Running Head: STATE PRE-K PROGRAMS

An Effectiveness-based Evaluation of Five State Pre-Kindergarten Programs using Regression-Discontinuity

Vivian C. Wong and Thomas D. Cook

Northwestern University

W. Steven Barnett and Kwanghee Jung

National Institute for Early Education Research, Rutgers University

June 2007

Abstract

This paper evaluates how five state pre-kindergarten (pre-K) programs affected children's receptive vocabulary, math, and print awareness skills. Taking advantage of each state's strict enrollment policy 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 vary by state and outcome.

Keywords: pre-kindergarten, cognitive development, regression-discontinuity

Introduction

Many evaluations of pre-kindergarten (pre-K) programs have appeared since the first and flawed evaluation of Head Start (Cicarelli, Evans, & Schiller, 1969). These studies vary in causal methodology and include some randomized experiments. They have mostly shown positive effects of the interventions aimed at children up to age five, with some effects being very long term and observed in multiple domains of adult life – e.g. Weikart, Bond, and McNeil, 1978, Campbell and Ramey, 1995, McCarton et al., 1997, Reynolds, Temple, Robertson, and Mann, 2001, Magnusson, Myers, and Ruhm, 2004, Johnson and Blumenthal, 2004, Loeb, Bridges, Bassok, Fuller, and Rumberger, 2005, and Magnuson, Ruhm, & Waldfogel 2007. Summaries of the relevant literature are in Barnett (1995), Currie (2001), Heckman and Masterov (2005) and Loeb and Bassok (2007). The efficacy (Flay, 1986) of pre-K programs is not in question.

Less clear is the effectiveness of such programs when mounted as large government initiatives, whether at the federal or state levels. A recent national evaluation of Head Start was based on a sampling frame of centers serving 84.5 percent of all Head Start enrollees. A probability sample of these centers was then drawn followed by random assignment of children to these centers versus to a wide variety of alternatives (Puma et al., 2005). In both intent to treat and treatment on treated analyses (Imbens & Rubin, 1987), the study showed consistently positive one-year trends in the cognitive domain, but they were only intermittently statistically significant. Positive trends were also observed in social and behavioral domains, but these were even more rarely reliable (Puma et al., 2005). Three recent studies with more of an effectiveness focus exist at the state level, and we review them below. However, none chooses states or pre-K centers within states using selection with known probabilities, as was the case with the national Head Start evaluation; and none has a causal design as strong as the random assignment used for

Head Start. So these studies of state pre-K programs are bound to be less complete for inferring the general effectiveness of state-level programs as they are currently implemented.

Xiang & Schweinhart (2002) evaluated six half-day pre-K sites in Michigan, matching children on demographic attributes like age and socioeconomic background but not on pretest measures of the main outcomes. Thus, their design is one that Cook and Campbell (1979) considered to be “generally uninterpretable” causally. Nonetheless, their claim was that, five years later, 24 percent more pre-K children passed the state literacy test, 16 percent more passed the mathematics test, and 35 percent fewer were retained a grade.

The Georgia Early Childhood Study (Henry, Henderson, Ponder, Gordon, Mashburn, & Rickman, 2003) compared learning outcomes for probability samples from state pre-K, Head Start, and private preschool programs, thus permitting within-state generalization to these three types. Pretests were administered in the fall of the preschool year, making selection differences better identified even if not perfectly so. Also, teachers and parents were surveyed. Significant pretest differences were observed between the programs. On average, the Georgia Head Start children had the lowest cognitive scores and lived in the most disadvantaged households. Children enrolled in private preschools had the highest scores 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 claim was that, after a year of intervention, children in the state and private programs did not perform differently. However, the Head Start children performed less well on three of the five cognitive outcomes. The problem here is to know how well selection was accounted for so as to rule out the possibility of residual bias.

The third study was conducted in Tulsa, Oklahoma, and is technically superior to the others from an internal but not external validity viewpoint (Gormley, Gayer, Phillips, & Dawson, 2005; Gormley & Phillips, 2005). A regression-discontinuity design (RDD) was used because it produces unbiased causal estimates both in theory (Goldberger, 1972a; 1972b; Robbins & Zhang, 1988) and empirical practice (Aiken, West, Schwalm, Carroll & Hsiung, 1998; Buddelmeyer & Skoufias, 2003; Black, Galdo, & Smith, 2005; summarized in Cook & Wong, in press). To implement RDD, the authors 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 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 treatments. Achievement test scores for 1567 city children entering pre-K were then compared with scores from 1461 kindergarteners who had just completed pre-K. The analysts claimed 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.

A concern with Gormley et al. (2005) is the external generalization of results. They are limited to Tulsa, and so it is unclear whether services offered there are representative of the state overall. It is, after all, the largest and most urban school district in Oklahoma. Moreover, some evidence indicates that Tulsa's pre-K program is of exceptionally high quality relative to other preschool programs nationally. A recent study examined the level of instructional and emotional support in Tulsa pre-K classrooms and the amount of time spent on pre-literacy and math activities (Philips, Gormley, & Lowenstein, 2007), comparing these findings to results from a multi-state study of pre-K classrooms (Early et al., 2005) using similar measures. They found that Tulsa pre-K classrooms scored significantly higher on all four dimensions of instructional support, on one of four dimensions of emotional support, and spent much more time engaged in reading and literacy, math, and science activities than in the national sample of pre-K classrooms. Indeed the authors characterize Tulsa as a national example of high quality preschool programs, thus limiting the extrapolation of results from Tulsa to state pre-K programs more generally.

One reason for the interest in state pre-K programs is their recent expansion and the need to know what they are accomplishing. Since 1980, the number of states with programs has more than doubled and, by 2006, 38 states were serving near 1,000,000 children (Barnett, Hustedt, Hawkinson & Robin, 2006). The number of 4-year-olds in these programs has now come to surpass even the number enrolled in Head Start. It is unfortunately not possible yet to produce an unbiased estimate of what state programs are accomplishing. We lack a study with random selection of pre-K sites from a national pool followed by random assignment of children to these sites or a control status. Instead, we must rely on the available but purposive sample of states for which pre-K data are available. The five we use here are Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia. Fortunately, the plan in four of these states was to select state-funded pre-K classrooms at random while in the fifth—New Jersey—the plan was to select classrooms at random from within the largest state program serving 79 percent of the age-eligible children in the urban areas it was offered. In all five states, four children were then randomly selected for study within each targeted classroom, creating a sampling design that is formally representative of state pre-K attendees in four cases, and in the fifth case, New Jersey, the design is representative of districts where the state pre-K program was offered. Since each state employs birthdates for assignment to pre-K, this permits use of a regression-discontinuity design and thus unbiased causal estimates if the design is implemented properly. The net result is that, for any state implementing its intended sampling design and age-based RDD assignment process correctly, the causal estimates produced by the RDD analysis should be both unbiased and generalize to the state or state program at large.

Another reason to be interested in state pre-K programs is that they vary. All seek to prepare young children for kindergarten and school, and there is broad consensus on the cognitive, social, behavioral and mental health attributes of school readiness. But consensus is much lower about the priority each outcome domain deserves, as also is consensus about the priority preschool efforts deserve relative to other state goals in education or other sectors. As a result, state pre-kindergarten programs vary. Some provide for one or two years of education prior to kindergarten; some fund services from various mixes of state and local school district resources and federal monies via Title I, Individuals with Disabilities Education Act (IDEA), TANF, and even Head Start. Typically, funds are administered through state departments of education, sometimes with or without cooperation from human services departments. Although all states require pre-K programs to meet higher standards of quality than for childcare centers, the standards nonetheless differ by state. The result is that states vary in the level of funding per child, the mix of services supported, and other aspects of the quality of programs offered. This variation makes it important to estimate how much states differ in their effects on preschool children, and the degree of variation that results will indicate how difficult it may be to extrapolate from five states in order to obtain meaningful estimates of program effectiveness at the national level.

Even more important than describing between-state variation in effect sizes is estimating the extent to which this variation depends on the quality of the services offered. The NICHD-funded Study of Early Child Care (Love et al., 2003) defined quality by looking at the structural attributes of care centers. The study found that safer, cleaner and more stimulating centers that had more favorable child-staff ratios also employed staff who were more sensitive to children's needs and provided more cognitively stimulating care. Children who attended these better child care centers also exhibited higher scores on cognitive and language development measures. Similar findings have been reported in studies with samples of minority and/or low-income families (Burchinal et al., 2000; Loeb, Fuller, Kagan, & Carrol, 2004), including when socio-emotional adjustment is the dependent variable rather than cognitive achievement (Votruba-Drzal, Coley, & Chase-Lansdale, 2004).

Another measure of quality examines states' policies for promoting quality pre-K programs, irrespective of how well these policies are implemented on the ground. Each year, the National Institute for Early Education Research (NIEER) assesses how well states meet criteria believed to promote quality early education programs. These criteria depend on states' requirements for teacher education and training; on their minimum staff-child ratios and classroom sizes; on having comprehensive early learning standards that cover socio-emotional, physical, and intellectual development; on states' provisions for meals, vision, hearing and health screenings; on requirements about teacher-parent conferences; on referrals to external social services; and on state monitoring of pre-K programs through site visits. In 2004—the present study year--all five states in this evaluation paid teachers on a public school scale; and they all required programs to employ teachers with four-year college degrees, though West Virginia did allow some teachers with only associate's degrees. Four of the five state programs were mature, established between 15 and 20 years ago. New Jersey was the exception. Its program was created in 1998 and its standards were substantially raised in 2002. So these five state programs rank "above average" on class size, staff-child ratios, teacher qualification, and compensation. Even so, they do vary in length of day, funding level, and eligibility requirements. While it is impossible to expect definitive statements about the association between state standards and child gains from a study with five states, we can at least explore how the two co-vary in hopes of inspiring further research on the topic.

The final purpose for state estimates of pre-K effects entails comparing the effectiveness of state and federal efforts to raise children's school readiness—the central purpose of Henry et al. (2003). Government officials and advocates have various motives for comparing effect sizes from Head Start with those from state pre-K programs. If states with higher quality standards outperform Head Start, policy commentators may argue for federal quality standards to be raised to the level found in these states (NIEER, 2005). However, comparisons of program effects may also be used to support efforts to increase states' authority over the $9 billion in Head Start funds, and in the extreme, block grant the federal program. Over the last decade, initiatives to block grant Head Start have been introduced by the Bush administration as well as in the House of Representatives.[1] 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. If the effect sizes from state pre-K evaluations seem larger than those from high quality Head Start evaluations, this would seem to support the congressman's goal. So this study will compare effect sizes from these five states and from Head Start, albeit in a context that emphasizes how difficult such comparisons are unless they have been deliberately and fairly built into a single experiment.

The present study has four purposes, then: (1) to calculate an average impact estimate across five states; (2) to describe the between-state variation in impact; (3) to identify clues that

[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.

might help explain this variation; and (4) to compare effect size estimates from these five state programs and Head Start, a federal pre-K program. To these ends the present study includes close approximations to probability samples of preschoolers from five states, albeit states that were themselves purposively selected. For study outcomes we use three academic achievement measures -- Peabody Picture Vocabulary Test (PPVT) scores, print awareness scores, and math scores. Academic achievement is not the sole rationale for pre-K programs, of course. However, many past reviews of pre-K effects have emphasized achievement because of its greater availability in the research record and its links to the widely shared policy goal of increasing human capital (Barnett 1995; Currie, 2001; Heckman, Stixrud, & Urzua, 2006). Even so, social, emotional, and physical development also contribute to human capital and welfare so that an exclusive focus on achievement is bound to describe only part of how Americans want young children to develop.

Methods

The Sampling Design

The sampling design has three levels: states, classrooms within states, and children within classrooms. Five states were purposively selected; for four of them we used a simple random sampling of classrooms and then randomly selected four students per class. For the fifth state, New Jersey, a stratified random sample was selected within the state's largest pre-K program. However, compliance was not perfect, with some districts, schools, and classrooms refusing to participate. Notable examples were in Michigan where the Detroit school district granted permission too late in the year to be included; and in West Virginia where 41 percent of those 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 pre-K programs within a given state, even though they are more heterogeneous than in most prior pre-K studies.

The Michigan School Readiness Program (MSRP). Targeting only at-risk four-year-olds, MSRP enrolled 24,729 children, or 19 percent of four-year-olds in the state. At each site, half or more of the children had to meet either an income eligibility criterion and have one other risk factor from a list of 25; or else they had to exhibit more than one of the 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. The program requires baccalaureate teachers, has a staff-child ratio under 1:8, and no more than 18 children per class. There was no comprehensive curriculum requirement and, 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 et al., 2005).[2] From K-12 spending patterns in Michigan, we estimate that *total* expenditure per child was approximately $5,000.

To obtain the Michigan sample, state-funded 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. Four children were then randomly selected within each pre-K or kindergarten class, making for 485 treatment and 386 comparison children. Nearly half qualified for free or reduced price lunch, a third were African American or Hispanic, and 53 percent were White or Asian (Table 2).

[2] 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.

New Jersey's Abbott Preschool Program. As a result of 1998 state Supreme Court ruling, the New Jersey Abbott Program provides voluntary preschool for three- and four-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, but is by far the largest and best funded. The state Supreme Court ruling resulted in the implementation of much higher standards in all programs beginning in 2002. These include: a maximum class size of 15, requirements for teachers to have a bachelor's degree and specialized training in early childhood education, and the provision of coaches to help teachers improve their classroom practice. The state has two other pre-K programs serving children in less disadvantaged communities, but these had lower standards and funding levels. The Abbott program served 19 percent of the state's four-year-olds while the other two pre-K programs served 7 percent. Our results apply only to the Abbott program.

About 21,286 four-year-old children were enrolled in the program. The program also served 17,397 (about 15 percent) three-year-olds. In addition to the state Department of Education funding pre-K programs for the 6 hour school day and 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.

To select classrooms, a random sample of 21 Abbott districts was selected after stratification on factors like district enrollment, geographic location, urban versus rural setting, and the percentage of bilingual students. Within these districts, pre-K classrooms were selected from an enumerated list of all Abbott-funded pre-K classrooms, and then an equal number of kindergarten classrooms were selected from within the same districts. Again, four children were randomly selected per classroom. The New Jersey sample is the largest of all five states, including 1,177 treatment children and 895 comparisons. Sixty-eight percent of the sample qualified for free or reduced price lunch, with 25 percent of the students being African American, 39 percent Hispanic, and 14 percent White or Asian (Table 2).

Oklahoma's Early Childhood Four-Year-Old Program. In 1980, Oklahoma began providing pre-K services for four-year-olds on a pilot basis. Ten years later, the program was broadened to include all four-year-olds eligible for Head Start. But in 1998, Oklahoma became only the second state to offer free voluntary preschool to all four-year-olds.[3] Enrollment increased steadily over the last decade and since 2002, Oklahoma enrolled a greater percentage of its four-year-olds than any other state. In the 2004-2005 school year, 30,180 four-year-olds were enrolled in the state preschool program, or 65 percent of state four-year-olds. State pre-K was not offered to any three-year-olds. Most children were served in public schools, though districts could also collaborate with private childcare or Head Start centers to provide services. Regardless of setting, all pre-K teachers were required to have a bachelor's degree and a certificate in early childhood learning. Open throughout the academic year, local centers could determine whether to offer half or full day services. Oklahoma had comprehensive curriculum standards and limited the staff-child ratio to 1:10, with a maximum class size of 20 (Barnett et al., 2005). In 2004-2005 school year, the state spent over $100 million on preschool education, approximately $3,500 per child, though the state school formula relies on local schools' support

[3] Georgia was the first state to enact legislation that offered voluntary universal pre-K to four-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 what monies could be made available through the state lottery system while Oklahoma funded any four-year-old that school districts could enroll. Thus, from 2004 to 2006, Georgia enrollment rates of four-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.

for a portion of their funding. Expenditure per child from all sources was estimated to exceed $6,100 per child (Barnett et al., 2005; Barnett, Hustedt, Hawkinson, & Robin, 2006).

The classroom sampling procedure was the same as in Michigan, with a random selection of state-funded pre-K classrooms and then of kindergarten classrooms within the same districts. Four students from each classroom were then chosen at random. In all, 431 children were included in the treatment group and 407 in the controls. Almost 5 percent of the sample children came from Tulsa, and the remainder from 51 other districts across the state. About half the sample received free or reduced price lunch, and most students were White (65 percent), though Native American students were 13 percent of the sample and African American and Hispanic students were each about 7 percent (Table 2).

<u>South Carolina’s Early Childhood Programs.</u> South Carolina’s state preschool initiative is comprised of 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 preschool classes or serving additional children in half-day classes. In the 2004-2005 school year, 17,821 of four-year-olds were in enrolled in the state pre-K program, or 32 percent of all four-year-olds. Although eligibility for the state pre-K program was determined at the district level, it was based on a list of risk factors identified by the state. Poverty was 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 preschool. South Carolina required that teachers have at least a bachelor’s degree and certification in early childhood education. The staff-child ratio was 1:10 with a maximum class

size of 20, and all curricula models used must be research-based. 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).

To select pre-K and kindergarten classrooms, the same sampling procedure was used as in Michigan and Oklahoma. The South Carolina sample included 353 treatment children and 424 comparison children. About 54 percent of the sample received free or reduced price lunch, and 44 percent were African American and 40 percent were White (Table 2).

<u>West Virginia Early Childhood Education Program</u>. The West Virginia state pre-K program began in 1983 when a revision in the school board code allowed local districts to create preschool programs. Currently, the state is in the process of expanding access with the goal of providing voluntary universal pre-K to all four-year-olds. In the 2004-2005 school year, 6,541 of four-year-olds were enrolled in state pre-K, or 33 percent of all four-year-olds. The state also served another 4 percent of three-year-olds. Eligibility for four-year-olds was determined at the local level, with some counties enrolling students on a first come/first serve basis or by lottery. Children were served in a variety of settings, including public schools, Head Start centers, and child care and private preschool centers. Preschool programs lasted for the academic year, but the hours of operation varied by site. Typical programs operated for nine months a year, two full days per week, or four full days with Fridays reserved for home visits and planning. The state had a comprehensive curriculum requirement, a staff-child ratio of 1:10, and limited class sizes to 20 students. Teachers were required to have either bachelors' or associates' degrees, and most teachers had to have training in early childhood development (Barnett et al., 2005). In the 2004-

2005 school year, West Virginia spent $34.5 million on state preschool education, or $4,323 per child, with total funding from all state and local sources amounting to at least $6,829 per child enrolled (Barnett et al., 2005). West Virginia classrooms were selected according to the same sampling procedure as in Michigan, Oklahoma, and South Carolina. The sample included 379 treatment children and 341 comparisons. Thirty-three percent of students in our sample qualified for free or reduced price lunch, and 89 percent were White (Table 2).

Looking across all five states, we see that they differ in some ways likely to affect achievement—whether they are limited to four-year-olds or not, are whole day, half day or mixed, how many days per year they are open, and whether teachers can have only an associate level degree. States also differ in more methodological features that can affect conclusions about their pre-K program. For instance, state variation in policies and practices is confounded with noncompliance with the request for random selection and with state variation in state sample sizes -- from 2,072 students in New Jersey to 720 students in West Virginia.

Data Collection Procedure

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

Children were tested in the 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 preschool 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, Lugo, Padilla, & 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). The PPVT-III and TVIP both have a mean standard score of 100 and a standard deviation of 15.

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 ruebas 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. The achievement subtests have been standardized with a mean of 100 and a standard deviation of 15.

Print Awareness abilities were measured using the print awareness subtest of the Preschool Comprehensive Test of Phonological & 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, 19999), 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 answer 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 Points

The basic function for identifying treatment effects of state pre-K programs is:

$$Y_{ij} = BX_{ij} + \beta_1(\text{Pre-K})_{ij} + \beta_2 Z_{ij} + \varepsilon_{ij} \quad [1]$$

Where Y is the test score of interest, X is a vector of student covariates, Pre-K is a binary variable that takes on the value of 1 if the student participated in pre-K and 0 if not, Z are unobserved factors that are correlated with children’s learning outcomes, ε is the error term, i is the individual subscript, and j is the teacher subscript. This analysis takes advantage of pre-K deterministic enrollment in each state that is supposed to depend only on a child’s birth date. Children with birth dates after the state cutoff can enroll but those before it are required to wait another year.

To check the adequacy of this process in real-world state applications, Figure 1 shows that the percentage of children enrolled in preschool increased precipitously at the cutoffs for all five states in 2004-2005. 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 these so-called fuzzy cases were relatively few (Trochim, 1984). Table 2 shows no single state achieved even 9 percent misallocation.

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 of the slope changes at the cutoff. Two internal validity threats have to be dealt with in this conceptualization – incorrect specification of the functional form of the regression line, and treatment misallocation near the cutoff. Figure 2 shows what happens when the non-parametric plot (2) indicates that the data should be modeled using a cubic function, but it is instead fitted with a linear one (1). When this is done, a significant effect is detected at the cutoff with the linear function, but it is a spurious product of fitting the wrong functional form. So the data analysis will deal in detail with the sources of evidence indicating we have correctly specified the functional form as well as with procedures for ensuring we have adequately dealt with the (modest amount of) treatment misallocation around the cutoff.

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

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 RDDs varies with other factors not in Goldberger's work, but in no circumstance has the design been shown to be as efficient as a randomized experiment (Shadish, Cook, & Campbell, 2002).

Data Analysis: Specifics

Our RDD analysis begins with efforts to model the functional form of the assignment and outcome variables. We then examine the sensitivity of our estimates to misallocated cases. A variety of analytic techniques are used for each purpose, and we describe their benefits and weaknesses below.

Assumption 1: Adequate specification of the functional form. To identify the proper functional form, the analytic plan has three components: a graphical analysis, a series of parametric regressions with alternate specifications, and non-parametric procedures using local linear kernel regression (Hahn, Todd, & van der Klaauw, 2001).

To gain an indication of 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 3, two types of lines are fitted onto the scatterplots each side of the cutoffs. Plot (1) depicts a linear regression line, and plot (2) shows a non-parametric regression line based on locally weighted scatterplot smoothing, called lowess, that is often used for data exploration because it 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 3 depicts linear regression and lowess plots for New Jersey's PPVT. The purpose is to ascertain their comparability and, in this case, we observe that the parametric model is a close approximation to the lowess. This suggests that a linear model may be an appropriate specification for PPVT in New Jersey. If we had observed evidence of non-linearity 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:

$$Y_{ij} = a + BX_{ij} + \beta_1(\text{Pre-K})_{ij} + g(AV)_{ij} + \varepsilon_i \qquad [2]$$

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 5 through 9 will provide impact estimates using linear, quadratic, and cubic models. 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 mis-specifying functional form. So we rerun the parametric analyses including only those children with birthdates within six months each side of the cutoff.[4] 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 mis-specified functional form is to conduct non-parametric analyses. For these, we use simple differences of smoothed versions of the kernel estimator generated by local linear regression (Hahn, Todd, & van der Klaauw, 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 non-parametric estimates depend on proper specification of the interval, or bandwidth, within which local regressions are carried out. The narrower these bandwidths, the less biased are the estimates they yield. But they 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 only estimates for the two bandwidth choices that appear to best balance the bias-efficiency tradeoff.

[4] 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.

Our non-parametric 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 non-parametric 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). We consider the non-parametric estimates as additional sensitivity tests for probing the functional form assumptions we are forced to make and on which the validity of RDD results depends.

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. 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 would expect response functions to differ, given various ways in which states can differ. For instance, if states varied in the distribution of children's ages, then floor effects might be evident for achievement when children are very young and ceiling effects when they are older, resulting in a cubic response function for any state with a large age distribution. In Oklahoma, the distribution of children's ages was bimodal (see Figure 8), possibly explaining why cubic response functions were found for two outcomes there. States may also vary by the SES of children included in the program, with higher SES children yielding quadratic functions because of ceiling effects. While we see little support for this hypothesis in the data, it must be admitted that our only measure of children's socio-economic status, free or reduced price lunch receipt, is fairly imprecise and has quite a bit of missing data. Finally, states may vary in the reliability of outcome measures, but since the same assessments were used in all five states and attempts were made to administer the tests consistently, this may not be a major concern. The truth is that we cannot be sure why response functions varied by state and outcome. All we know for sure is that graphical, parametric, and non-parametric evidence points to heterogeneous response functions, and to ignore this heterogeneity would bias the causal results achieved wherever the functional form is mis-specified.

Assumption 2: Adherence to the cutoff. While states aspired to error-free treatment assignment based on birth dates alone, there was some misallocation in each state and hence a "fuzzy discontinuity" (Trochim, 1984). South Carolina and Michigan had the fewest fuzzy cases (1 percent and 2 percent, respectively), West Virginia had the most (8 percent), and Oklahoma and New Jersey were intermediate at 4 percent each.

To determine the sensitivity of causal estimates to this degree of treatment misallocation we calculated OLS effects for both the "full sample" of all children and a "restricted sample" purged of the misallocated cases. The order of the polynomial used in these analyses is in Table 3 and was determined by the functional form tests described earlier. When fuzzy cases are fewer than 5 percent — as in all but one state — experience is that excluding the misclassified participants makes little difference (Judd & Kenny, 1981; Trochim, 1984; Shadish, Cook, & Campbell, 2002). In any event, the full sample results we present constitute an intent to treat analysis (ITT), and the restricted sample results a treatment on treated analysis (TOT).

Our second approach to fuzzy discontinuity treats it as a problem of omitted variable bias (Barnow, Cain, & Goldberger, 1978) and requires identifying an instrumental variable (IV) that is correlated with treatment assignment but not with errors in the outcome. In practice, it is difficult to find cases where this assumption clearly holds except when the IV is either random assignment (Angrist, Imbens & Rubin, 1996) or one side of the cutoff versus the other in RDD (see Hahn et al., 2001 for the theoretical justification and van der Klaauw,2002; Angrist and Lavy,1999; and Jacob and Lefgren, 2004a, 2004b for applications). So states' enrollment rules allow us to treat students' true assignment into pre-K as an "instrument" for their actual participation, allowing us to estimate the following first stage equation:

$$\text{Pre-K}=BX_{ij} + \gamma_1(\text{TrueAssignment})_{ij} + g(AV)_{ij} + \eta_{ij} \quad [3]$$

Where Y, X, Pre-K, and g(AV) are the same for student i and classroom j as in Equation 1, and TrueAssignment is a dichotomous variable for the treatment condition that student i should be assigned to based on his/her birthday and the state's assignment rule. The second stage equation is identified in Equation 2.

The underlying assumption for using pre-K assignment as an instrument is that all other effects of children's age on test scores are adequately controlled by the covariates in the two stage least squares model. To probe this assumption, we consider how children's ages might be related to test score performance other than through admittance into pre-K. For instance, do older children have more opportunities for out-of-school learning, or do younger children receive more attention from parents and siblings at home? We check the legitimacy of our instrument by presenting IV results using two specifications – one with and one without student covariates. As we will see later, comparable estimates result for IV models with and without these covariates, suggesting that the covariates add nothing over and above what is added by knowing whether a child's birth date makes her or him eligible for pre-K access.

Summary of analytic strategy. To deal with the crucial functional form and misallocation assumptions we present 10 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, of course, the order that best fits the data in Column (1). 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 columns (6) and (7) we present non-parametric estimates for boundary groups at various bandwidths. This is in case the functional form assumptions made in the parametric analyses are marginally flawed. Taking the best model of functional form into account from column (1), Column (8) then presents regression impact estimates for the full sample including the fuzzy cases. Column (9) provides results from the same model but without these same cases. Column (10) presents IV estimates without student covariates in the model, while column (11) controls for student ethnicity, free-lunch status, gender, and whether the child took assessments in Spanish or English. We interpret the IV estimates presented in column (11) as our best single summary estimate. This is because they take advantage of information from the full sample; they respect the best assessment of functional form for a given outcome in a given state; and the IV analysis controls for the relatively few misallocated cases. Technically, column (11) is a TOT estimate, but given the low misallocation it should not differ much from ITT estimates without any IV adjustment. 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, on sometimes using IV instead of OLS, and on whether parametric regression models include higher order terms or not. In Tables 9 and 10, we summarize the findings by listing the preferred ITT and TOT estimates both in the original metric and as standardized effect sizes. The latter are calculated using standard deviation data from each state’s comparison group and not from test developer publications using broader samples. State differences in standard deviations could make it difficult to interpret state differences in effect sizes, but 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 non-parametric estimates for PPVT are small and not significant. Since only 2 percent of Michigan’s students were misallocated, columns (8) through (11) are nearly identical and reveal no influence of misallocation. 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 on the Woodcock-Johnson Applied Problems subtest and answered 22.14 percent more items correctly on the print awareness measure.

New Jersey. 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 non-parametric 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 non-parametric models. For math, the estimate is .72 ($p<.05$) in the linear model, but .08 and .38 in the other two models, each non-significant. Although the quadratic model has a slightly larger adjusted R-square (.351 versus .350 for each of the others), graphical analyses and the lack of significant higher order terms in the regression analyses suggest that the response function is best modeled as linear. For print awareness, there is clear evidence of non-linearity in Figure 4 and in the reliable quadratic term in the analysis. So we over-fit the model by including a cubic term in the parametric estimate. Columns (8) through (11) show that the 4 percent misallocated cases were not a problem. It seems, then, that New Jersey resulted in positive and significant impacts on children's receptive vocabulary, math, and print awareness skills. For receptive vocabulary, scores were 6.10 raw points higher at the cutoff; in math, scores were .87 raw points higher; and for print awareness 13.02 percent more items were answered correctly.

Oklahoma. Graphical, parametric and non-parametric analyses provide strong evidence that the response function was linear for Oklahoma's PPVT outcome, and cubic for math and print awareness. Figures 5-7 plot Oklahoma's assignment variable against the three outcome variables as a linear model, a non-parametric lowess line, our “best fit” parametric regression model, and as a local linear regression line. For PPVT, we choose the linear specification because of evidence from graphical plots (Figure 5), lack of statistical significant in the higher order terms, and a higher adjusted R-square value. 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 (Figures 6 and 7), 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 non-parametric estimates with the smallest optimal bandwidth (column (6)). Four percent of the sample could be identified as fuzzy, and columns (9) through (11) show that estimates are generally robust to variations designed to probe misallocation effects.

One issue with Oklahoma’s estimates is that the PPVT results are somewhat sensitive to specification and sample choices, and so we view these estimates with more uncertainty than the PPVT results from elsewhere. Another concern is 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 (see Figure 8). So there are fewer children than expected with birthdays just above the cutoff, the very place where they are most needed in RDD analysis. The most reasonable estimates assume a linear model for PPVT and cubic models for math and print awareness, and given these specifications, positive trends are indicated across the board but they are only reliable for PPVT. On average, treatment children scored 5.12 raw points higher than controls on the PPVT; 1.36 raw points higher on the Woodcock-Johnson math assessment; and they obtained 11.46 percent 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 non-parametric analyses consistently indicate that the assignment and outcome variables were linearly related. For PPVT, all the estimates were small and non-significant (columns (2)–(7) of Table 7); for print awareness, estimates were generally large and significant across all methods of estimation; and the controls for misallocation suggest that it again had no real effect. So the South Carolina program had little or no effect on children's receptive vocabulary, with treatment students scoring only .80 raw points above controls (see column (11) of Table 7). But the program did have a reliable impact on print awareness, with treatment students answering 21.01 percent more items correctly.

West Virginia. Table 8 describes the West Virginia results. Graphical, parametric, and non-parametric results provide evidence of linearity for PPVT and print awareness but not for math where the graphical and regression analyses indicate a quadratic functional form. 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 (columns (2) through (7)), with the reliable estimates falling within five percentage points of each other. With 8 percent misallocated cases, West Virginia had the largest number of fuzzy cases, but this still made little difference to the results—see columns (8) and (9). So the impact estimates for both math and receptive vocabulary were positive, but small and non-significant — .44 and 2.42 respectively — while a positive and significant effect emerged for print awareness where treatment students correctly answered 20.15 percent more items.

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 population of four-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 that 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 results clearly establish that state-level programs can have positive short-term effects on cognitive development even when (1) local programs are heterogeneous within a state; (2) there is no hovering program developer, and (3) residual selection threats are ruled out that might be due to studying local programs quasi-experimentally or to analyzing non-experimental survey data in a fashion deemed causal. The case for pre-K being generally effective rests on consistent results across these five states where 13 of the 14 causal coefficients were positive and eight of them statistically significant—far more than would be expected by chance. We prefer to emphasize the direction of effects than statistical significance levels, given that RDD is less statistically powerful than an experiment. All three effects were reliable in the state with the largest number of sampled children (New Jersey), and reliability was less frequent in states with smaller samples. Also, higher order functional forms require adding quadratic and cubic terms to models, thus also increasing standard errors. Indeed, Table 6 clearly shows how standard errors increased for Oklahoma's math and print awareness estimates as higher order terms were added; and neither of these effects reached conventional levels of statistical significance.

We should also not forget that the composition of control groups has changed in pre-K research compared to earlier days when experiments or strong quasi-experiments with small samples were able to show reliable effects. If we take the five states here and divide their sample sizes by 2.75, 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 was needed to show cognitive effects when Sesame Street began (Minton, 1976) and in the famous Ypsilanti-Perry preschool study. Yet 40 years ago, the control groups had more children without any alternative center-based care, creating a lower counterfactual hurdle than we find today and hence the need for fewer cases. Who knows how many of the control children in these five states were attending some kind of center-based care when they were two or three? So the case for state pre-K programs being generally effective rests on the striking consistency in the direction of effects more than on the less strong (but still respectable) pattern of statistically significant results.

The main factor limiting a conclusion about *general* state-level effectiveness is that the five states in this study are among the best in the country in terms of pre-K quality standards. At least this is the conclusion suggested by an analysis of their policies in terms of quality attributes that seem plausible. It is not clear how well these programs are implemented on the ground, but they are definitely among the better conceptualized and staffed in the country. As encouraging as these results are, it is difficult extrapolating from them to the nation at large. But what is not difficult to conclude is that effective programs can be found across the range of variation found in these particular states and, as Table 1 indicates, this is itself considerable even if truncated at the lower end.

The between-state variation in effect sizes described in Tables 9 and 10 requires some explanation. A key issue for contemporary preschool policy is how much the true between-state variation depends on the quality of state programs. However, we do not have direct measures of quality from an empirically corroborated theory of quality; and the state standards that we do have set limits but may not index quality as children directly experience it. Our estimates of total expenditures per child from all sources (federal, state, and local) could be used as a very crude proxy for quality. This would lead to the following rank order of total spending per state-- New Jersey first, then West Virginia, Oklahoma, Michigan, and South Carolina in that order (see Table 1). We then note that: (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 -- it achieved a statistically significant result only for print awareness and not PPVT. This analysis is crude in some ways, but it is very clear that no strong relationship holds between state funding levels and the magnitude of results. Of course, the populations served and options available to “control” children varied considerably across the states as well, making variations in effect sizes across states difficult to interpret.

Effect sizes also varied across the three outcomes examined, being lowest for PPVT (.14 across states, unweighted from the ITT analyses), next highest for math (.29) and highest for print awareness (.70). Of these tests, the vocabulary-based PPVT is the most general in the cognitive skills tested, while print awareness is probably the most specific, tapping into letter recognition, associating sounds with letters, and distinguishing print from pictures. Prior studies have shown pre-K children to be particularly open to learning alphabet-related concepts rather than the larger PPVT skill repertoire (Cook, 1975; Minton, 1975 for learning from Sesame Street). Did these state pre-K programs teach best the specific set of alphabet-related skills to which children between 3 and 5 are particularly primed in our culture, while achieving less across a broader range of early cognitive skills? An alternative explanation is that larger effects tend to be achieved when the assessment is closely matched to what is taught (Cook, 1974), and teaching letters and symbols is a core component of all preschool classrooms. Were effects for print awareness larger because the relevant skills were taught more often and more explicitly in pre-K classes than were vocabulary and math skills? Finally, with only 36 items on the print awareness measure and 204 items on the PPVT, the difference in effect sizes may reflect how relatively easy it is to obtain large differences when the assessment contains few items measuring a narrow domain as opposed to more items measuring a larger domain. We are not certain why the effect sizes varied so much, but they clearly varied considerably in ways that have been demonstrated in earlier work on Sesame Street.

The final issue we address is contentious in the current policy context—how large are the effects of these state programs relative to results from other recent studies of preschool programs? The two most currently discussed comparisons are with Head Start and pre-K services in Tulsa, for the claim has been made that the Tulsa estimates are of an especially high quality program and so indicate what pre-K programs are capable of (Gormley et al., 2005). For the Tulsa estimates we rely on Gormley et al. (2005), and for the Head Start effects we turn to the national evaluation with random selection followed by random assignment (Puma et al., 2005). Since the Head Start Impact Study report does not present treatment on treated impact estimates for non-significant results, we use TOT estimates calculated by Ludwig & Phillips (2007). The authors report that their treatment on treated estimates address selection issues from treatment children not showing up for the Head Start program, and from control children "crossing over" into participation of program services.[5] Table 11 provides the relevant TOT effect sizes from overlapping tests. On Table 12, we include ITT effect sizes for the Head Start Impact Study and the state pre-K evaluation, but not for the Tulsa study because ITT estimates are not available.

Let us begin with comparing the treatment on treated PPVT results from Head Start and the five states. For the five states averaged without weighting, the TOT effect size is .14 and for Head Start it is .08. For math, the TOT estimate is .29 across this sample of states against .15 for Head Start, and for print awareness the unweighted average state effect size is .70 against .36 for Head Start. The states seem to outperform Head Start on all three outcome domains. The pattern of results is similar for the intent to treat estimates where the Head Start effects appear even smaller. No comparison with Tulsa is possible for PPVT, but for both math and print awareness the effect sizes are somewhat larger than in our states combined (.38 versus .29; and .79 versus .70).

Such comparisons are beset with inferential problems. Methodological differences between studies are a serious confound, as are those within states when we compare our

[5] The procedure used by Ludwig & Phillips (2007) 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 (pg. 22).

Oklahoma results with Gormley et al.’s (2005) Tulsa estimates. First, Gormley et al.’s sample is 3.6 times larger than ours, and these differential sample sizes may well contribute to greater uncertainty about the functional forms of the regression lines for our data. Second, Gormley et al.’s study uses quadratic models, not the cubic functional form that we obtained for math and print awareness after graphical, parametric, and non-parametric analyses. In our data, linear and quadratic models tended to overestimate results and produce larger estimates than those based on the more descriptively accurate cubic functional form. However, when both studies used the same Woodcock-Johnson Applied Problems test and each generated quadratic estimates the results were quite similar (ours in Oklahoma=2.17 raw points; Tulsa=1.94 raw points, p<.05 for both). Our study used a different print awareness measure than Gormley et al.’s, but if we took our quadratic estimate and converted it into an effect size, our print awareness impact would be .58 (p<.05) and Gormley et al.’s .79 (p<.05)—not identical but close and each reliable. So the main discrepancy in results appears to result from the different functional forms in each dataset. For math and print awareness, our data clearly fit a cubic form better than a quadratic one, but our analysis has fewer cases and a reduced density of cases near the cutoff. Are the differences between the Oklahoma and Tulsa results real or artifacts of the models used?

Turning to the Head Start Study (Puma et al., 2005), we note that this evaluation is national, whereas the five states studied here have among the highest quality standards in the nation and are thus not nationally representative. The Head Start Study was explicitly conducted as an effectiveness study, and so is this five-state study except for how the states were selected. Programs that operated as both state pre-K and Head Start are another possible confound. Obviously, a clean contrast of Head Start and state services should omit Head Start centers from the state comparison. Fortunately, for the states in our sample, while some state services were provided under Head Start, the percentage was small and not more than 10 percent. 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. At least 10 percent of the slots are also reserved for children with disabilities whose families may have incomes above the poverty threshold. In this state pre-K sample, Oklahoma offers universal access to services, and West Virginia is expanding its program to serve all four-year-olds. Michigan offers services to those up to 185 percent of the poverty rate, New Jersey's program serves children who reside in districts that had 40 percent or more of its children receive 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 in worse material straights than state pre-K families. A fair comparison would examine state pre-K effects using the same Head Start eligibility criteria.

There is also a difference in the emphasis given to cognitive achievement gains. They are included in Head Start goals and are becoming ever more central to that program. Head Start is unique in its emphasis on health and nutrition programming, parental involvement and education, and coordination of 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 of vision, hearing, and health screenings, referrals to social service, meals and snacks, and parental education. While we know how well Head Start did in non-cognitive areas—nearly all coefficients are positive but quite small and rarely reliable—we do not know how well the state programs did in these areas. Given the emphasis of state pre-K programs on school readiness, one might speculate that any changes the state programs achieved in other domains may not have been large. If we averaged all the Head Start effects across all the cognitive and non-cognitive domains, how different would the state and Head Start results be if their goals are made more equal? The sad truth is that a clean comparison of Head Start and state programs requires random assignment to each within the same study. Since no such study currently exists, all between-study comparisons of average effect sizes are fraught with confounds.

This research project used RDD for its acknowledged theoretical and empirical advantages in justifying unbiased causal inference. RDD is an important tool in the developmental sciences and public policy whenever resources are distributed by merit, need, first come first served or -- as here -- by date of birth. RDD is not as useful as an experiment, however. 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 method of choice, with RDD being an acceptable causal alternative if done carefully

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*(4), 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*(3), 25-50.

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. (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.

Black, D., Galdo, J., & Smith, J. C. (2005). Evaluating the regression discontinuity design using experimental data [Electronic Version]. *Working paper* 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.

Burchinal, M. R., Roberts, J. E., Riggins, R., Zeisel, S. A., Neebe, E., & Bryant, D. (2000). Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development, 71*(2), 339-357.

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal, 32*, 1051-1058.

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

Cicarelli, V. G., Evans, I. W., & Schiller, T. S. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. Athens, OH: Westinghouse Learning Corporation, Ohio University.

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*(403), 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., 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, IL: Rand McNally.

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

Currie, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives, 15*(2), 213-238.

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.

Finkelstein, M., Levin, B., & Robbins, H. (1996). Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *Journal of Public Health, 86*(5), 691-695.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine, 15*(5), 451-474.

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. Unpublished Unpublished manuscript. Institute for Research on Poverty.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The Effects of Universal Pre-K on Cognitive Development. *Developmental Psychology, 41*(6), 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*(1), 65-81.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209.

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.

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

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.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2), 467-475.

Imbens, G. W., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, *64*(4), 555-574.

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*(1), 50-79.

Jacob, B., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, LXXXVI*(1), 226-244.

Johnson, D. L., & Blumenthal, J. (2004). The Parent Child Development Centers and School Achievement: A Follow-Up. *The Journal of Primary Prevention*, *25*(2), 195-209.

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

Loeb, S., & Bassok, D. (2007). Early childhood and the achievement gap. *Working paper*.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*, 52-66.

Loeb, S., Fuller, B., Kagan, S. L., & Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality and stability. *Child Development, 75*(1), 47-65.

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.

Love, J. M., Harrison, L., Sagi-Schwartz, A., van Hzendoorn, M. H., Ross, C., Ungerer, J. A., et al. (2003). Child care quality matters: How conclusions may vary with context. *Child Development, 74*(4), 1021-1033.

Ludwig, J., & Phillips, D. (2007). The benefits and costs of Head Start. *Working Paper 12973*.

Magnuson, K. A., Ruhm, C. J., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26*, 33-51.

Magnusson, K. A., Meyers, M., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41*(1), 115-157.

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. *JAMA, 277*(2), 126-132.

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

National Institute for Early Education Research (2005). New study shows high quality state pre-K programs improve language and math abilities of children of all backgrounds. Retrieved from http://nieer.org/mediacenter/index.php?PressID=46 on 6/11/2007.

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., Tobertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest. *JAMA, 285*(18), 2339-1026.

Robbins, H., & Zhang, C.-H. (1988). Estimating a treatment effect under biased sampling. *Proceedings from the National Academy of Sciences USA, 85*, 3670-3672.

Rock, D. A., & Stenner, J. A. (2005). Assessment issues in the testing of children at school entry. *Future of Children, 15*(1), 15-34.

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.

van der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach. *International Economic Review, 43*(4), 1249-1287.

Votruba-Drzal, E., Coley, R. L., & Chase-Lansdale, P. L. (2004). Child care and low-income children's development: Direct and moderated effects. *Child Development*, *75*(1), 296-312.

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.

Xiang, Z., & Schweinhart, L. J. (2002). *Effects five years later: The Michigan School Readiness Program evaluation through age 10*. Ypsilanti, MI: High/Scope Educational Research Foundation.

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 served by child's age	% of 4 year olds served	Teacher/ child ratio	Maximum class size	Duration	Teacher education	Comprehensive curriculum standard
Michigan	1985	$5,031	24,729 age 4	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 age 4 16,725 age 3	79% of Abbott children*	2:15	15	Full day	BA degree with training in early	Yes
Oklahoma	1990 universal in 1998	$6,167	30,180 age 4	65%	1:10	20	Varied	BA degree with training in early	Yes
South Carolina	1984	$3,219	17,821 age 4 740 age 3	32%	1:10	20	Half-day	BA degree with training in early	No
West Virginia	1983 universal by 2010	$6,829	6,541 age 4 1,370 age 3	33%	1:10	20	Varied	BA or AA degree with training in early	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.

Table 2: Summary statistics

	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)	

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

Table 4: Michigan

		Parametric models used in analysis				Non-parametric estimates by bandwidth		OLS estimates		IV estimates with and without covariates	
	Empirically identified functional form	Linear	Quadratic	Cubic	Truncated at 6 months	50 BW	75 BW	Full sample (ITT)	Restricted sample (TOT)	IV w/o covariates (TOT)	IV w/ covariates (TOT)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PPVT	Quadratic	0.332	-1.911	-3.181	-3.741	-4.990	-1.655	-2.203	-1.911	-1.507	**-2.747**
		(2.488)	(4.153)	(5.617)	(6.096)	(4.400)	(3.914)	(3.637)	(4.153)	(4.878)	**(4.531)**
Math	Linear	2.032*	2.251*	2.474	1.905*	2.990*	2.349*	2.069*	2.032*	1.869*	**1.820***
		(0.562)	(0.902)	(1.289)	(0.872)	(0.863)	(1.017)	(0.549)	(0.562)	(0.509)	**(0.483)**
Print Awareness	Linear	24.978*	21.579*	21.745*	21.790*	19.313*	22.187*	25.210*	24.978*	22.232*	**22.139***
		(3.578)	(5.679)	(7.766)	(5.582)	(5.993)	(5.155)	(3.483)	(3.578)	(3.185)	**(3.105)**
Student covariates		Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
Fuzzy cases		No	No	No	No	No	No	Yes	No	Yes	Yes

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	Parametric models used in analysis				Non-parametric estimates by bandwidth		OLS estimates		IV estimates with and without covariates	
		Linear	Quadratic	Cubic	Truncated at 6 months	30 BW	40 BW	Full sample (ITT)	Restricted sample (TOT)	IV w/o covariates (TOT)	IV w/ covariates (TOT)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PPVT	Linear	5.705*	5.368*	5.256	4.975*	8.094*	7.955*	6.293*	5.705*	8.701*	**6.101***
		(1.438)	(2.019)	(2.715)	(1.925)	(3.861)	(2.637)	(1.519)	(1.438)	(1.789)	**(1.436)**
Math	Linear	0.715*	0.077	0.377	0.268	.392	.494	0.893*	0.715*	1.217*	**0.867***
		(0.352)	(0.469)	(0.596)	(0.463)	(0.707)	(0.710)	(0.380)	(0.352)	(0.393)	**(0.363)**
Print Awareness											
	Cubic	17.159*	11.921*	9.252	6.299	8.704	8.250	8.464*	9.252	16.533*	**13.019***
		(2.471)	(3.726)	(4.828)	(6.679)	(5.646)	(6.532)	(3.844)	(4.828)	(6.277)	**(5.848)**
Student covariates		Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
Fuzzy cases		No	No	No	No	No	No	Yes	No	Yes	Yes

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	Parametric models used in analysis				Non-parametric estimates by bandwidth		OLS estimates		IV estimates with and without covariates	
		Linear	Quadratic	Cubic	Truncated at 6 months	50 BW	75 BW	Full sample (ITT)	Restricted sample (TOT)	IV w/o covariates (TOT)	IV w/ covariates (TOT)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PPVT	Linear	5.648*	5.074	1.710	0.268	1.263	4.333	4.936*	5.648*	5.771	**5.117***
		(2.350)	(3.599)	(4.563)	(0.463)	(5.601)	(3.895)	(2.218)	(2.350)	(3.100)	**(2.308)**
Math	Cubic	2.011*	2.167*	0.483	0.296	.337	1.204	1.334	0.483	1.256	**1.358**
		(0.557)	(0.740)	(1.040)	(1.268)	(1.389)	(0.984)	(0.885)	(1.040)	(0.932)	**(0.903)**
Print Awareness	Cubic	21.013*	15.549*	9.247	0.465	1.039	10.065	11.270	9.247	8.405	**11.464**
		(3.516)	(4.841)	(6.907)	(9.700)	(12.379)	(7.029)	(5.883)	(6.907)	(6.289)	**(6.001)**
Student covariates		Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
Fuzzy cases		No	No	No	No	No	No	Yes	No	Yes	Yes

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	Parametric models used in analysis				Non-parametric estimates by bandwidth		OLS estimates		IV estimates with and without covariates	
		Linear	Quadratic	Cubic	Truncated at 6 months	50 BW	75 BW	Full sample (ITT)	Restricted sample (TOT)	IV w/o covariates (TOT)	IV w/ covariates (TOT)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PPVT	Linear	0.985	-0.187	-0.219	1.362	.088	-1.818	0.788	0.985	0.547	**0.795**
		(2.327)	(3.176)	(4.356)	(3.033)	(3.864)	(5.483)	(2.333)	(2.327)	(2.525)	**(2.351)**
Print Awareness	Linear	21.072*	21.716*	22.831*	25.318*	21.239*	18.512*	20.833*	21.072*	20.252*	**21.005***
		(2.909)	(4.380)	(5.966)	(4.153)	(5.017)	(6.402)	(2.967)	(2.909)	(3.102)	**(2.928)**
Student covariates		Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
Fuzzy cases		No	No	No	No	No	No	Yes	No	Yes	Yes

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	Parametric models used in analysis				Non-parametric estimates by bandwidth		OLS estimates		IV estimates with and without covariates	
		Linear	Quadratic	Cubic	Truncated at 6 months	50 BW	75 BW	Full sample (ITT)	Restricted sample (TOT)	IV w/o covariates (TOT)	IV w/ covariates (TOT)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PPVT	Linear	3.655	2.527	8.249	3.994	4.884	6.39	2.747	3.655	2.686	**2.422**
		(2.387)	(3.469)	(4.526)	(3.112)	(5.792)	(4.250)	(2.227)	(2.387)	(2.022)	**(1.940)**
Math	Quadratic	1.937*	1.530	1.244	0.769	0.764	1.743	0.263	1.530	0.754	**0.435**
		(0.634)	(0.940)	(1.349)	(1.444)	(1.495)	(0.953)	(0.845)	(0.940)	(1.444)	**(1.393)**
Print Awareness	Linear	24.491*	28.024*	28.445*	27.015*	30.488*	30.950*	22.252*	24.491*	20.670*	**20.150***
		(3.496)	(5.032)	(6.381)	(5.097)	(5.471)	(4.969)	(3.586)	(3.496)	(3.099)	**(2.980)**
Student covariates		Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
Fuzzy cases		No	No	No	No	No	No	Yes	No	Yes	Yes

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 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.

Table 11: Comparison of TOT effect size estimates from State Pre-K study (2007), Gormley et al. (2005), and the Head Start Impact Study (2005)

	State pre-K study (2007)						Gormley et al. (2005)	Head Start (2005)**
	Unweighted ES average	Michigan	New Jersey	South Carolina	West Virginia	Oklahoma	Tulsa, OK	Nationally representative
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PPVT	0.14	-0.16	.36*	0.05	0.14	0.29*		0.08
Math	0.29	.47*	.23*		0.11	0.35	0.38*	0.15
Print Awareness	0.70	.96*	.50*	.79*	.83*	0.43	0.79*	.36*

* significant at 5%, ** TOT estimates are from Ludwig and Phillips (2007)
Effect sizes are calculated using sample standard deviations.

Table 12: Comparison of ITT effect size estimates from State Pre-K study (2007), Gormley et al. (2005), and the Head Start Impact Study (2005)

	State pre-K study (2007)						Gormley et al. (2005)	Head Start (2005)***
	Unweighted ES average	Michigan	New Jersey	South Carolina	West Virginia	Oklahoma	Tulsa, OK	Nationally representative
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PPVT	0.14	-0.13	.36*	0.04	0.16	.28*		0.05
Math	0.29	.53*	.23*		0.06*	0.34	NA	0.10
Print Awareness	0.70	1.09*	.32*	0.78*	.92*	0.42	NA	.25*

* significant at 5% *** ITT estimates are from Puma et al. (2005)
Effect sizes are calculated using sample standard deviations.

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

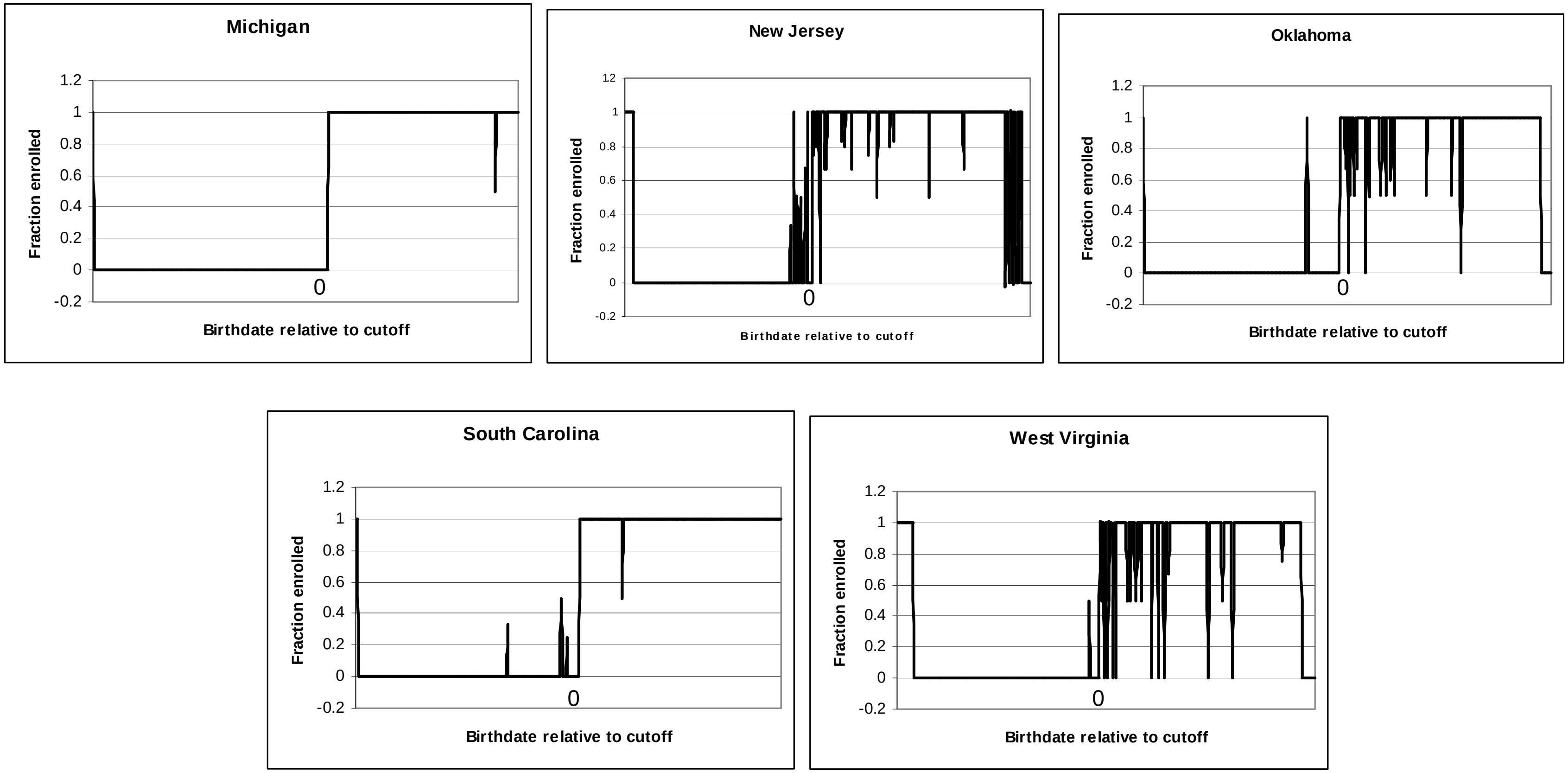
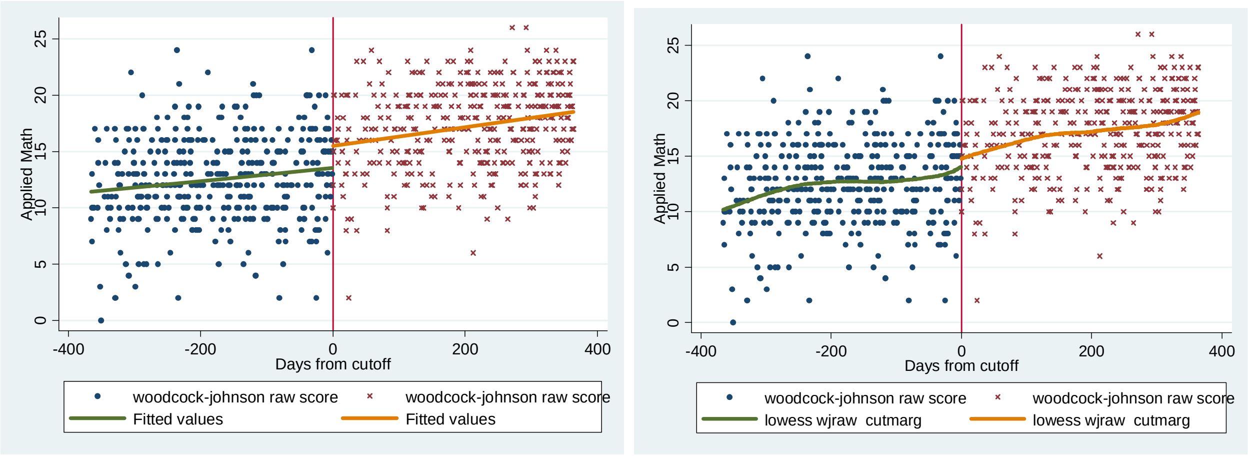


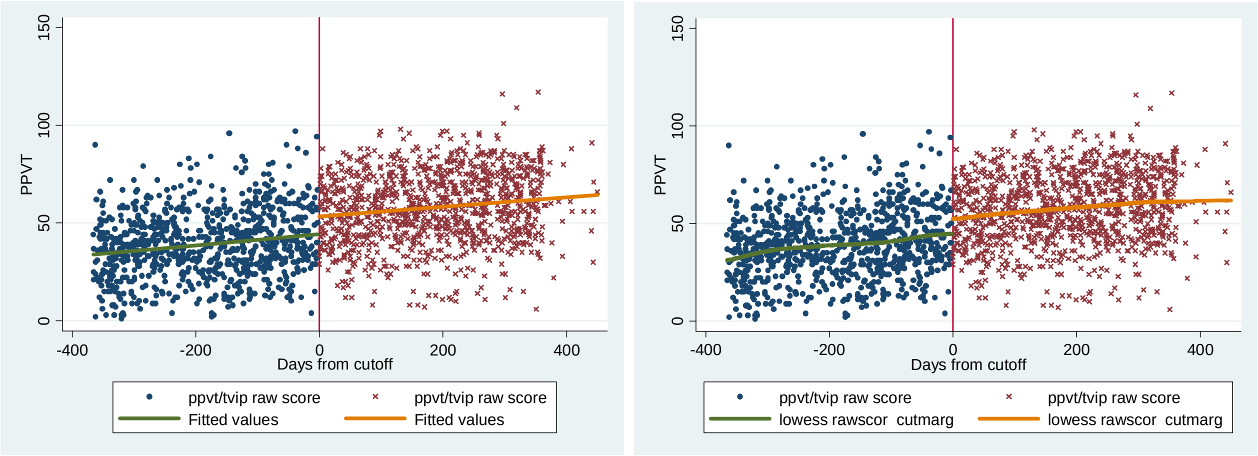
Figure 2: Incorrectly modeled functional form using Oklahoma’s math outcome



(1) Linear plot

(2) Lowess plot

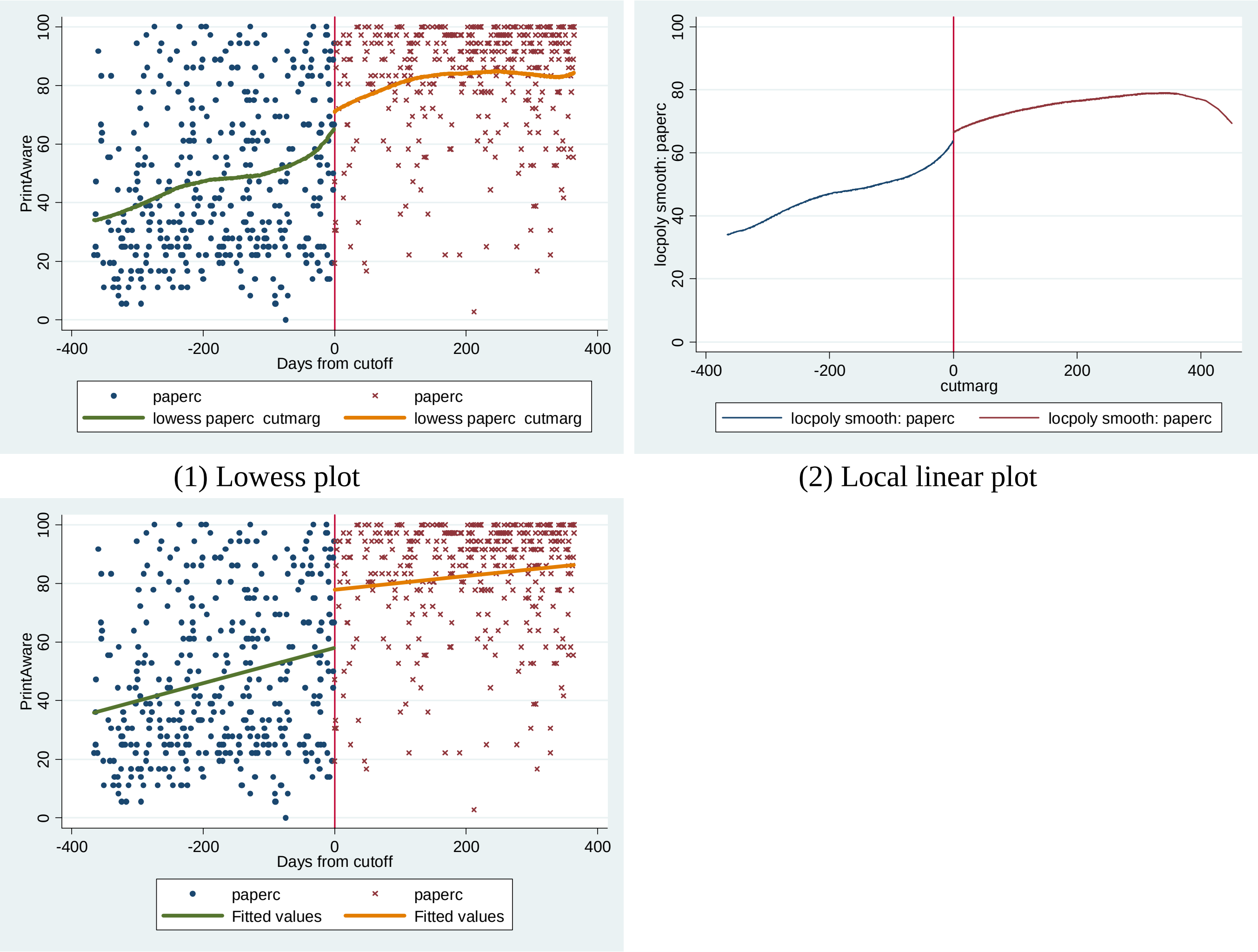
Figure 3: Examples of lowess and linear plots of New Jersey’s PPVT



(1) Linear regression plot

(2) Lowess plot

Figure 4: Examples of lowess, local linear, and linear plots of New Jersey's Print Awareness

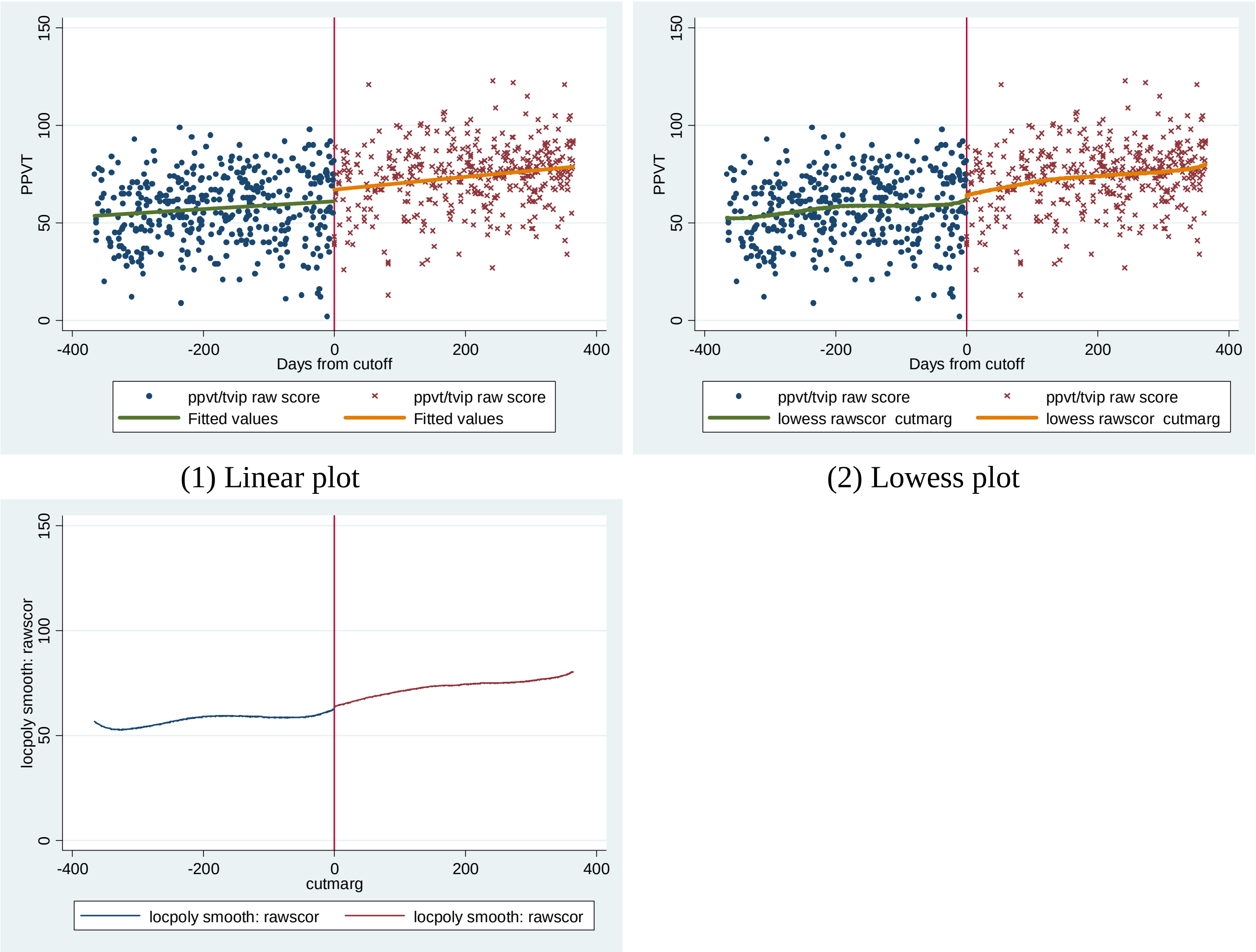


(1) Lowess plot

(2) Local linear plot

(3) Linear plot

Figure 5: Lowess, local linear, and linear plots of Oklahoma's PPVT

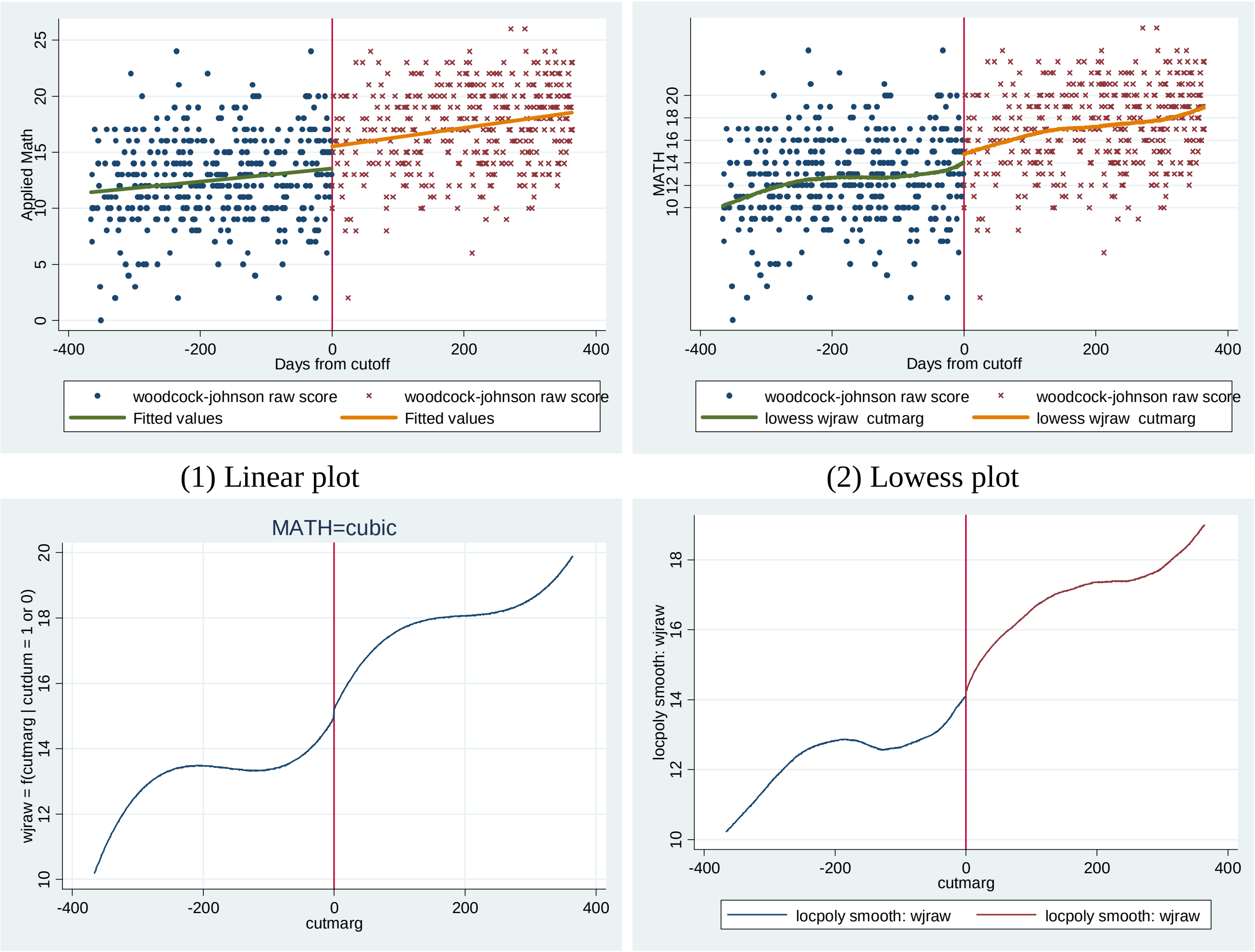


(1) Linear plot

(2) Lowess plot

(4) Local linear plot

Figure 6: Lowess, local linear, linear, and cubic plots of Oklahoma’s Math



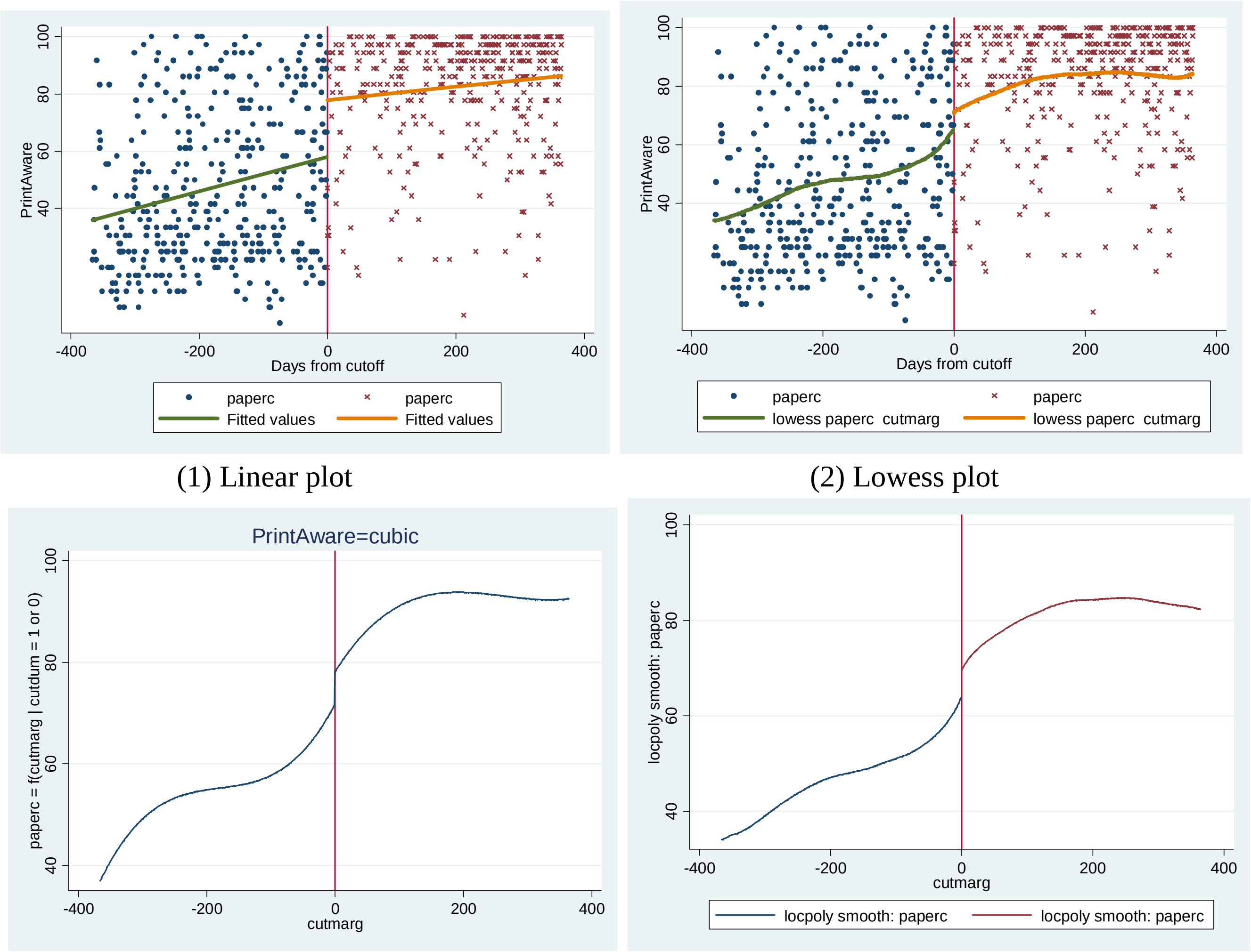
(1) Linear plot

(2) Lowess plot

(3) Cubic model plot

(4) Local linear plot

Figure 7: Lowess, local linear, linear, and cubic plots of Oklahoma’s Print Awareness



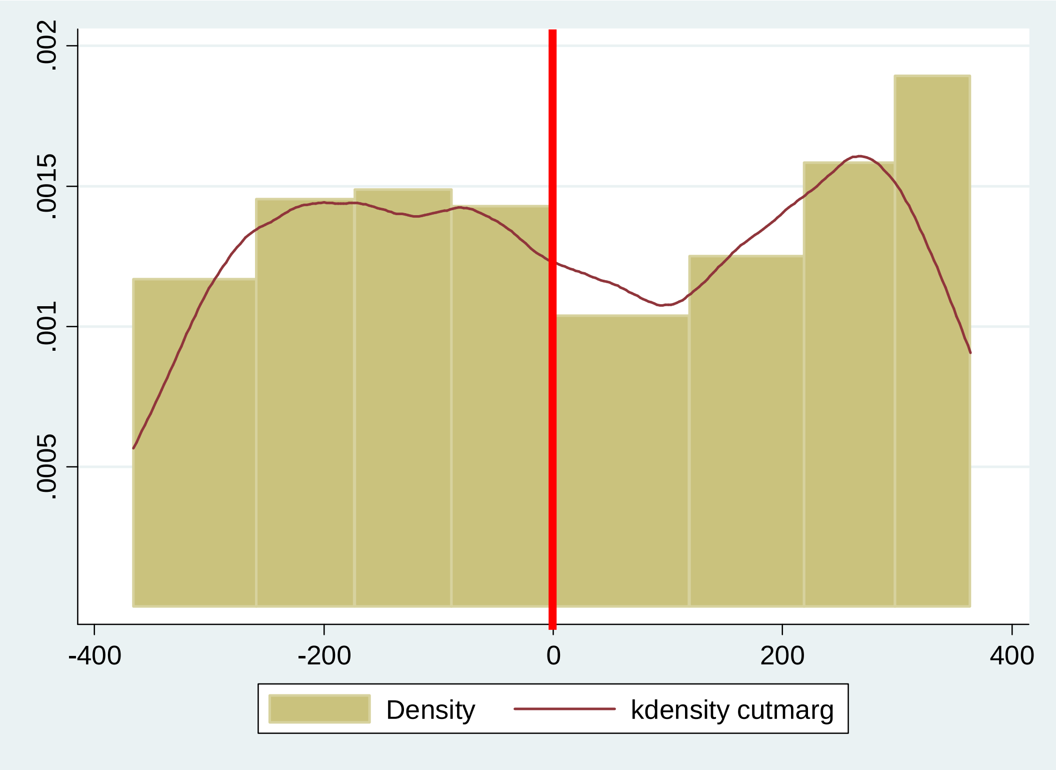
(1) Linear plot

(2) Lowess plot

(3) Cubic model plot

(4) Local linear plot

Figure 8: Oklahoma density plot of observations


Density
kdensity cutmarg