigan, Wagner, Torgeson & Rashotte, 2002). The Pre-CTOPPP was designed as a downward extension of the Comprehensive Test of Phonological Processing (CTOPP) (Wagner, Torgeson & Rashotte, 1999), which measures phonological sensitivity in elementary school-aged children. Although not yet published, the Pre-CTOPPP has been used with middle-income and low-income samples and includes a Spanish version. Print awareness items measure whether children recognize individual letters and letter-sound correspondences, and whether they differentiate words in print from pictures and other symbols. The percentage of items answered correctly out of 36 total subtest items is reported. As the Pre-CTOPP has only been very recently developed, very little technical information is available about its performance and psychometrics properties.

DATA ANALYSIS: GENERAL APPROACH

For each state, the analysis takes advantage of the pre-K program's deterministic enrollment depending only on a child's birth date. Children with birthdates after the state cutoff were permitted to enroll in pre-K, but those with birthdays before it were required to wait another year. So the treatment groups consisted of children who had completed pre-K in spring 2004 and were starting kindergarten in fall 2004. Comparison children were 4-year-olds just starting pre-K in fall 2004. Table 2 provides descriptive statistics for sample members in each state, including the number of treatment and comparison group members, and the percentages of students who were minorities or received free/reduced-price lunch. To check the fit between this desired assignment process and the assignment actually achieved, Figure 1 shows that the percentage of children enrolled in pre-K increased precipitously at the cutoffs for all five states. More than 90 percent of students with birthdates after the cutoff entered their state pre-K program, and fewer than 6 percent of those with birthdays before the cutoff were enrolled. So, the cutoff rules were well implemented. Even so, implementation was not perfect and some treatment misallocation occurred, though only from 1 percent to 8 percent across the states, as Table 2 indicates.

One way to conceptualize RDD is in terms of modeling the selection process via the regression line that describes how the assignment and outcome variables are related. In the untreated portion of the assignment variable, this regression line serves as the counterfactual against which to interpret whether the level or the slope changes at the cutoff. Two internal validity threats have then to be dealt—incorrect specification of the functional form of the regression line, and treatment misallocation near the cutoff. When the response function is incorrectly specified, such as when a true cubic model is fitted with a linear one, a spurious effect may be detected at the cutoff. So the data analysis has to provide evidence that the functional form has been correctly specified and that treatment misallocation around the cutoff is minor or has been well modeled.

The second conceptualization of RDD views it as akin to a randomized experiment near the cutoff. The relevant justification is that the difference between

Table 2. Summary Statistics for State Samples.

	N	PPVT	Math	Print Awareness	Fuzzy Cases	Black	Hispanic	Native American	White/ Asian	Other	Race Missing	Girl	No Free Lunch	Free Lunch	Free Lunch Missing	TVIP
Michigan	871	58.87	13.03	53.59	2%	22%	10%		53%	4%	10%	54%	28%	49%	23%	
		(19.14)	(4.85)	(30.35)	(0.15)	(0.41)	(0.31)		(0.50)	(0.21)	(0.30)	(0.50)	(0.45)	(0.50)	(0.42)	
Comparison	386	51.31	10.54	35.17	0%	26%	8%		53%	5%	7%	54%	28%	50%	22%	
		(16.93)	(3.91)	(23.05)	(0.05)	(0.44)	(0.27)		(0.50)	(0.23)	(0.26)	(0.50)	(0.45)	(0.50)	(0.42)	
Treatment	485	68.28	16.19	76.41	5%	17%	13%		53%	3%	13%	53%	27%	48%	25%	
		(17.50)	(4.02)	(21.52)	(0.22)	(0.38)	(0.34)		(0.50)	(0.18)	(0.34)	(0.50)	(0.45)	(0.50)	(0.43)	
New Jersey	2072	49.60	11.84	62.33	4%	25%	39%		14%	2%	19%	51%	22%	68%	10%	6%
		(19.97)	(4.56)	(28.90)	(0.20)	(0.44)	(0.49)		(0.35)	(0.15)	(0.39)	(0.50)	(0.41)	(0.47)	(0.30)	(0.24)
Comparison	895	39.21	9.39	44.15	4%	28%	44%		12%	3%	14%	50%	17%	71%	12%	7%
		(17.26)	(3.84)	(26.52)	(0.19)	(0.45)	(0.50)		(0.32)	(0.17)	(0.35)	(0.50)	(0.37)	(0.45)	(0.33)	(0.26)
Treatment	1177	57.45	13.68	75.07	5%	24%	36%		16%	2%	22%	51%	26%	65%	9%	6%
		(18.22)	(4.17)	(23.11)	(0.21)	(0.43)	(0.48)		(0.37)	(0.14)	(0.42)	(0.50)	(0.44)	(0.48)	(0.28)	(0.23)
Oklahoma	838	65.97	14.89	65.30	4%	7%	7%	13%	65%	1%	8%	51%	32%	50%	18%	2%
		(18.88)	(4.47)	(29.27)	(0.19)	(0.26)	(0.25)	(0.33)	(0.48)	(0.10)	(0.26)	(0.50)	(0.47)	(0.50)	(0.39)	(0.14)
Comparison	407	57.59	12.53	47.72	0%	7%	5%	12%	68%	1%	7%	54%	34%	44%	22%	2%
		(17.50)	(3.90)	(26.94)	(0.07)	(0.26)	(0.23)	(0.32)	(0.47)	(0.10)	(0.25)	(0.50)	(0.47)	(0.50)	(0.42)	(0.14)
Treatment	431	73.79	17.12	81.69	7%	7%	8%	13%	61%	1%	8%	47%	30%	55%	15%	2%
		(16.65)	(3.77)	(20.57)	(0.25)	(0.26)	(0.28)	(0.34)	(0.49)	(0.11)	(0.28)	(0.50)	(0.46)	(0.50)	(0.36)	(0.14)
South Carolina	777	58.55	NA	62.18	1%	44%			40%	4%	13%	51%	35%	54%	11%	
		(19.28)	NA	(29.90)	(0.09)	(0.50)			(0.49)	(0.19)	(0.34)	(0.50)	(0.48)	(0.50)	(0.31)	
Comparison	424	50.44	NA	45.17	1%	45%			37%	4%	15%	51%	35%	50%	15%	
		(17.62)	NA	(26.79)	(0.12)	(0.50)			(0.48)	(0.19)	(0.36)	(0.50)	(0.48)	(0.50)	(0.36)	
Treatment	353	68.12	NA	82.07	0%	42%			44%	3%	10%	52%	35%	59%	6%	
		(16.59)	NA	(19.14)	(0.05)	(0.49)			(0.50)	(0.18)	(0.30)	(0.50)	(0.48)	(0.49)	(0.24)	
West Virginia	720	68.01	14.62	62.08	8%				89%	5%	6%	50%	14%	33%	53%	
		(18.43)	(4.85)	(30.53)	(0.27)				(0.31)	(0.22)	(0.24)	(0.50)	(0.35)	(0.47)	(0.50)	
Comparison	341	58.78	11.88	40.52	6%				87%	6%	7%	54%	13%	39%	48%	
		(17.32)	(4.12)	(24.31)	(0.24)				(0.33)	(0.24)	(0.25)	(0.50)	(0.34)	(0.49)	(0.50)	
Treatment	379	76.27	17.04	80.45	10%				90%	4%	6%	46%	15%	28%	57%	
		(15.21)	(4.11)	(22.12)	(0.30)				(0.30)	(0.20)	(0.23)	(0.50)	(0.36)	(0.45)	(0.50)	

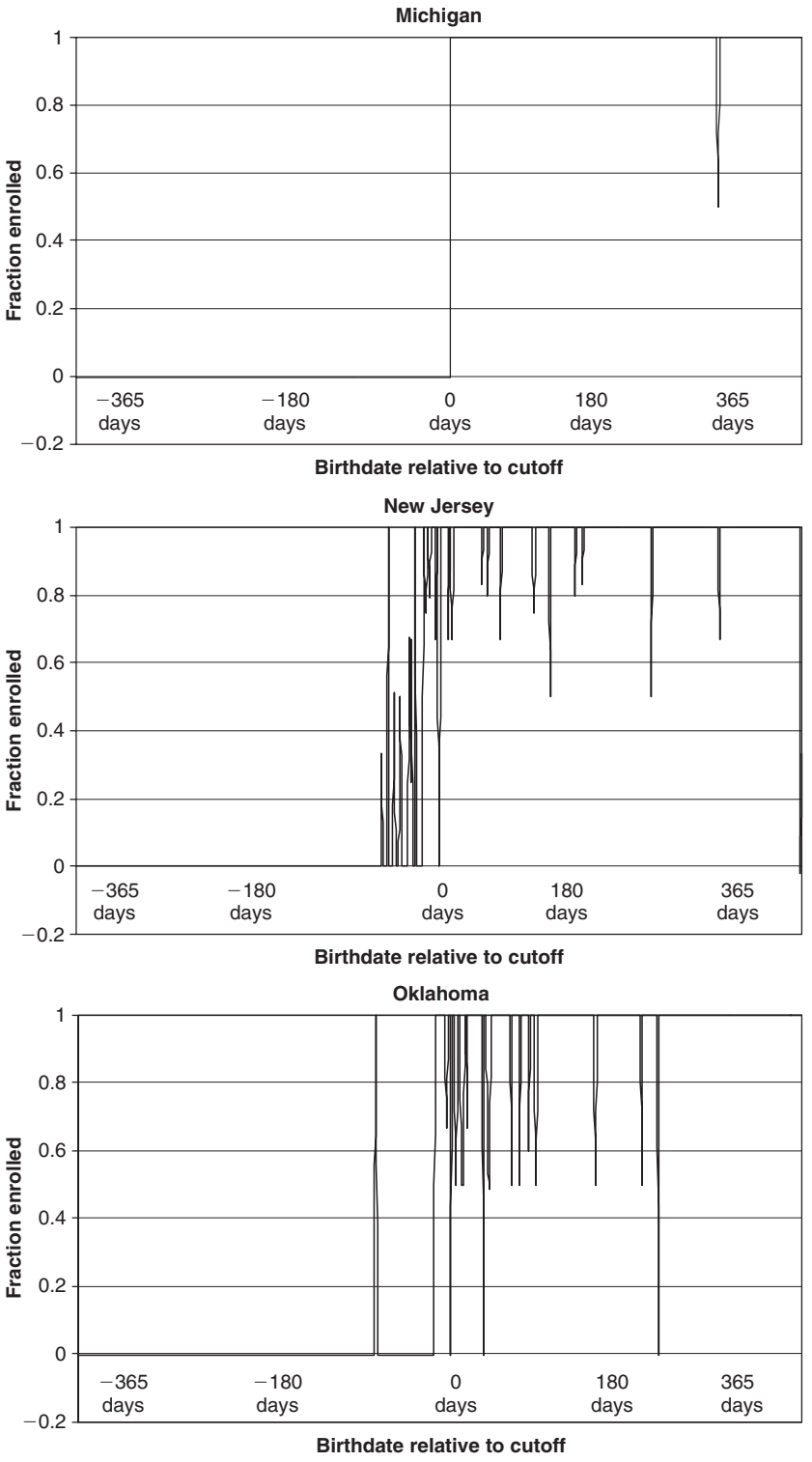

Figure 1. Relationship between children's birthdates relative to cutoff and state preschool enrollment.

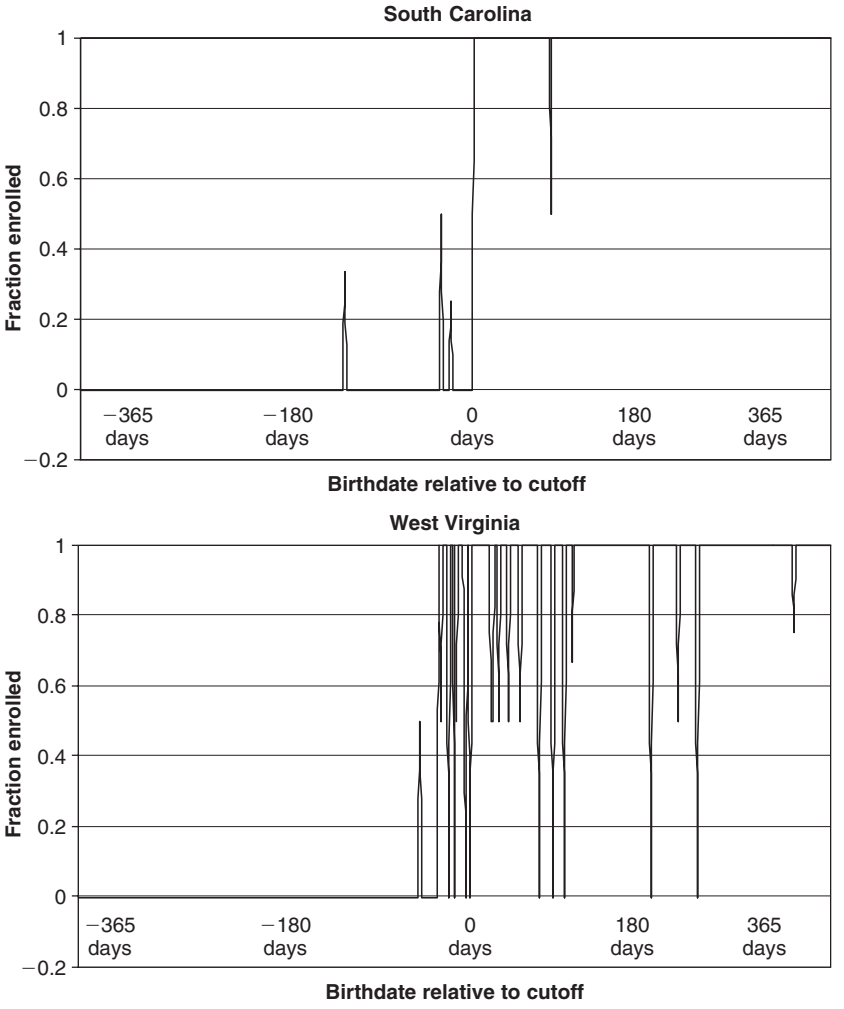

Figure 1. Continued.

students with birthdays one day apart on different sides of the cutoff is almost entirely due to chance—the very treatment assignment mechanism from which the randomized experiment draws its interpretative power. Impact estimates can then be calculated as mean differences immediately each side of the cutoff, or as close to it as is required for a well powered test. This approach severely reduces the need to specify the functional form linking the assignment and outcome variables along all the assignment range, but it depends on treatment misallocation being minimal, on dense sampling around the cutoff, and on a strong justification for the local average treatment effect (LATE) that is estimated at the cutoff, for it cannot be generalized elsewhere along the assignment variable.

However conceptualized, RDD is less efficient than a randomized experiment for detecting the same treatment effect (Cappelleri, Darlington, & Trochim, 1994). Holding sample size constant, RDD will have higher standard errors and so reject

the null hypothesis less often. Under the conditions incorporated into his simulation, Goldberger (1972b) found that randomized experiments are more efficient than RDD by a factor of 2.75. The power of RDD also varies with other factors not included in Goldberger's work, but the design has never been shown to be as efficient as a randomized experiment (Shadish, Cook, & Campbell, 2002). We will have to bear this power issue in mind when interpreting the results from states with smaller samples.

DATA ANALYSIS: SPECIFICS

Our RDD analysis focuses on efforts to model the functional form of the assignment and outcome variables, and below we present a variety of analytic techniques used to identify the correct response function for each outcome in each state. The analytic plan has three components: a graphical analysis, a series of parametric regressions with alternate specifications, and nonparametric procedures using local linear kernel regression (Hahn, Todd, & van der Klaauw, 2001).

To indicate the true functional form, detailed graphical analysis is essential (Trochim, 1984). We begin with simple graphs of each outcome in each state. As shown in Figure 2, two types of lines are fitted onto the scatterplots each side of the cutoff. Plot 1 depicts a linear regression line, and plot 2 shows a nonparametric regression line based on locally weighted scatterplot smoothing, called lowess, that relaxes assumptions about the form of the relationship between the assignment and outcome (Cleveland & Devlin, 1988). For each y_i, a smoothed value is obtained by weighted regressions involving only those observations within a local interval. Observations closer to y_i are weighted more heavily than those farther away. Figure 2 depicts linear regression and lowess plots for New Jersey's PPVT. We observe that the parametric model is indeed a close approximation to the lowess, suggesting that a linear model is likely appropriate for PPVT in New Jersey. If we had observed evidence of nonlinearity in the lowess, we would then have compared it with graphs of quadratic or cubic models as part of a plan to determine whether these higher order models are better specification choices.

We next run a series of regressions to obtain parametric estimates of the treatment effect. To describe the causal relationship of state *pre-K* participation on children's achievement scores, we model the latter. For the ith individual in classroom *j*, we write:

$$Y_{ij} = BX_{ij} + \beta_1(Pre\text{-}K)_{ij} + g(AV)_{ij} + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is student i's outcome, X_{ij} is a vector of student characteristics including gender, race/ethnicity, whether the child receives free or reduced price lunch, and whether the child took English or Spanish versions of tests. Pre-K_{ij} is a dichotomous indicator variable such that $T = 1$ for treatment and $T = 0$ for no treatment, and $g(AV)_{ij}$ is a smooth function of the continuous assignment variable.

We check for robustness of our estimates by considering a number of alternative specifications for $g(AV)_{ij}$, including polynomials and interaction terms. The order of the polynomial approximation to the $g(AV)_{ij}$ function is determined by examining the statistical significance of the higher order and interaction terms. Following Trochim (1984), when the functional form of the regression model is ambiguous, we overfit the model by including more polynomial and interaction terms than needed, yielding unbiased but less efficient estimates. In presenting the actual results later, Tables 4 through 8 will provide impact estimates using linear, quadratic, and cubic models.

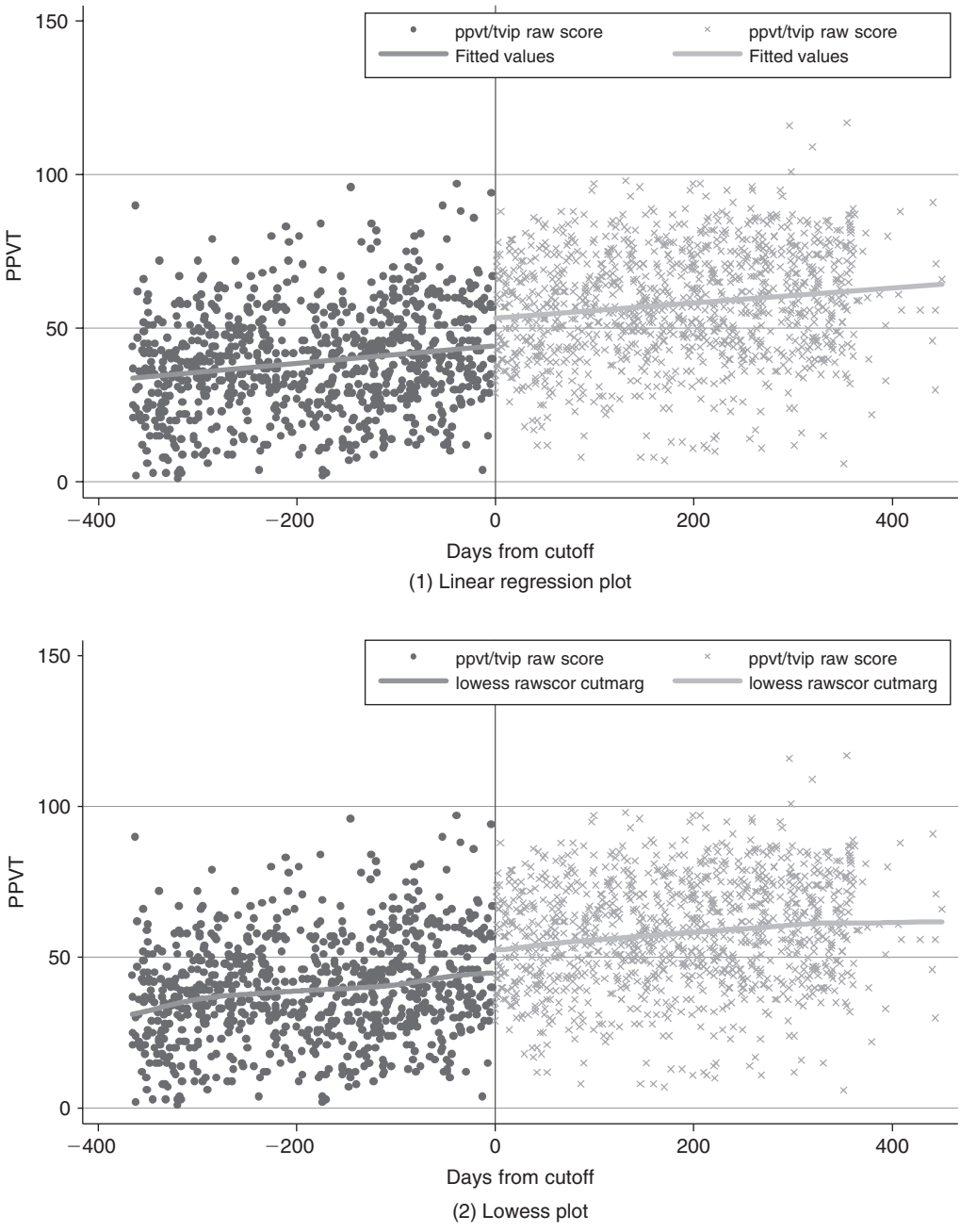

Figure 2. Examples of lowess and linear plots of New Jersey's PPVT.

As a final parametric check on functional form, we truncate the dataset to include only observations near the cutoff. In placing greater weight on these, we eliminate the influence of extreme assignment variable values that often play a disproportionate role in misspecifying functional form. So we rerun the parametric analyses

Table 3. Functional Form of Parametric Estimates.

	PPVT	Math	Print Awareness
Michigan	quadratic	linear	linear
New Jersey	linear	linear	cubic
Oklahoma	linear	cubic	cubic
South Carolina	linear		linear
West Virginia	linear	quadratic	linear

including only those children with birthdates within six months each side of the cutoff.[5] In all the parametric analyses, we use Huber-White standard errors adjusted for clustered data at the classroom level.

The final strategy to deal with misspecified functional form is to conduct nonparametric analyses. For these, we use simple differences of smoothed versions of the kernel estimator generated by local linear regression (Hahn et al., 2001) rather than simple differences of kernel estimates generated each side of the discontinuity (as in Buddelmeyer & Skoufias, 2003). These estimates require that, within a given interval on the assignment variable, weighted regressions are run using the same weights as for kernel estimates but including an additional linear term in the weight so as to converge more quickly at the boundaries and produce less biased estimates at the cutoff (Pagan & Ullah, 1999; Hahn et al., 2001). Unbiased nonparametric estimates depend on proper specification of the interval, or bandwidth, within which local regressions are carried out. The narrower these bandwidths are, the less biased are the estimates they yield. But the estimates are then also less efficient because only observations close to the point at which the predicted mean is calculated receive weight. Wider bandwidths use more observations to calculate the bandwidth mean, but the estimates they produce may be less consistent. So we estimated treatment impacts using a variety of bandwidths, but present here estimates for just the two bandwidth choices that appear to best balance the bias-efficiency tradeoff. Our nonparametric impact estimates are simple mean differences of smooth outcomes on each side of the discontinuity. These are the predicted means immediately on the right and left sides of the cutoff, with each mean computed using weighted observations in the chosen bandwidth interval on the assignment variable. Standard errors for predicted means were calculated using bootstrapping techniques (500 repetitions). Significant differences for the treatment and comparison groups were determined through a series of t-tests of predicted means for observations near the cutoff. The state-of-the-art is still uncertain for some nonparametric issues in RDD, especially as concerns hypothesis testing and the consistency of estimates at the boundaries. In general, we attempted to follow procedures used by Black et al. (2005).

Going back to the parametric estimates, Table 3 summarizes the regression models we ultimately determined to be most appropriate for each outcome in each state. In 13 of 14 cases, we chose the functional form best predicting the outcome—with the largest, or equal to largest, adjusted R-square value. The exception (New Jersey Math) involved a miniscule difference between the linear and quadratic models (.0009) because additional analyses indicated that a linear specification was more appropriate. For the PPVT outcome, a linear specification described the response function best for all states except Michigan, where a quadratic function prevailed.

[5] We also truncated the sample to include children only three months each side of the cutoff, but there were too few observations to reliably estimate the regression line.

Table 4. Michigan.

	Empirically Identified Functional Form (1)	Parametric Models Used in Analysis			Truncated at 6 Months (5)	Non-parametric Estimates by Bandwidth		IV Adjusted Estimates (8)
		Linear (2)	Quadratic (3)	Cubic (4)		50 BW (6)	75 BW (7)	
PPVT	Quadratic	0.332	−1.911	−3.181	−3.741	−4.990	−1.655	**−2.747**
		(2.488)	(4.153)	(5.617)	(6.096)	(4.400)	(3.914)	**(4.531)**
Math	Linear	2.032*	2.251*	2.474	1.905*	2.990*	2.349*	**1.820***
		(0.562)	(0.902)	(1.289)	(0.872)	(0.863)	(1.017)	**(0.483)**
Print Awareness	Linear	24.978*	21.579*	21.745*	21.790*	19.313*	22.187*	**22.139***
		(3.578)	(5.679)	(7.766)	(5.582)	(5.993)	(5.155)	**(3.105)**

Robust standard errors in parentheses.
* significant at 5%
Our preferred TOT estimates are in bold.
We used Epanechnikov kernel function for non-parametric estimates. Non-parametric estimates have bootstrapped standard errors (reps = 500).

Table 5. New Jersey.

	Empirically Identified Functional Form (1)	Parametric Models Used in Analysis			Truncated at 6 Months (5)	Non-parametric Estimates by Bandwidth		IV Adjusted Estimates (8)
		Linear (2)	Quadratic (3)	Cubic (4)		30 BW (6)	40 BW (7)	
PPVT	Linear	5.705*	5.368*	5.256	4.975*	8.094*	7.955*	**6.101***
		(1.438)	(2.019)	(2.715)	(1.925)	(3.861)	(2.637)	**(1.436)**
Math	Linear	0.715*	0.077	0.377	0.268	.392	.494	**0.867***
		(0.352)	(0.469)	(0.596)	(0.463)	(0.707)	(0.710)	**(0.363)**
Print Awareness	Cubic	17.159*	11.921*	9.252	6.299	8.704	8.250	**13.019***
		(2.471)	(3.726)	(4.828)	(6.679)	(5.646)	(6.532)	**(5.848)**

Robust standard errors in parentheses.
* significant at 5%.
Our preferred TOT estimates are in bold.
We used Epanechnikov kernel function for non-parametric estimates. Non-parametric estimates have bootstrapped standard errors (reps = 500).

Table 6. Oklahoma.

	Empirically Identified Functional Form (1)	Parametric Models Used in Analysis			Truncated at 6 Months (5)	Non-parametric Estimates by Bandwidth		IV Adjusted Estimates (8)
		Linear (2)	Quadratic (3)	Cubic (4)		50 BW (6)	75 BW (7)	
PPVT	Linear	5.648*	5.074	1.710	0.268	1.263	4.333	**5.117***
		(2.350)	(3.599)	(4.563)	(0.463)	(5.601)	(3.895)	**(2.308)**
Math	Cubic	2.011*	2.167*	0.483	0.296	.337	1.204	**1.358**
		(0.557)	(0.740)	(1.040)	(1.268)	(1.389)	(0.984)	**(0.903)**
Print Awareness	Cubic	21.013*	15.549*	9.247	0.465	1.039	10.065	**11.464**
		(3.516)	(4.841)	(6.907)	(9.700)	(12.379)	(7.029)	**(6.001)**

Robust standard errors in parentheses.
* significant at 5%.
Our preferred TOT estimates are in bold.
We used Epanechnikov kernel function for non-parametric estimates. Non-parametric estimates have bootstrapped standard errors (reps = 500).

Table 7. South Carolina.

	Empirically Identified Functional Form (1)	Parametric Models Used in Analysis			Truncated at 6 Months (5)	Non-parametric Estimates by Bandwidth		IV Adjusted Estimates (8)
		Linear (2)	Quadratic (3)	Cubic (4)		50 BW (6)	75 BW (7)	
PPVT	Linear	0.985 (2.327)	−0.187 (3.176)	−0.219 (4.356)	1.362 (3.033)	.088 (3.864)	−1.818 (5.483)	**0.795** **(2.351)**
Print Awareness	Linear	21.072* (2.909)	21.716* (4.380)	22.831* (5.966)	25.318* (4.153)	21.239* (5.017)	18.512* (6.402)	**21.005*** **(2.928)**

Robust standard errors in parentheses.
* significant at 5%.
Our preferred TOT estimates are in bold.
We used Epanechnikov kernel function for non-parametric estimates. Non-parametric estimates have bootstrapped standard errors (reps = 500).

Table 8. West Virginia.

	Empirically Identified Functional Form (1)	Parametric Models Used in Analysis			Truncated at 6 Months (5)	Non-parametric Estimates by Bandwidth		IV Adjusted Estimates (8)
		Linear (2)	Quadratic (3)	Cubic (4)		50 BW (6)	75 BW (7)	
PPVT	Linear	3.655 (2.387)	2.527 (3.469)	8.249 (4.526)	3.994 (3.112)	4.884 (5.792)	6.39 (4.250)	**2.422** **(1.940)**
Math	Quadratic	1.937* (0.634)	1.530 (0.940)	1.244 (1.349)	0.769 (1.444)	0.764 (1.495)	1.743 (0.953)	**0.435** **(1.393)**
Print Awareness	Linear	24.491* (3.496)	28.024* (5.032)	28.445* (6.381)	27.015* (5.097)	30.488* (5.471)	30.950* (4.969)	**20.150*** **(2.980)**

Robust standard errors in parentheses.
* significant at 5%.
Our preferred TOT estimates are in bold.
We used Epanechnikov kernel function for non-parametric estimates. Non-parametric estimates have bootstrapped standard errors (reps = 500).

For math, response functions were linear for Michigan and New Jersey and cubic and quadratic for Oklahoma and West Virginia, respectively. For print awareness, the response function was linear in three states (Michigan, South Carolina, and West Virginia) and cubic in two others (New Jersey and Oklahoma).

One might expect such variation in functional forms, given the many ways in which state programs differ. For instance, if states varied in the distribution of children's ages, then achievement floor effects might be evident for the very youngest children and ceiling effects for the oldest ones, resulting in a cubic response function in those states with a broad age distribution but not elsewhere. In Oklahoma, the distribution of children's ages was bimodal, and this may explain why cubic response functions were found for two outcomes there and somewhat less frequently elsewhere. States may also vary by the SES of children included in the program, with higher SES children yielding quadratic functions because of ceiling effects. We see little support for this in the data, but receipt of free or reduced price lunch is our only measure of SES. So we cannot be totally certain why response functions varied by state and outcome, though we are certain they did. To ignore this heterogeneity would bias the causal results achieved.

In subsequent analyses,[6] we address the second threat to validity in RDD: misallocation around the cutoff. While states aspired to error-free treatment assignment based on birthdates alone, there was a modest amount of misallocation in each state and hence a "fuzzy discontinuity" (Trochim, 1984). Experience suggests that when the number of misallocated cases is smaller than 5 percent, excluding misclassified participants makes little difference to the results (Judd & Kenny, 1981; Trochim, 1984; Shadish, Cook, & Campbell, 2002). Because only one state, West Virginia, exceeded 5 percent of misallocation, we present results that correct for misallocation in only one column (8) of Tables 4–8. The adjustment used there treats the cutoff as an instrumental variable (IV) for pre-K participation (see Hahn et al. [2001] for the proof of using IV in RDD to address misallocation; and Angrist and Lavy [1999] and Jacob and Lefgren [2004a, 2004b] for prior examples of such use).

Summary of analytic strategy

To deal with the functional form and misallocation assumptions, we present eight estimates of state pre-K effects for each outcome in each state. These estimates are in Tables 4 through 8. Column 1 presents the order of the polynomial that best models the relationship between the selection and outcome variables, given the descriptive analyses of functional form. In columns 2 through 4, we present parametric estimates that control for first-, second-, and third-order polynomials of the assignment variable. Of special interest here is the order that best fits the data in column 1, for it is the least likely to yield biased causal results. In column 5, we truncate the sample to six months on each side of the cutoff to reduce the role of outliers in determining the obtained functional form. In case the functional form assumptions

[6] To determine the sensitivity of causal estimates to treatment misallocation, we first calculated OLS effects for both the "full sample" of all children and a "restricted sample" purged of the misallocated cases, comparing the robustness of estimates using both samples. We then treated fuzzy discontinuity as a problem of omitted variable bias (Barnow, Cain, & Goldberger, 1980) and used an instrumental variable (IV) approach where children's true pre-K assignment was treated as an IV for their actual pre-K participation. The underlying assumption here was that all other effects of children's age on test scores were adequately controlled by covariates in the model. We checked the legitimacy of our instrument by presenting IV results using two specifications—one with and one without student covariates. Our analyses suggested that the relatively few instances of misallocation did not make much difference in estimates, that both approaches for handling misallocation yielded comparable results, and that the cutoff served as a valid IV instrument.

made in the parametric analyses are marginally flawed, columns 6 and 7 present nonparametric estimates for boundary groups at various bandwidths. Column 8 shows our preferred treatment estimates because they depend on the correctly modeled functional form and are also adjusted for the low levels of misallocation. We present both magnitude estimates and statistical significance patterns, though the latter are less informative since they depend on irrelevant state differences in sample size, on deliberately omitting cases in some analyses, and on whether parametric regression models include higher order terms or not.

Because policy makers are interested in the effect of a policy as implemented, we summarize the intent to treat (ITT) estimates in Table 9. These are OLS estimates based on full sample of cases, taking the best model of functional form into account from column 1 of Tables 4 through 8. The treatment on treated estimates (TOT) are summarized in Table 10. These are the IV adjusted estimates presented in column

Table 9. Intent to Treat Estimates of Outcomes by States.

	PPVT		Math		Print Awareness	
	Raw Score	Effect Size	Raw Score	Effect Size	Raw Score	Effect Size
	(1)	(2)	(3)	(4)	(5)	(6)
Michigan	−2.20	–0.13	2.07*	0.53*	25.21*	1.09*
New Jersey	6.29*	0.36*	0.89*	0.23*	8.46*	0.32*
Oklahoma	4.94*	0.28*	1.33	0.34	11.27	0.42
South Carolina	0.79	0.04			20.83*	0.78*
West Virginia	2.75	0.16	0.26	0.06	22.25*	0.92*
Unweighted average	2.51	0.14	1.14	0.29	17.61	0.70
Weighted average**	3.03	0.17	1.01	0.26	16.70	0.68

* significant at 5%.
** Weighted averages are calculated by weighting the number of enrolled state pre-K children by state. Effect sizes are calculated using sample standard deviations.

Table 10. Treatment on Treated Estimates of Outcomes by States.

	PPVT		Math		Print Awareness	
	Raw Score	Effect Size	Raw Score	Effect Size	Raw Score	Effect Size
	(1)	(2)	(3)	(4)	(5)	(6)
Michigan	−2.75	–0.16	1.82*	0.47*	22.14*	0.96*
New Jersey	6.10*	0.36*	0.87*	0.23*	13.02*	0.50*
Oklahoma	5.12*	0.29*	1.36	0.35	11.46	0.43
South Carolina	0.80	0.05			21.01*	0.79*
West Virginia	2.42	0.14	0.44	0.11	20.15*	0.83
Unweighted average	2.34	0.14	1.12	0.29	17.56	0.70
Weighted average**	2.80	0.16	0.99	0.26	16.95	0.68

* significant at 5%.
** Weighted averages are calculated by weighting the number of enrolled state pre-K children by state. Effect sizes are calculated using sample standard deviations.

8 of Tables 4 through 8. Given the low misallocation, the ITT and TOT estimates will be quite comparable. Both tables present results in the original metric and as standardized effect sizes. Effect sizes are calculated using standard deviation data from each state's comparison group and not from test developer publications using broader samples. While state differences in standard deviations could make it difficult to interpret state differences in effect sizes, Table 2 shows that there were no such variance differences.

RESULTS

Michigan

Table 4 presents results of the Michigan School Readiness program. Columns 2 through 7 show that linear models are appropriate for math and print awareness and that the estimates remain robust even when we overfit the regression model or truncate the sample to 6 months or use local linear regression at two different bandwidths. For PPVT, both graphical analysis and statistical analysis of higher order terms indicate that the response function is quadratic. However, regardless of the method used for estimation, all parametric and nonparametric estimates for PPVT are small and not significant. To summarize the Michigan effects is easy. PPVT scores were not affected, but math and print awareness scores rose because of pre-K. Students in the program scored about 1.82 points higher—or .47 standard deviation (SD) of the comparison group—on the Woodcock-Johnson Applied Problems subtest and answered 22.14 percent (.96 SD) more items correctly on the print awareness measure.

New Jersey Abbott program

Columns 2 through 7 of Table 5 examine the sensitivity of our New Jersey estimates. Because of the state's large sample size, we are able to use smaller bandwidths for the nonparametric estimates than elsewhere, thus weighting observations closer to the cutoff more heavily. For PPVT a linear form fits well, and the results are generally positive and consistent across all parametric and nonparametric models. For math, the estimate is .72 ($p < .05$) in the linear model, but .08 and .38 in the other two models, each nonsignificant. For print awareness, the graphical analysis and a reliable quadratic term in the regression analysis indicate clear evidence of nonlinearity. So we over-fit the model by including a cubic term in the parametric estimate. Positive and significant impacts were then observed, entailing reliable effects for all outcomes in New Jersey. For receptive vocabulary, scores were 6.10 raw points (.36 SD) higher at the cutoff; in math, scores were .87 raw points (.23 SD) higher; and for print awareness 13.02 percent (.50 SD) more items were answered correctly.

Oklahoma

Graphical, parametric, and nonparametric analyses provide strong evidence that the response function was linear for Oklahoma's PPVT outcome, and cubic for math and print awareness. For PPVT, the linear specification is obvious. For math and print awareness, the impact estimates in columns 2 through 4 of Table 6 decrease with the inclusion of higher order terms, implying that linear and quadratic specifications do not model the response functions well. The appropriateness of the cubic function is suggested through graphical analyses, the larger adjusted R-squares, the robustness of the estimates when the dataset is truncated to 6 months

each side of the cutoff (column 5), and the nonparametric estimates with the smallest optimal bandwidth (column 6). However, it is a concern that the density of cases inexplicably drops between 0 and 80 days after the cutoff relative to the density found in other areas of the age distribution, entailing fewer children than expected with birthdays just above the cutoff, the very place where they are most needed in RDD analysis. So we are less certain about the Oklahoma results than for other states. Positive impacts are indicated across the board, but they are only reliable for PPVT. On average, treatment children scored 5.12 raw points (.29 SD) higher than comparisons on the PPVT; 1.36 raw points (.35 SD) higher on the Woodcock-Johnson math assessment; and they obtained 11.46 percent (.43 SD) more print awareness items correct.

South Carolina

Due to a desire to limit testing time and costs, math measures were not administered in the first year of the South Carolina evaluation. Graphical, parametric, and nonparametric analyses consistently indicate that the assignment and outcome variables were linearly related. For PPVT, all the estimates were small and nonsignificant (columns 2–8 of Table 7). However, print awareness estimates were generally large and significant across all methods of estimation. So the program had little or no effect on children's receptive vocabulary, with treatment students scoring only .80 raw points (.05 SD) above comparisons. But it did have a reliable impact on print awareness, with treatment students answering 21.01 percent (.79 SD) more items correctly.

West Virginia

Table 8 describes the West Virginia results. Graphical, parametric, and nonparametric results provide evidence of linearity for PPVT and print awareness but not for math where a quadratic functional form was indicated instead. So for this one outcome we chose to include a quadratic term in our final parametric model. The print awareness estimate was comparable across all models, with the reliable estimates falling within 5 percentage points of each other. Treatment students correctly answered 20.15 percent (.83 SD) more items than controls. In contrast, the impact estimates for both math and receptive vocabulary, while positive, were small and nonsignificant—.44 (.11 SD) and 2.42 (.14 SD) respectively.

Summary of results across states

Tables 9 and 10 present estimates for each state in the raw score metric and as standardized effect sizes, both for the ITT and TOT analyses. Because misallocation was low, the ITT and TOT estimates hardly differ. Three things stand out. First, with the exception of PPVT in Michigan, all the coefficients are positive, illustrating the general effectiveness of these particular state pre-K programs. For PPVT, the mean ITT unweighted effect size is .14; for math it is .29, and for print awareness it is .70. Weighting each state by its sample of 4-year-olds yields estimates of .17 for PPVT, .26 for math, and .68 for print awareness. Second, the between-state variation in the size of effects seems large for each outcome, impelling one to ask whether a summary average effect size makes much sense in light of the state differences in effects. And finally, it is striking how different the effect sizes are across the three outcomes. They are very large for the print awareness measure, which is basically a test of knowledge of letters of the alphabet. They are quite modest for the more

general and vocabulary-based PPVT measure. And the math impact falls between the other two.

DISCUSSION

The main purpose of this study was to examine whether state pre-K programs have generally positive short-term effects on cognitive aspects of school readiness. The results clearly establish that state-level programs can have such effects on vocabulary, pre-reading, and early math skills even when (1) local programs are heterogeneous within a state; (2) program implementation is in the hands of routine managers and teachers rather than program developers; and (3) selection bias is ruled out.

The case for pre-K's general effectiveness rests on the consistency of results. Thirteen of the 14 causal coefficients were positive in direction, and 8 of them were statistically significant—far more than would be expected by chance. RDD is less powerful than an experiment, and so it is perhaps not surprising that all three effects were reliable in the state with most children (New Jersey) but reliability was less frequent in the states with smaller samples. Moreover, higher order functional forms require adding quadratic and cubic terms to models and thus increasing standard errors and reducing the chances of rejecting the null hypothesis. Thus, in Oklahoma, standard errors clearly increased as higher order terms had to be added to the model to respect the nonlinear functional forms found (Table 6). The result was that neither of these effects reached conventional levels of statistical significance, though both were positive. It is more important, therefore, to consider the direction of effects than statistical significance patterns, and the direction of effects was consistently positive.

However, effect sizes did vary by outcome. They were lowest for PPVT (.14 across states, unweighted from the ITT analyses), next highest for math (.29), and highest for print awareness (.70). The vocabulary-based PPVT measure is probably the most general in the range of cognitive skills tested, and print awareness is probably the most specific, tapping into just 26 lower-order and capital letters and the sounds associated with them. Prior studies have shown pre-K children to be particularly adept at learning alphabet-related concepts rather than the larger PPVT skill repertoire. This has now been the case for learning from *Sesame Street* (Cook et al., 1975; Minton, 1975), Head Start (Puma et al., 2005), and Early Reading First (Jackson et al., 2007). All three produced reliable effects for print awareness/letter recognition but not for the PPVT total. Are children in our culture particularly primed to learn alphabet-related skills between the ages of 3 and 5; or do larger effects tend to be achieved when the assessment is closely aligned to what is taught (Cook, 1974), and teaching letters is central to all preschool classrooms? We are not certain why effect sizes varied so much by outcome, and consistently so across states. But the pattern of outcome variation is consonant with past evaluative results on other pre-K programs. The state pre-K programs achieved effect sizes for alphabet skills that far surpass the minimal detectable effect levels currently specified in educational research funded by the Institute for Educational Sciences; the math results also surpass the usual .20 criterion; but the PPVT does not when aggregated across states, though it does in some individual states. Future analysis is required, therefore, of the differential weight that should be assigned to each of these outcomes, particularly in regards to their capacity to predict to the changes in early adulthood that make pre-K research so promising for public policy.

In considering this variation in effect sizes by outcome, it is also important to remember that the composition of control groups has probably changed in pre-K research compared to 40 years ago, when experiments first began showing reliable

effects with small samples. If we take the five states here and divide their sample sizes by 2.75—the approximate efficiency difference between an experiment and RDD—then the five sites are roughly equivalent to randomized experiments with sample sizes of 317 for Michigan, 753 for New Jersey, 305 for Oklahoma, 283 for South Carolina, and 262 for West Virginia. These are all larger than what was needed to show cognitive effects both when *Sesame Street* began (Minton, 1975) and in the famous Ypsilanti-Perry preschool study. But we might presume that, at that time, the control groups contained more children who had no center-based care options, thus creating a lower counterfactual hurdle, and the need for fewer cases, than would be the case today. In our New Jersey sample, more than half the children participated in pre-K when they were three (the comparison group). We did not collect comparable data in the other four states, but many were eligible to enroll in Head Start or private preschool programs (Yarosz & Barnett, 2001). If current pre-K programs are indeed required to surpass no-cause baseline criteria that are higher than in the past, this suggests how robustly effective state pre-K programs are.

Because results are consistently positive does not mean that they can be confidently extrapolated to the nation at large. To be able to do this requires either a census of all states or the random selection of states from among the 38 with pre-K programs. However, the five states studied here were selected opportunistically and they are among the best in the country in terms of pre-K quality standards (Barnett et al., 2005, 2006), obfuscating extrapolation to the nation at large. Nonetheless, it is possible to conclude that effective programs can be found across these states and across the range of variation found in them. As Table 1 indicates, this variation is considerable even if it is truncated at the lower end. Moreover, some of the states in our sample have not been historically noted for their commitment to education. Even so, positive results tended to be achieved.

Our second purpose was to document state variation in effect sizes. This variation is described in Tables 9 and 10. It seems impressionistically large and so partially undercuts the utility of computing a single average effect across all five states. The state variation in effect sizes speaks to a key issue for contemporary pre-K policy. Does the variation in outcomes depend on state program attributes that are usually considered to be central in a theory of quality preschool education? There is still no universally agreed on measure of state pre-K quality, or even an empirically corroborated theory of quality from which researchers can derive measures. Nonetheless, our estimates of total expenditures per child from all sources (federal, state, and local) can be used as one crude proxy for state program quality. When states are rank-ordered by their spending patterns, New Jersey is first, then West Virginia, Oklahoma, Michigan, and South Carolina (see Table 1). However, (1) New Jersey spends the most on pre-K per student and produces the largest effect size for PPVT but the smallest for print awareness; (2) West Virginia has the second highest funding but scored lowest in math and produced medium size effects for the other two outcomes; (3) Oklahoma ranked third, but yielded reliable results only for PPVT, though the point estimates for PPVT and math were the second largest of all; (4) Michigan ranked fourth and had the smallest PPVT effect size but the largest math and print awareness effect sizes; and, (5) South Carolina ranked lowest in funding and among the lowest in outcomes, a statistically significant result only for print awareness and not PPVT. This analysis is crude and confounded with dosage since states that spend less may offer fewer hours of pre-K services and also differ in the options available to the (younger) comparison children. But even so some things are clear: First, if there is a relationship between total state funding levels and the magnitude of results, it is not so strong as to stand out cleanly. Second, there is no plausible correction for state dosage differences that would cause state outcomes per unit of class time to closely mirror the state rank-ordering of

Table 11. Comparison of Average ITT Effect Size Estimates from State Pre-K Study (2007) and the Head Start Impact Study (2005)

	State pre-K Study (2007)	Head Start (2005)***
	Unweighted ES Average (1)	Nationally Representative (2)
PPVT	0.14	0.05
Math	0.29	0.10
Print Awareness	0.70	0.25

*** ITT estimates are from Puma et al. (2005).
Effect sizes are calculated using sample standard deviations.

Table 12. Comparison of Average TOT Effect Size Estimates From State Pre-K Study (2007) and the Head Start Impact Study (2005)

	State pre-K Study (2007)	Head Start (2005)***
	Unweighted ES Average (1)	Nationally Representative (2)
PPVT	0.14	0.08
Math	0.29	0.15
Print Awareness	0.70	0.36

*** TOT estimates are from Ludwig and Phillips (2007).
Effect sizes are calculated using sample standard deviations.

expenditure levels. States with half-day programs do not do noticeably worse than those with full-day programs. And third, we do not know how the states differed in the services available to comparison children. So with only five states, we cannot be totally certain whether state quality differences are correlated with state outcomes. Even so, advocates of quality pre-K services can get little solace from comparing results across different states, though they may derive more solace from the generally positive results in these five states that, after all, meet high quality standards. Unfortunately, we do not know what level of results would be found in low quality standard states.

The next issue is incidental to our main purposes. The state-level effect sizes we have produced can be averaged and then compared to the results from competing programs, of which Head Start is the perennial contender. To estimate the Head Start effects we turn to the national evaluation with random selection followed by random assignment (Puma et al., 2005). Table 11 provides the relevant ITT effect sizes. Since the Head Start Impact Study report does not present treatment on treated impact estimates for nonsignificant results, in Table 12 we report TOT estimates calculated by Ludwig and Phillips (2007) that are sensitive to treatment children not showing up for the Head Start program and to control children "crossing over" into the program.[7] For PPVT, the five states averaged TOT without weighting

[7] The procedure used by Ludwig and Phillips (2007, pg. 22) requires the following three assumptions are met: (1) that random assignment was successful and treatment group assignment had no effect on children who did not participate in the program; (2) that there were no defiers, or children who would not participate if assigned into the treatment and vice versa; and (3) that the average quality of Head Start programs attended by treatment and control children is comparable.

is .14 while it is .08 for Head Start. For math, the TOT average state estimate is .29 against .15 for Head Start. For print awareness, the unweighted average state effect size is .70 against .36 for Head Start. Thus, the states seem to outperform Head Start in all domains by a factor of about two. This same conclusion is also supported when we examine the ITT estimates where the Head Start effects appear even smaller but the state ones do not, creating effect sizes for the states that are three times those for Head Start.

Like others (Cook, 2006; Ludwig & Phillips, 2007), we do not much like comparisons of the above kind, though we acknowledge that some policy analysts will make them (Besharov & Higney, 2007b). One problem stems from confounds due to methodological differences in how each program is evaluated. The Head Start Study (Puma et al., 2005) is national, whereas the five states studied here are not nationally representative and have some of the highest quality standards in the nation. Programs that operated as both state pre-K and Head Start are another possible confound, and a clean contrast should omit Head Start centers that are cofunded from state sources, though in this study the percentage of such centers was never more than 10 percent in any state. Another difference clouding state and Head Start comparisons is the difference in the population served. Head Start's eligibility guidelines require that at least 90 percent of children served come from families at or below the poverty line, and at least 10 percent of the slots are reserved for children with disabilities whose families may have incomes above the poverty threshold. In contrast, Oklahoma offers universal access to services; West Virginia is expanding its program to serve all 4-year-olds; Michigan offers services to those whose income is up to 185 percent of the poverty rate; New Jersey's program serves children who reside in districts where 40 percent or more of its children received subsidized lunch in 2002; and South Carolina does not use poverty as a criterion but includes it as a possible risk factor. Comparison of results across states is confounded if Head Start families are on average worse off than state pre-K families. A difference also exists in the emphasis given to cognitive achievement gains. They are included among Head Start's goals and are becoming ever more central to that program. But Head Start emphasizes health and nutrition programming, parental education and involvement, and coordination with social services. Four of the five states in our sample set comprehensive standards for physical well-being and social and emotional development, but they varied in their provisions around vision, hearing, and health screenings; referrals to social service; meals and snacks; and parental education. While we know how well Head Start did in noncognitive areas—nearly all coefficients are positive but quite small and rarely reliable—we do not know how well the state programs did in these non-tested areas. As Cook (2006) pointed out, the sad truth is that a clean comparison of Head Start and state programs requires random assignment to each within the same study. But no such study currently exists except for that of Henry et al. (2003), which involves a single state and where selection bias is not obviously ruled out.

This project used RDD because of its acknowledged theoretical and empirical advantages in justifying unbiased causal inference. RDD is an important tool in public policy whenever resources are distributed by merit, need, first come first served or—as here—by date of birth. However, RDD is not as useful as an experiment. It is less statistically powerful. Its assumption about functional form is particularly stringent. In many situations, the local average treatment effect that RDD estimates is less general than the average treatment effect from an experiment. And we have not yet had as much experience in discovering and solving problems with RDD's implementation as we have had with understanding the implementation of experiments. So experiments are still the causal method of choice, with RDD being an acceptable causal alternative if done carefully. We tried here to model a careful

RDD analysis, particularly in order to deal with the considerable state variation in functional form since the amount of "fuzziness" was quite limited. Our analysis was complex and had to be so in order to convince readers that the method's assumptions were met. RDD is a tool that can and should be used more often—but only with sensitivity to its functional form requirements, its proclivity to fuzzy allocation around the cutoff score, its lower statistical power than an experiment, and its very local average treatment effect around the cutoff.

VIVIAN C. WONG is a doctoral student in Human Development and Social Policy, Northwestern University.

THOMAS D. COOK is the Joan and Sarepta Harrison Chair in Ethics and Justice and an IPR faculty fellow at the Institute for Policy Research, Northwestern University.

W. STEVEN BARNETT is Director of the National Institute for Early Education Research, Rutgers University.

KWANGHEE JUNG is Assistant Research Professor, National Institute for Early Education Research, Rutgers University.

REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. Evaluation Review, 22, 207–244.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics, 144, 533–576.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. The Future of Children, 5, 25–50.

Barnett, W. S. (2007). Surprising agreement on Head Start: Compli/ementing Currie and Bersharov. Journal of Policy Analysis and Management, 26, 685–686.

Barnett, W. S., Hustedt, J. T., Hawkinson, L. E., & Robin, K. B. (2006). The state of preschool: 2006 state preschool yearbook. New Brunswick, NJ: The National Institute for Early Education Research.

Barnett, W. S., Hustedt, J. T., Robin, K. B., & Schulman, K. L. (2004). The state of preschool: 2004 state preschool yearbook. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W. S., Hustedt, J. T., Robin, K. B., & Schulman, K. L. (2005). The state of preschool: 2005 state preschool yearbook. New Brunswick, NJ: The National Institute for Early Education Research.

Barnow, B. S., Cain, G. C., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. W. Stormsdorfer & G. Farkas (Eds.), Evaluation studies review annual (Vol. 5). Beverly Hills, CA: Sage Publications.

Besharov, D. J., & Higney, C. A. (2007a). Head Start: Mend it, don't expand it (yet). Journal of Policy Analysis and Management, 26, 678–681.

Besharov, D. J., & Higney, C. A. (2007b). Response to Barnett and Currie. Journal of Policy Analysis and Management, 26, 686–687.

Black, D., Galdo, J., & Smith, J. C. (2005). Evaluating the regression discontinuity design using experimental data [Electronic Version]. Working paper retrieved August 6, 2007, from http://www.personal.ceu.hu/departs/personal/Gabor_Kezdi/Program–Evaluation/Black–Galdo–Smith–2005–RegressionDiscontinuity.pdf.

Buddelmeyer, H., & Skoufias, E. (2003). An evaluation of the performance of regression discontinuity design on PROGRESA. Bonn, Germany: IZA.

Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. Child Development, 65, 684–698.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. Applied Developmental Science, 6, 42–57.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. Evaluation Review, 18, 141–152.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association, 83, 596–610.

Cook, T. D. (1974). The medical and tailored models of evaluation research. In J. G. Albert & M. Kamrass (Eds.), Social experiments and social program evaluation (pp. 28–37). Cambridge, MA: Ballinger.

Cook, T. D. (2006). What works in publicly funded pre-kindergarten education? Paper presented at the Institute for Policy Research Congressional Policy Briefing.

Cook, T. D. (in press). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics.

Cook, T. D., Appleton, H., Conner, R., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). "Sesame Street" revisited. New York: Russell Sage Foundation.

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

Cook, T. D., & Wong, V. C. (in press). Empirical tests of the validity of the regression discontinuity design. Annales d'Economie et de Statistique.

Currie, J. (1995). Does Head Start make a difference? American Economic Review, 85, 341–364.

Currie, J. (2001). Early childhood education programs. Journal of Economic Perspectives, 15, 213–238.

Currie, J. (2007). How should we interpret the evidence about Head Start? Journal of Policy Analysis and Management, 26, 681–684.

Dunn, L. M., & Dunn, L. M. (1997). Peabody Picture Vocabulary Test–Third Edition (PPVT-3). Circle Pines, MN: AGS Publishing.

Dunn, L. M., Padilla, Lugo, & Dunn, L. M. (1986). Test de Vocabulario en Imagenes Peabody (TVIP). Circle Pines, MN: AGS Publishing.

Early, D. M., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., Clifford, R., et al. (2005). Pre-kindergarten in eleven states: NCEDL's multi-state study of pre-kindergarten and study of state-wide early education programs (SWEEP). Chapel Hill, NC.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine, 15, 451–474.

Gilliam, W. S., & Zigler, E. F. (2001). A critical meta-analysis of all impact evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery and program evaluation. Early Childhood Research Quarterly, 15, 441–473.

Goldberger, A. S. (1972a). Selection bias in evaluating treatment effects: Some formal illustrations. Institute for Research on Poverty.

Goldberger, A. S. (1972b). Selection bias in evaluating treatment effects: The case of interaction. Institute for Research on Poverty.

Gormley, W. T. (2007). Early childhood care and education: Lessons and puzzles. Journal of Policy Analysis and Management, 26, 633–671.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. Developmental Psychology, 41, 872–884.

Gormley, W. T., & Phillips, D. (2005). The effects of universal pre-K in Oklahoma: Research highlights and policy implications. The Policy Studies Journal 33, 65–81.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69, 201–209.

Heckman, J., & Masterov, D. V. (2005). The productivity argument for investing in young children. University of Chicago.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics, 24, 411–482.

Henry, G. T., Henderson, L. W., Ponder, B. D., Gordon, C. S., Mashburn, A. J., & Rickman, D. K. (2003). Report on the findings from the early childhood study: 2001–2002. Atlanta, GA: Georgia State University.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., et al. (2007). National evaluation of Early Reading First: Final report. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Jacob, B., & Lefgren, L. (2004a). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. Journal of Human Resources, 39, 50–79.

Jacob, B., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity analysis. Review of Economics and Statistics, 86, 226–244.

Judd, C. M., & Kenny, D. A. (1981). Estimating the effects of social interventions. New York: Cambridge University Press.

Lonigan, C., Wagner, R., Torgeson, J., & Rashotte, C. (2002). Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPP): Department of Psychology, Florida State University.

Ludwig, J., & Miller, D. L. (2005). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. NBER Working Paper No. 11702.

Ludwig, J., & Phillips, D. (2007). The benefits and costs of Head Start. NBER Working Paper No. 12973.

McCarton, C. M., Brooks-Gunn, J., Wallace, I. F., Bauer, C. R., Bennett, F. C., Bernbaum, J. C., et al. (1997). Results at age 8 years of early intervention for low-birth-weight premature infants: The Infant Health and Development Program. Journal of the American Medical Association, 277, 126–132.

Minton, J. H. (1975). The impact of "Sesame Street" on reading readiness of kindergarten children. Sociology of Education, 48, 141–151.

Pagan, A., & Ullah, A. (1999). Nonparametric econometrics. Cambridge, UK: Cambridge University Press.

Phillips, D., Gormley, W. T., & Lowenstein, A. (2007). Classroom quality and time allocation in Tulsa's early childhood programs. Georgetown University.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). Head Start impact study: First year findings. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention. Evaluation Review, 19, 347–373.

Rock, D. A., & Stenner, J. A. (2005). Assessment issues in the testing of children at school entry. Future of Children, 15, 15–34.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). The High/Scope Perry Preschool study through age 40. Ypsilanti, MI: High/Scope Press.

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

Trochim, W. M. K. (1984). Research design for program evaluation. Beverly Hills, CA: Sage Publications.

Wagner, R., Torgeson, J., & Rashotte, C. (1999). Comprehensive Test of Phonological Processing (CTOPP). Austin, TX: Pro-Ed.

Weikart, D. P., Bond, J. T., & McNeil, J. T. (1978). The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade. Yipsilanti, MI: High/Scope Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson tests of achievement. Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Munoz, A. (1990). Bateria Woodcock-Munoz pruebas de aprovechamiento-revisados. Itasca, IL: Riverside Publishing.

Yarosz, D., & Barnett, W. S. (2001). Who reads to young children? Identifying predictors of family reading activities. Reading Psychology, 22, 67–81.