

State Prekindergarten Effects on Early Learning at Kindergarten Entry: An Analysis of Eight State Programs

W. Steven Barnett

Kwanghee Jung

Allison Friedman-Krauss

Ellen C. Frede

Milagros Nores

National Institute for Early Education Research

Rutgers University

Jason T. Hustedt

University of Delaware

Carollee Howes

University of California, Los Angeles

Marijata Daniel-Echols

Michigan Public Health Institute

State-funded prekindergarten (preK) programs are increasingly common across the country. This study estimated the effects of eight state-funded preK programs (Arkansas, California, Michigan, New Jersey, New Mexico, Oklahoma, South Carolina, and West Virginia) on children's learning using a regression discontinuity design. These programs vary with respect to the population served, program design, and context. Weighted average effect sizes from instrumental variables analyses across these states are 0.24 for language (vocabulary), 0.44 for math, and 1.10 for emergent literacy. Differences in effect sizes by domain suggest that preK programs should attend more to enhancing learning beyond simple literacy skills. State preK programs appear to differ in their effects. We offer recommendations for more rigorous, regular evaluation.

Keywords: *state-funded prekindergarten, preK, preschool, regression discontinuity*

STATE-FUNDED prekindergarten (preK) programs are now established in 43 states and the District of Columbia (Barnett et al., 2017) that are home to more than 96% of the nation's 3- and 4-year-olds. A primary goal of these state-funded preK programs is enhancing the learning and development of young children to better prepare them for school success. Rigorous studies have found that high-quality preK education programs can improve children's learning and development in both the short and long term, but public preK programs have on average performed less well than researcher-designed model programs (Camilli, Vargas, Ryan, & Barnett, 2010; Duncan & Magnuson, 2013). The extent to which today's state preK programs produce the types and magnitudes of effects associated with persistent academic improvements in smaller-scale studies remains unclear (Bailey, Duncan, Odgers, & Yu, 2017).

This study seeks to expand knowledge about contemporary public preK by addressing four questions: What are the impacts of eight diverse state-funded preK programs on language, literacy, and mathematics? To what extent do effects vary across

these outcomes and states? More tentatively, we ask if patterns across these states suggest any explanations for the observed variations. Finally, what practical advice for future research and evaluation can we offer based on an assessment of the strengths and weaknesses in this study? Although our study examines only effects at kindergarten entry, these effects have some relevance for those concerned with longer-term outcomes. Some studies indicate that larger initial gains predict larger long-term gains, and initial gains for some skills may be more strongly associated with persistent gains (Bailey et al., 2017; Barnett & Frede, 2017; Hill, Gormley, & Adelstein, 2015).

Over the past 15 years, state-funded preK programs have more than doubled in size to enroll about 1.5 million children. Nearly one third of the nation's 4-year-olds enroll in these programs, which primarily serve children at age 4 (Barnett et al., 2017). Across states, programs vary greatly with respect to funding, eligibility criteria, and virtually every other program feature that can be regulated. The result is a wide range of structural (e.g., teacher



qualifications, teacher pay rates, and teacher–child ratio) and process (as measured by standardized classroom observations) quality as well as variations in hours and age of entry (Barnett et al., 2017). State preK programs all share an expressed intent to provide education typically codified in early learning standards, and in some states preK programs are connected to the public schools more strongly than others (Barnett et al., 2017).

The diversity among state preK programs, together with constantly evolving program standards, participation rates, and contexts, present challenges for evaluating effectiveness (Barnett et al., 2017; Phillips et al., 2017). Evaluations typically have focused on single states or even local examples within states. This study seeks to expand knowledge regarding similarities and differences in the effects of contemporary state preK programs. It goes well beyond an earlier study of five state preK programs—in Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia (Wong, Cook, Barnett, & Jung, 2008)—by using more recent data for three of those states and adding programs in Arkansas, California, and New Mexico. The prior study relied entirely on data from 2004; in the current study the data extend from 2004 to 2015. We also introduce methodological improvements following advances in the field (e.g., Lipsey, Weiland, Yoshikawa, Wilson, & Hofer, 2015). The additional states and new data stretching over a decade provide greater generalizability of findings and a better view of cross-state variation than previously available. Programs studied range from low to high in per-child funding and vary in standards relating to quality as well as program duration and population served. Using common measures and methods across the states, we provide a stronger basis for comparison of a more mature set of programs in a contemporary context.

Findings From Prior Research

A large number of studies have investigated the effects of preK education on the learning and development of young children, with an average effect size of about 0.25 standard deviations (Bailey et al., 2017; Camilli et al., 2010). This is equivalent to about 3 months of learning gains or one quarter to one third of the achievement gap between Black or Hispanic children and their White non-Hispanic peers at kindergarten entry (Friedman-Krauss, Barnett, & Nores, 2016; Yoshikawa, Weiland, & Brooks-Gunn, 2016). However, there is considerable variation around that average effect size. Effects were larger at program completion than in later follow-up assessments, and some programs produced much larger effects than others (Bailey et al., 2017). Plausible explanations for differences in effect estimates include differences in program design; populations served; the broader context, including the availability of alternative programs; and research methods (Barnett et al. 2017; Duncan & Magnuson, 2013; Phillips et al., 2017).

A critical concern for the field is the apparent divergence between past findings and the estimated effects of today’s large-scale public programs. Intensive small-scale programs from past decades had relatively large impacts on language and general cognitive abilities at kindergarten entry, 0.50 to 0.75 standard deviations. Recent studies of large-scale programs have tended to find much smaller effects on similar measures at kindergarten entry, typically about 0.10 standard deviations (Bailey et al., 2017; Magnuson, Ruhm, & Waldfogel, 2007; Puma et al., 2010). However, estimated effects vary substantially even among recent evaluations of large-scale public programs (Phillips et al., 2017). For example, an evaluation of the preK program offered by the Boston Public Schools in 2008–2009 found relatively large positive effects at kindergarten entry for language (0.44 *SD*), letter-word identification (0.62 *SD*), and math (0.59 *SD*) (Weiland & Yoshikawa, 2013).

What can we generalize about state-funded preK from prior research? A recent comprehensive review concluded that a diverse array of public preschool programs has produced positive short-term effects, particularly in the academic areas of literacy and numeracy (Phillips et al., 2017). Evidence of long-term effects is less robust. Yet, given the variation in program design, that review also concluded that “it is not meaningful to talk about state-sponsored pre-k as if it were a single intervention for which we would expect research to reach a general conclusion about whether it ‘works’” (Phillips et al., 2017, p. 25).

Another concern that emerges from a review of the literature is the methodological limitations of public program evaluations. Evaluation designs often have been weak with respect to internal validity and generalizability to a state-wide program (Gilliam & Ripple, 2004; Gilliam & Zigler, 2001). The past decade has seen more attention to this issue, particularly problems of selection bias. A large randomized controlled trial (RCT) has been conducted for the state preK program in Tennessee (Lipsey, Farran, Hofer, 2015; Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013). That study employed propensity score matching with a subsample of children for whom parental consent was received and found effects at preK exit that are broadly consistent with the national impact study of Head Start (also an RCT) and quasiexperimental studies of state preK. However, in a subsequent follow-up, effects disappear by the end of kindergarten and reverse in all domains measured by the end of second grade. In second and third grades, a number of outcomes are significantly negative. These unusual findings increase concern that other designs may have overestimated preK impacts. It is somewhat reassuring in this regard that an earlier, nonexperimental evaluation of Tennessee’s preK program also found long-term negative effects after initial positive results and—as in the RCT—that preK students were more likely to be in special education later (Strategic Research Group, 2008).

Nevertheless, concerns regarding research design and selection bias remain salient because self-selection and eligibility restrictions, such as means testing, are common in state preK. RCTs provide added value, but they face substantial practical hurdles and are not possible for universal programs. Recently, this has led to greater use of “second-best” methods, particularly the age-cutoff variant of the regression discontinuity (RD) design (Cook, Shadish, & Wong, 2008; Lee & Lemieux, 2010; Lipsey, Weiland, et al., 2015). Typically, state and city preK studies using this age-cutoff RD approach have found positive effects at kindergarten entry. The earliest was a study of universal preK ($N = 2,756$) in Tulsa, Oklahoma (Gormley, Phillips, & Gayer, 2008). This study found effects of 0.99 standard deviations for literacy and 0.36 standard deviations for math at kindergarten entry. As with the broader body of research on public preK, estimated effect sizes using this RD approach have varied widely across programs, with somewhat more consistency in positive findings for simple literacy skills than other outcomes (Hustedt, Jung, Barnett, & Williams, 2015; Lipsey, Farran, Bilbrey, Hofer, & Dong, 2011; Manship et al., 2015; Peisner-Feinberg & Schaaf, 2011; Weiland & Yoshikawa, 2013; Wong et al., 2008).

Overall, the age-cutoff RD approach appears to produce somewhat larger estimates of effects than other methods. Three studies permit direct comparisons of estimates between the RD approach and relatively rigorous alternatives that compare those who did and did not attend the preK program in the same cohort for the same year (Barnett & Frede, 2017; Hill et al., 2015; Lipsey et al., 2013). In all three, the RD estimates are larger, perhaps because they measure a different type of preK effect, as we discuss later.

Our study is designed to increase knowledge regarding consistency and variation in the short-term impacts of contemporary state preK programs. We also seek insights into the sources of variation in outcomes, and we hope that future studies will expand the basis for such insights by expanding the number of rigorous state preK program evaluations. To facilitate this, we also seek practical lessons for the improvement of future research and evaluation on state preK.

Method

Using an age-cutoff RD approach, we estimated the effects of a single year of state-funded preK at age 4 in eight states: Arkansas, California, Michigan, New Jersey, New Mexico, Oklahoma, South Carolina, and West Virginia. These states provide a diverse sample of programs operating between 2004 and 2015. To the extent possible, we followed the same procedures in each state. As explained below, it was sometimes necessary to depart from the common procedure, but overall methodology was highly consistent across the states.

State Program Descriptions

Salient features of each state’s program are reported in Table 1 for the year in which data were collected. Programs in Arkansas, California, Michigan, New Mexico, and South Carolina were means tested, and programs in Oklahoma and West Virginia were open to all age-eligible children. New Jersey’s program was open to all age-eligible children living in 31 relatively high-poverty school districts. All programs served 4-year-olds and most served some 3-year-olds, but only New Jersey’s program had most children enter at age 3. Most programs used mixed-delivery systems including both public and private providers. However, our California sample includes only public school classrooms, although the program served some children in private providers. The vast majority of study classrooms were in public schools in Oklahoma and South Carolina, reflecting distribution of settings in the programs overall. All programs employed a teacher and an assistant teacher in each classroom. Maximum class size was 20, except in Michigan, where it was 18; New Jersey, where it was 15; and California, where there was no maximum class size but a 1-to-8 staff-to-child ratio. Most teachers in these programs had BA degrees with specialized training in early childhood, except in California, where just under half had a BA degree and most others had AA degrees. PreK schedules (hours per day and per week) varied across and within programs. Six states had comprehensive early learning standards.

Research Design

This study employs a variant of the RD design (Bloom, 2012; Lee & Lemieux, 2010) that takes advantage of each state preK program’s strict determination of eligibility based on the child’s date of birth. By relying on this assignment rule, one that is unlikely to be related to other child and family characteristics, this approach seeks to reduce the likelihood of selection bias. Typically, preK program effects are estimated by comparing the test scores of children who attended a program with the scores of similar children from the same age cohort who did not attend. Where programs are universal, it can be impractical to find a “comparable” group of children who did not attend. Yet, even where programs target only some children, a problem remains: Those who enroll in state preK are *not* the same as those who do not enroll. PreK programs that target specific types of children create these differences through their eligibility criteria, but differences also come about because some parents choose to enroll their children whereas others do not, or some parents are aware of the program whereas others are not. When randomized trials are not possible, the concern is that such differences will be confounded with, or bias, the estimates of program impact.

The RD approach compares two groups of children who select, and are selected by, a state preK program and takes

TABLE 1

Characteristics of State-Funded PreK Programs Included in Analyses

State	Year established	Spending per child (2015 dollars)	Number 4-year-olds enrolled	% of 4-year-olds in state served	Staff:child ratio	Maximum class size	Duration	Teacher education	Comprehensive learning standards	Means tested
Arkansas 2005	1991	\$6,064 ^a	4,462	12%	1:10	20	7 hours	Mostly BA degree with ECE training	Yes	Yes
California 2006	1965	\$3,928 ^a	52,849	10%	1:8	No limit	3 hours or 6.5 hours	CDA	No	Yes
Michigan 2008	1985	\$4,691 ^a	23,134	18%	1:8	18	Half day or full day	BA degree with ECE training	Yes	Yes
New Jersey Abbott ^b 2005	1998; upgraded in 2002	\$11,293	21,410	79% of Abbott Children	2:15	15	Full day	Mostly BA degree with ECE training	Yes	Not in district
New Mexico 2008	2005	\$2,857 ^a	3,570	13%	1:10	20	Varied	Mostly BA degree in ECE	Yes	Yes
Oklahoma 2004	1990; universal in 1998	\$3,474 ^a	30,180	64%	1:10	20	Varied	BA degree with ECE training	Yes	No
South Carolina 2004	1984	\$1,905	17,821	33%	1:10	20	Mostly half day	BA degree with ECE training	No	Yes
West Virginia 2015	1983; universal by 2010	\$9,898	13,779	68%	1:10	20	Varied	BA degree with ECE training	Yes	No

Note. Data from the annual survey of state prekindergarten programs conducted by National Institute for Early Education Research and reported in the State of Preschool yearbooks (e.g., Barnett et al., 2017). Ratios and class sizes are maximums allowed. Duration often is determined locally. ECE = early childhood education; CDA = Child Development Associate Credential.

a. Represents an incomplete amount of spending per child as it does not include federal and local spending shares.

b. New Jersey's Abbott districts include about one quarter of the state's children.

advantage of the stringent birthdate cutoff that states use to define the groups. One way to interpret this design is to view it as similar to a randomized trial near the age cutoff. RD creates groups that *at the margin* differ only in that some children were born a few days before the age cutoff and others a few days after the cutoff. When these children are about to turn 5 years old, the slightly younger children will enter the preK program and the slightly older children will enter kindergarten having already attended the preK program. If all children are tested at that time, the difference in their scores can provide an unbiased estimate of the preK program's effect under reasonable circumstances. Obviously, if only children with birthdays a few days on either side of the age cutoff were included in a study, the sample size would be unreasonably small. Alternatively, the RD design can be viewed as modeling the relationship between the assignment variable (age) and children's outcomes. The younger sample models the relationship without treatment. The older sample is used to model the relationship after treatment. This approach can be applied to wider age ranges around the

cutoff. However, its internal validity depends on correctly modeling the relationship.

Sample

In each state included, we sampled two groups of children. The comparison group was composed of children who were just entering the state preK program (having missed the birthdate cutoff for preK the prior year). The treatment group was composed of children who were just beginning kindergarten, after having completed preK the prior year. Total sample size varied across states based on the budget constraints for each state's evaluation. Participation was based on passive consent except in Michigan and West Virginia, where active parental consent was required. We sought to assess both groups as early in the academic year as possible.

In four states, random sampling was designed to represent the statewide population served by state preK. In New Jersey, the sample was selected from the 21 largest school

districts within the state's court-ordered "Abbott" preK program (enrolling the vast majority of children) in order to minimize travel costs. Smaller New Jersey districts had higher observed quality at the time, so this sample may have represented less effective programs than the initiative as a whole. In California, the sample was collected from four regions (Los Angeles, San Diego, Fresno, and Sacramento) in order to minimize travel costs. In Michigan, Detroit could not be persuaded to participate in the study. In West Virginia, the study was limited to seven of 55 counties.

Differences in size between the treatment (entering kindergarten, having completed preK) and comparison (entering preK) groups primarily reflect chance variations, including weather conditions, holiday schedules, illness (e.g., a flu outbreak when data collection was scheduled), and bureaucratic hurdles (sometimes a principal or program director would deny data collectors access despite prior agreement at the district level). In California, a larger comparison group sample was intentionally collected as these children participated in another study. In Michigan, researchers had more difficulty securing access to kindergarten classrooms, and in Arkansas, they had more difficulty accessing preK classrooms.

Comparison groups. As it can be difficult to obtain advance class rosters for preK at the beginning of the school year, we developed a sampling strategy that did not require student lists in advance. We gathered information on the number and location of state preK students, sites, and classrooms. Then, we randomly selected preK classrooms to generate the desired sample size assuming four randomly selected children per classroom. Random-number lists generated for each classroom were used to select the specified number of students from class rosters obtained when data collectors visited classrooms.

Treatment groups. Treatment group sampling procedure varied somewhat across states, depending on details such as relationships between preK attendance zones and school districts and access to kindergarten classrooms. In New Mexico, children who had completed preK were selected from master state enrollment lists and tracked to their kindergarten classrooms. In other states, in each school district, we sampled the same number of kindergarten classrooms as preK classrooms, matched against lists of those who had attended preK in the prior year, and selected at random when more than four former preK students were available in a class.

Sample characteristics. Table 2 summarizes descriptive statistics for the state samples, which are quite diverse and include states from every region of the country. These states vary considerably from each other with respect to children's family background and preK entry ability levels. As judged

by scores at preK entry, New Jersey's program served the most educationally disadvantaged population. Oklahoma and West Virginia served the least educationally disadvantaged populations, as might be expected given that their programs are open to all children regardless of family income. Within states, the demographics of treatment and comparison groups are for the most part similar (note that percentages for race and free-lunch eligibility in Table 2 include *missing* as a category, so these are not percentages of nonmissing).

Few significant differences in demographics between groups were found in analyses of whether each variable differs at the cutoff (see online Supplementary Tables A1 through A8). Differential success in obtaining official demographic data between preK and kindergarten in some states resulted in differences in missing data. In this study, preK programs were less likely to have data on free-lunch status than were kindergartens.

Child Assessment Procedures

Children were tested in the fall of the school year, as soon as feasible given the reluctance of some school districts to permit assessment in the first weeks of school. Assessments were conducted in English or Spanish depending on the child's strongest language as ascertained from the classroom teacher, though all children were given the assessment of English vocabulary. 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 by experienced assessors employed and trained by the research project, and assessments were scheduled to avoid meal, nap, and outdoor play times. Testing sessions lasted 20 to 40 minutes.

Measures of Learning

Our study focused on cognitive development using relatively low-cost, easily administered, direct assessment measures. We did not want to rely on teacher assessments as preK and comparison groups would be assessed by different types of teachers with different expectations, possibly introducing bias confounded with the age cutoff. Cost constraints and the lack of inexpensive direct measures of social-emotional development limited the range of outcomes we measured.

Children's language, specifically, receptive vocabulary, was measured by the Peabody Picture Vocabulary Test (3rd ed.; PPVT-III; 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 orally to identify. The PPVT-III directly measures vocabulary size. This test is also used as a quick indicator of general cognitive ability, correlates reasonably well with other measures of linguistic and cognitive development related to

TABLE 2

Characteristics of Sample Children in Each State-Funded Prekindergarten Program

State	N	Language	Math	Literacy	Fuzzy cases	Black	Hispanic	Native American	White/Asian	Other	Race missing	Female	No free lunch	Free lunch	Lunch status missing	Assessed in Spanish
Arkansas	901	61.55 (19.72)	13.62 (4.65)	64.96 (29.89)	2%	36%	6%	0.3%	57%	2%		48%	19%	44%	37%	1.2%
Comparison	390	50.69 (18.86)	11.10 (4.02)	44.47 (27.08)	0.3%	35%	6%	0.5%	58%	0.3%		48%	15%	37%	48%	0.8%
Treatment	511	69.42 (16.32)	15.51 (4.17)	79.83 (21.99)	3.3%	36%	6%	0.2%	56%	3.5%		48%	22%	48%	30%	1.6%
California	1630	48.22 (22.08)	11.27 (5.16)	56.15 (31.11)	8%	6.5%	58%		32%	2%	2%	50%				33.3%
Comparison	1071	40.44 (19.93)	9.21 (4.46)	40.64 (25.21)	3%	5.5%	59%		32%	1.5%	1.8%	49%				34.6%
Treatment	559	62.25 (18.56)	15.20 (4.00)	86.24 (15.49)	17%	8.4%	55%		32%	2.5%	1.6%	51%				30.6%
Michigan	634	62.86 (18.55)	13.32 (4.79)	337.73 ^a (30.54)		6.6%	3.2%		53%	2.5%	34.5%	52%	41%	25%	34%	
Comparison	469	57.5 (16.81)	11.67 (4.07)	326.23 ^a (23.86)		6.8%	3%		46%	3%	42%	52%	36%	22%	42%	
Treatment	165	77.42 (14.93)	17.77 (3.58)	368.72 ^a (24.45)		6.1%	5%		74%	1%	14%	50%	53%	34%	13%	
New Jersey	1538	46.32 (19.24)	11.19 (4.63)	58.43 (28.84)	1.5%	39%	53%	0.2%	5.8%	1.9%		53%				29%
Comparison	780	37.04 (16.35)	8.86 (3.79)	43.90 (25.25)	0.9%	39%	54%	0.3%	5.5%	1.8%		54%				34%
Treatment	758	55.54 (17.37)	13.55 (4.19)	73.31 (24.35)	2.1%	40%	52%	0.1%	6%	2.0%		51%				24%
New Mexico	1333	53.17 (22.43)	12.04 (5.14)	50.12 (32.33)	3%	2%	58%	14%	24%	0.7%	1%	54%				11%
Comparison	685	45.19 (20.70)	9.48 (4.36)	30.84 (24.71)	3%	1%	56%	15%	26%	0.7%	1.5%	55%				13%
Treatment	648	61.60 (21.08)	14.75 (4.47)	70.47 (26.42)	2%	2%	59%	14%	23%	0.6%	0.9%	52%				10%
Oklahoma	836	66.07 (18.80)	14.92 (4.41)	65.41 (29.22)	4%	7%	7%	12%	64%	1%	8%	51%	32%	50%	18%	2%
Comparison	406	57.70 (17.40)	12.58 (3.82)	47.79 (26.93)	0.5%	7%	5%	12%	68%	1%	7%	54%	34%	44%	22%	2%
Treatment	430	73.87 (16.58)	17.14 (3.76)	81.83 (20.37)	7%	7%	8%	13%	61%	1%	9%	47%	30%	55%	15%	2%
South Carolina	777	58.55 (19.28)	NA	62.18 (29.90)	0.9%	44%			40%	4%	13%	51%	35%	54%	11%	
Comparison	424	50.44 (17.62)	NA	45.17 (26.79)	1.4%	45%			37%	4%	15%	51%	35%	50%	15%	
Treatment	353	68.12 (16.59)	NA	82.07 (19.14)	0.3%	42%			44%	3%	10%	52%	35%	59%	6%	
West Virginia	1048	88.52 (22.69)	14.33 (4.91)	21.84 ^b (11.16)	2.19%	4%			93%	3%		49%		73% ^c		
Comparison	475	76.48 (21.17)	11.59 (4.30)	13.56 ^b (9.37)	0.4%	5%			91%	4%		49%		74% ^c		
Treatment	573	98.48 (18.74)	16.59 (4.18)	28.68 ^b (7.18)	4%	3%			94%	3%		49%		72% ^c		

a. Woodcock-Johnson Tests of Achievement (3rd ed.) Letter-Word Identification W score.

b. Test of Preschool Early Literacy raw score.

c. Low-income group designation, which differs somewhat from free lunch status.

Fuzzy cases are those that violate the birthdate assignment rule. For race and free lunch status percentages include missing as a category.

school success, and has good test-retest and split-half reliability. Spanish-speaking children were also tested in Spanish with the Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, Lugo, Padilla, & Dunn, 1986). Although we do not report the results here, alternative analyses conducted using the TVIP instead of the PPVT or the highest standard

score between TVIP and PPVT for these children do not substantively alter findings.

Children’s early mathematical skills were measured with the Woodcock-Johnson Tests of Achievement (3rd ed.; WJ III; Woodcock, McGrew, & Mather, 2001) Subtest 10: Applied Problems. Spanish speakers were given the Bateria

Woodcock-Muñoz Pruebas de Aprovechamiento–Revisado (Woodcock & Muñoz, 1990) Prueba 25: Problemas Aplicados. Woodcock-Johnson achievement subtests have good reliability and have been widely and successfully used in studies of the effects of preschool programs, including Head Start.

Early literacy abilities were measured in six states using the Print Awareness subtest of the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP; Lonigan, Wagner, Torgeson, & Rashotte, 2002) and the equivalent Spanish version. Items measure recognition of individual letters, letter-sound correspondences, and whether children can differentiate words in print from pictures and other symbols. We used the percentage of items answered correctly out of the 36 Print Awareness subtest items. The Pre-CTOPPP is a precursor of the Test of Preschool Early Literacy (TOPEL), which has been reported to have adequate reliability and consistency (Lonigan, Wagner, Torgeson, & Rashotte, 2007). In West Virginia, we used the TOPEL. In Michigan, we used the Letter-Word Identification subtest of the WJ III (Woodcock et al., 2001) as the measure of early literacy.

Statistical Analyses

We estimated treatment effects of preK for each state using the following basic function:

$$Y_{ij} = \beta_0 X_{ij} + \beta_1 (\text{PreK})_{ij} + \beta_2 g(\text{AV})_{ij} + \beta_3 D_k + \varepsilon_{ijk}.$$

In this equation, Y is one of three cognitive test scores measured at program entry, X is a vector of student covariates (gender, ethnicity, free-lunch status, whether the child was tested in English or Spanish, and number of days between school entry and test administration), PreK is a binary variable that takes on the value of 1 if the student participated in preK and 0 if not (i.e., the student just entered preK), $g(\text{AV})$ is a smooth function of the age variable that may include polynomials and interactions, D represents school district fixed effects, ε is the error term, i is the individual subscript, j is the teacher subscript, and k is the district subscript. Analyses were conducted using raw scores. Dummy variables were included as missing data indicators for demographic variables. Most often there was baseline equivalence between groups on demographics including their missing data indicators, a condition necessary for valid use of the dummy-variable approach (Groenwold et al., 2012). We also conducted our preferred analyses using 50 imputations to adjust for missing data. We find no meaningful differences in outcomes between the two approaches to missing data. For California, where testing extended past December and more of the comparison group than treatment group was assessed after December, we also reestimated all models using a sample restricted to those tested by December.

Huber-White standard errors for parametric analyses were used to account for clustering by classroom.

To aggregate and examine effects across states, we employed random-effects meta-analysis of our preferred estimates (instrumental variable [IV] estimates using multiple imputation and 6-months bandwidth). This provided precision-weighted means, a homogeneity test (Q), and tests of significant differences in effect sizes across states.

General Assumptions and Sensitivity Analysis

The validity of the RD approach depends on several assumptions. First, the cutoff must be exogenously imposed and not manipulated. Second, the impact models must include the assignment variable (age), and the functional form must be correctly specified. Bandwidth (the age interval over which data are analyzed) must not be manipulated to influence the results. Third, program participation is strictly determined by the assignment variable, and there is minimal misallocation around the cutoff. Fourth, the observed and unobserved characteristics of participants should vary smoothly and continuously across the cutoff. We examined each of these assumptions in depth.

Exogeneity of Birthdate

The RD approach assumes the “running variable” is not controlled and manipulated at the cutoff to affect program participation. There is evidence of family background differences in season of birth based on expected weather at delivery, reducing the number of planned winter births (Buckles & Hungerman, 2013). However, fall births do not differ appreciably by family background, and there is no evidence that relationships between family or infant characteristics and birthdate vary abruptly around the cutoff (Buckles & Hungerman, 2013).

We consider it implausible that parents would manipulate birthdate (or records of birthdate) to alter eligibility around the cutoff. Moreover, the cutoff varies over time, across states, and within states (e.g., in New Jersey, district cutoffs varied from September 1 to December 1), limiting parental ability to even know the cutoff when planning a pregnancy. Nevertheless, we used the McCrary (2008) test to empirically examine whether participants may have manipulated determination of the assignment variable around the birthdate cutoff. We found indications of discontinuity in the density-of-assignment variable at the cut point in California, Michigan, and New Mexico. We also generated density distribution graphs of the difference between children’s birthdate and the birthdate cutoff. If there is no accumulated density around the cutoff point, we can be more confident that participants did not manipulate their positions relative to the cutoff point. These graphs appear as online Supplementary Figures A1 to A8. In each state, there is no obvious piling up of density close to the cutoff point,

and the density is roughly distributed uniformly. In contrast to the McCrary test, these suggest no obvious sign of participants manipulating their positions relative to the cutoff point in any state.

Model Specification and Bandwidth

To guard against model misspecification, we employed two strategies: a graphical analysis and a series of parametric regressions with alternate specifications of the functional form. To investigate the appropriate functional form, we began by graphing test scores for each state. For each of the three assessments, two types of lines are fitted onto scatterplots on each side of the birthdate cutoffs. The first is a linear regression line. The second is a nonparametric regression line based on locally weighted scatterplot smoothing, called lowess, which relaxes assumptions about the form of the relationship between the assignment and outcome (Cleveland & Devlin, 1988). For each test score observation (Y_{ij}), a smoothed value is obtained by weighted regressions involving only those observations within a local interval. Observations closer to Y_{ij} are weighted more heavily than those farther away. In only a few instances did the graphical analyses suggest possible problems (see online Supplementary Figures A9 to A16). However, these analyses did not employ covariates adjusting for characteristics of children other than age. Further information on functional form was obtained from parametric analyses that did control for other student characteristics.

In a series of parametric analyses, we first estimated alternative single-equation models with higher-order polynomial forms of each equation, including squared and cubic transformations of the selection variable (the difference between birthdate and cutoff date) and its interaction with the cutoff dummy variable. The inclusion of higher-order terms entails some loss of statistical power but does not bias estimates of treatment effects if they are in fact irrelevant. In addition, we examined the potential influence of bandwidth by estimating these linear and polynomial models on subsets of the data truncated at plus or minus 3 and 6 months around the birthdate cutoff, as well as for the full 12 months, and by optimal bandwidth selection for a local polynomial estimator (Calonico, Cattaneo, & Titiunik, 2014). Consistency across these alternate estimates instills confidence that our estimates do not suffer from misspecification of the functional form. Whenever higher-order terms were statistically significant, we evaluated the alternate models against comparable estimates from the truncated samples and examined their consistency with the graphical analyses, using all of this information to select the best model. In reporting findings, we place greatest confidence in those where there is clear agreement across parametric, local polynomial, and graphical nonparametric analyses.

Adherence to Assignment Rule

The RD approach also relies on the assumption that within each state sample the determination of preK participation depended only on a child's birthdate. Children with birthdates before the state cutoff can enroll in a given year, but those born after the cutoff date must wait until the following year. Table 2 reports the percentage of cases that violate this assumption (considered "fuzzy" cases). In no state except California did this exceed 8%, and in most, it was far lower. In California, the treatment group contained an exceptionally high percentage (17%) of cases violating the assumption. California appears to allow exceptions to its cutoff date fairly frequently. When there are few fuzzy cases, it has been found that excluding the participants violating the assignment rule makes little difference (Shadish, Cook, & Campbell, 2002). However, in several states, the percentage of fuzzy cases exceeds 5%, so the potential for this violation of policy to create problems cannot be ignored.

To address the issue of children whose enrollment does not follow the birthdate cutoff, we conducted the parametric analyses described above with the full sample, including the cases that violate the assignment rule. Results from these analyses are reported in the Supplementary Materials (online Supplementary Tables A17 to A25). In addition, we conducted IV analyses in which we treat the problem as one of omitted-variable bias that requires identifying an instrument correlated with treatment assignment but not with error in the outcome (Barnow, Cain, & Goldberger, 1980). States' enrollment rules allow us to treat students' age eligibility for preK as an instrument for their actual participation (Lee & Lemieux, 2010; van der Klaauw, 2008). Again, we evaluated alternate functional forms using higher-order polynomials and performed a graphical review of the data to select the most appropriate functional form. The IV results are presented as our preferred estimates (except for Michigan, which had no fuzzy cases).

Continuity Around the Cutoff

As no other characteristics of the participants should vary sharply at the cutoff, an important test for the validity of an RD is whether there are observed discontinuities in baseline measures at the cutoff. We examined ethnicity, gender, home language, and free-lunch status and found few significant discontinuities at the birthdate cutoff for these measures. Five states had no significant differences in a 6-month age span on either side of the cutoff. We found one significant (but small) difference in three states, each different. Full test results are provided in online Supplementary Tables A1 to A8.

Data collection timing should be similar for the treatment and comparison groups as differences could bias the results. In California, Michigan, New Jersey, and New Mexico, kindergarten children (children in treatment group) were more likely to have been tested at a later date than children in the

control group, which could bias estimates of preK impacts if not taken into account. We include in our analyses the number of school days prior to assessment to account for such differences. California had the most extreme differences in timing, and we reanalyzed that state program using only fall assessments from September through December. Tests of differences in data collection timing are presented in online Supplementary Tables A9 to A16.

Specific Threats to Validity of the Birthdate Cutoff RD Design

Additional concerns have been raised specific to the birthdate cutoff RD and potential violations of basic assumptions (Lipsey, Weiland, et al., 2015). These include that (a) the age difference may interact with assessment to introduce biases, (b) differences in cohort composition and experiences may be confounded with treatment, and (c) effect estimates may not be comparable to those generated by RCTs and other RD approaches. We address each below.

Assessment Issues

The age difference between treatment and comparison groups could lead to bias, either because of floor or ceiling effects or because adaptive measures that vary the starting point based on age would favor the older group. We see evidence of floor or ceiling effects only for print awareness, where estimated effects are nevertheless uniformly large (online Supplementary Figures A9 to A16). A rationale for adaptive assessment is to avoid floor and ceiling problems. In Arkansas and West Virginia, every child was assessed from the beginning on the PPVT, whereas standard procedures were followed for the other states. We see no distinct differences in the data or findings as a result. Another age cutoff RD study that examined this issue found only a small percentage of scores affected and no differential effects by group (Weiland & Yoshikawa, 2013).

Potential Confounds

The age-cutoff approach as we have implemented it has the potential to introduce differences between the two groups that are unrelated to preK participation. Cohort differences can arise for various reasons, including preK program expansion that allows more of the eligible population to attend or changes in the availability of alternative programs. The samples may differ due to differential attrition. The treatment group includes only children who remain through kindergarten entry, whereas the comparison group is composed of children just arriving for preK who have not had the opportunity to attrite. Finally, in anticipation of kindergarten entry, parents might increase home learning activities, thereby creating a home education difference between the two groups that was discontinuous at the birthdate cutoff.

We investigated the potential attrition problem in two ways. First, we assessed the extent to which cohort and attrition differentially affect the samples by comparing the demographics of the treatment and comparison groups. As noted earlier, we find few significant differences in family background that would substantiate such a problem. Second, data from two states allowed us to reanalyze the data after equating attrition between the two groups (see online Supplementary Tables A17b and A25b). For West Virginia, we estimated effects with the full data set and then again using only data for children in both groups who could be located at the beginning of kindergarten. For Arkansas, we estimated effects with the full data set and then again using only data for children in both groups who could be located at the end of kindergarten. Estimates from the reanalysis were nearly identical to the originals with one exception: The West Virginia results for language were more consistent across different bandwidths and with the graphical analysis in the restricted sample. Therefore, we use estimates based on samples equated for attrition as our preferred results for West Virginia.

Some have speculated that parental efforts to educate children might be higher for the children entering kindergarten and could vary sharply at the cutoff (Lipsey, Weiland, et al., 2015). We could not test this hypothesis with our data but did consider other evidence. National surveys find the extent to which parents read to children, tell them stories, and teach numbers, letters, and words varies hardly at all from age 3 to 5 (Snyder, de Brey, & Dillow, 2016). However, Weiland, McCoy, Grace, and Park (2017) found evidence that parents in the national Head Start Impact Study modestly increased language and literacy activities in anticipation of kindergarten entry. Finally, the age-cutoff RD study of California's Transitional Kindergarten (TK), in which both groups enter kindergarten at the same time, nevertheless found substantive positive effects of TK on literacy and mathematics (Manship et al., 2015).

Interpretation

The age-cutoff RD approach creates several issues for clear interpretation of the estimated effects (Lipsey, Weiland, et al., 2015). As Weiland and Yoshikawa (2013) explain in detail, estimates from the age-cutoff RD are neither intent-to-treat nor treatment-on-treated (TOT) estimates, as these are typically understood. The samples are not composed of all eligible children but contain only children who enter the program. Children in the treatment group are lost through attrition (as discussed earlier), and both groups likely contain some children who did not regularly attend or complete the program. From this perspective, our results are closest to TOT estimates. Furthermore, our estimates are local average effects for the sample around the cutoff. Children around the cutoff could differ from others with respect to family background, ability, and other characteristics that might interact

Estimated Effect Sizes

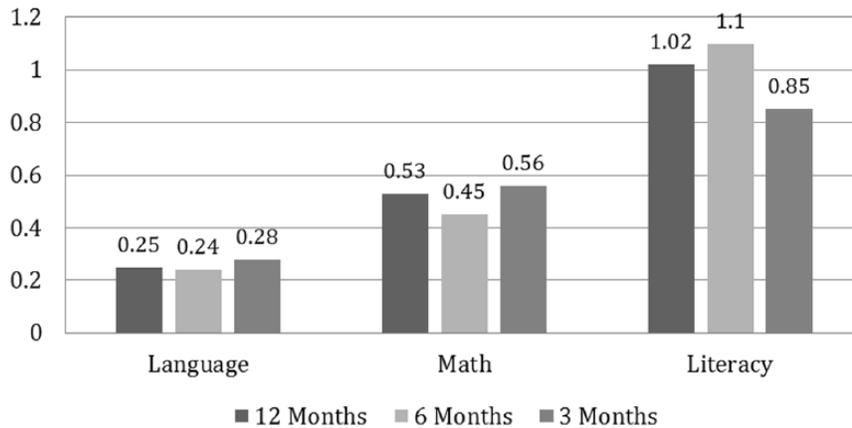


FIGURE 1. Unweighted average estimated state prekindergarten effect sizes for language, math, and literacy by bandwidth using instrumental variable estimates. Language was measured using the Peabody Picture Vocabulary Test (3rd ed.). Math was measured using the Woodcock-Johnson Tests of Achievement Applied Problems subtest. Literacy was measured using the Preschool Comprehensive Test of Phonological and Print Processing except in West Virginia, where the Test of Preschool Early Literacy was used, and Michigan, where the Woodcock-Johnson Tests of Achievement Letter-Word Identification test was used.

with treatment. However, children born in the fall are fairly representative, whereas those born in winter are slightly more disadvantaged than those born in summer (Buckles & Hungerman, 2013).

One interpretation of the age-cutoff design is that it estimates the effect of an added year of preK over typical experience. Across the full sample, the comparison group has the same experiences as the preK group did in the prior year if all assumptions hold. By contrast, an RCT estimates impacts relative to the alternatives to state preK available at the same age. However, children's alternative experiences likely vary with age in the RD, with the oldest children in the comparison group having more experience of other care and education outside the home. Parental need for childcare is presumably very nearly the same on both sides of the age cutoff, and alternatives to state preK, such as private preschools and Head Start, do not employ the same birthdate rules for entry as state preK. Following this logic, the comparison group's average participation rate in other care and education would increase to be more similar to that in an RCT as the bandwidth around the cutoff shrinks.

Results

Our findings are broadly summarized in Figure 1 as unweighted average IV effect sizes from across the eight states for each outcome domain and using each of the three bandwidths (3, 6, 12 months). Effects roughly average 0.25 standard deviations for language, 0.50 standard deviations for math, and 1 standard deviation for emergent literacy skills regardless of the bandwidth around the cutoff.

Table 3 reports IV estimates for each state with a 6-month bandwidth. It offers a much narrower span around the cutoff than 12-month estimates but a much larger sample size than 3-month estimates. Table 3 includes two sets of estimates for California, one based only on data from the fall and one using all data. For Table 3, meta-analysis provided precision-weighted average effect sizes with 95% confidence intervals. Average effect sizes were robust with respect to the fall-only California sample or dropping three states that did not pass the McCrary test. Heterogeneity tests (Cochran's Q) rejected the hypothesis of one true effect across all eight states for each outcome. Significant differences in effects across states are discussed by type of outcome below.

Complete results including all sensitivity analyses are provided for each state in online Supplementary Tables A17 to A25, including results for each functional form and bandwidth and for IV estimates. Results of nonparametric analyses are presented in online Supplementary Figures A9 to A16. Graphical analyses informed the selection of the functional form and were weighed in our overall conclusions regarding each state's outcomes. Findings using the empirically identified functional forms are summarized in online Supplementary Tables A26 to A28 for each bandwidth. The chief difference by bandwidth is that statistical significance is less likely as the age span is reduced and sample size declines. We report estimated effects and statistical significance for each state from both 12- and 6-month analyses for each measure in Figures 2 to 4.

Language

Estimated effect sizes for state-funded preK programs on the PPVT-III averaged 0.24 standard deviations in the IV

TABLE 3

Estimate Effect Sizes of State-Funded Prekindergarten Participation (6 Months Instrumental Variable Analysis)

State	Sample size		Effect size		
	n_{PreK}	n_{K}	Language	Math	Literacy
Arkansas	204	252	0.25	0.26	1.08***
California (all months ^c)	572	239	0.40	0.43	1.40***
California (fall months ^d)	566	109	0.11	0.42	1.39***
Michigan	263	66	0.52	1.10***	1.34***
New Jersey	384	374	0.34*	0.35*	0.50***
New Mexico	350	300	0.07	0.24	1.04***
Oklahoma	214	173	0.33	0.52*	0.82***
South Carolina	230	162	0.12	NA	0.93***
West Virginia	234	257	0.21	0.25 ^b	1.72*** ^a
Weighted average		Effect	0.28	0.44	1.10
		CI	[0.183, 0.367]	[0.243, 0.643]	[0.855, 1.349]
Weighted average with California		Effect	0.24	0.44	1.10
fall months data ^d		CI	[0.141, 0.336]	[0.241, 0.642]	[0.854, 1.346]
Weighted average without		Effect	0.26	0.34	1.01
California, Michigan, New Mexico		CI	[0.187, 0.326]	[0.242, 0.440]	[0.654, 1.363]

Note. Effect sizes are calculated using sample standard deviations. Linear functional forms are used unless otherwise noted. Weighted averages from random effects meta-analysis. PreK = prekindergarten; K = kindergarten; CI = 95% confidence interval.

a. Quadratic functional form

b. Cubic functional form.

c. California all months data include all of the children tested.

d. California fall months data are restricted to children assessed in the fall (September to December).

* $p < .05$. *** $p < .001$.

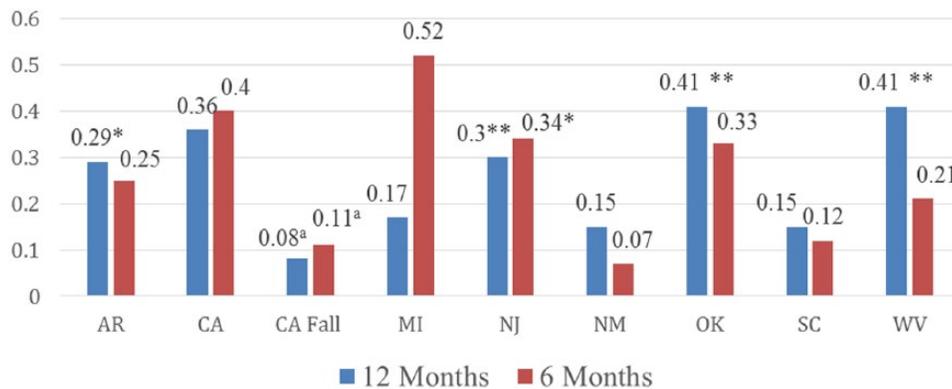


FIGURE 2. Estimated effects sizes for language (Peabody Picture Vocabulary Test, 3rd ed.) by bandwidth, using instrumental variable estimates.

a. California fall months data are restricted to children assessed in the fall (September to December).

* $p < .05$. ** $p < .01$. *** $p < .001$.

analyses with 6-month bandwidth (restricting the California analysis to children assessed in the fall). Two states, New Mexico and South Carolina, had effects of about 0.10 standard deviations, as did California when the sample is restricted to data collected in the fall. Only New Jersey's effect is significantly different from zero in the 6-month analyses, although effects for three other states are significant in the 12-month analyses. Michigan's estimated effect

is significantly above average and New Mexico's significantly below average.

With two exceptions, the graphical analyses are consistent with the parametric model. For Arkansas, graphical analysis suggests no effect on language around the cutoff, but that is inconsistent with point estimates from all of the parametric analyses regardless of functional form and bandwidth. For West Virginia, graphical analysis and analysis

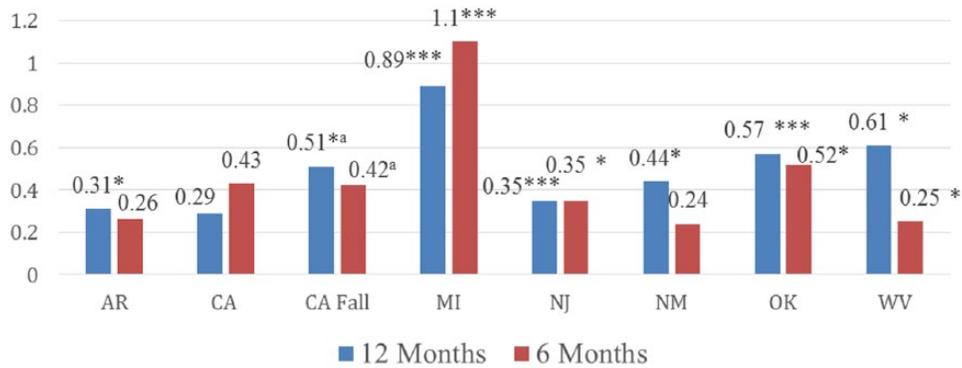


FIGURE 3. Estimated effects sizes for math (Woodcock-Johnson Tests of Achievement Applied Problems), by bandwidth using instrumental variable estimates.

a. California fall data are restricted to children assessed in the fall (September to December).

* $p < .05$. ** $p < .01$. *** $p < .001$.

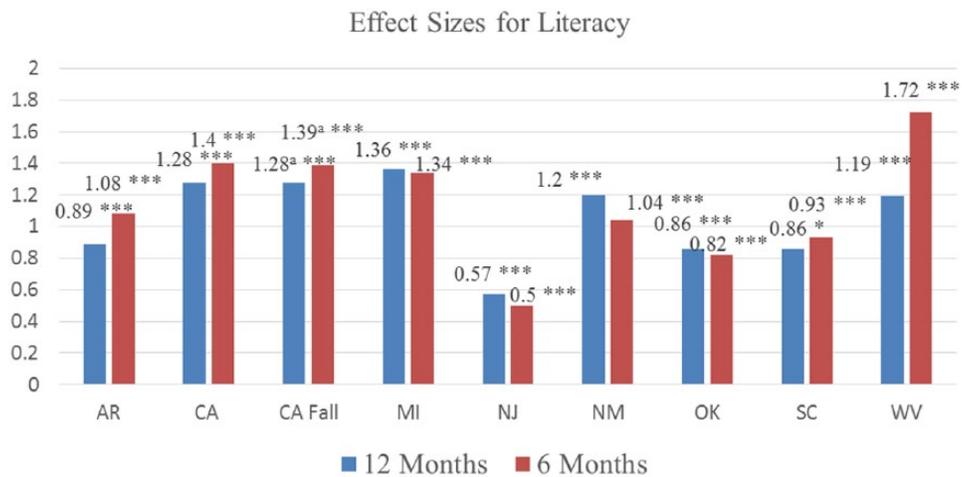


FIGURE 4. Effects on literacy by bandwidth using instrumental variable estimates. Literacy was measured using the Preschool Comprehensive Test of Phonological and Print Processing except in West Virginia, where the Test of Preschool Early Literacy was used, and Michigan, where the Woodcock-Johnson Tests of Achievement Letter-Word Identification test was used.

a. California fall months data are restricted to children assessed in the fall (September to December).

* $p < .05$. ** $p < .01$. *** $p < .001$.

with a 3-month bandwidth suggest a smaller estimate than the other analyses.

Mathematics

The estimated average effect of state-funded preK on math was 0.44 standard deviations but dropped to 0.34 standard deviations excluding California, Michigan, and New Mexico. A math test was not administered in South Carolina. Michigan had a significantly-larger-than-average effect, whereas New Mexico and West Virginia had significantly-below-average effects. Michigan's exceptionally large estimate could reflect sampling limitations. All estimated effects on math were statistically significant in the 12-month analyses. Graphical analyses and parametric analyses were qualitatively consistent for all states.

Literacy

All of the estimated effects on children's emergent literacy (primarily, print awareness) were large and statistically significant for even the 6-month bandwidth, averaging greater than 1.0 standard deviation. New Jersey's program effect is significantly smaller than average. California, Michigan, and West Virginia had significantly-larger-than-average effects, but Michigan used a different measure, and West Virginia used a different version of the assessment used elsewhere. The average effect size for literacy effects was statistically significantly higher than for language and math. Graphical analyses and parametric analyses were consistent. If anything, graphical analyses suggested ceiling problems for the treatment group that might bias the estimated impacts downward.

Discussion

We estimated the effects of eight state preK programs serving 4-year-olds on measures of children's early learning in language, literacy, and mathematics using an age-cutoff RD approach. On average, programs were found to produce broad gains in children's learning at kindergarten entry for all outcomes, but there were significant departures from the average effects. Overall, these results were qualitatively robust with respect to potential threats to the validity of our analytical approach, including nonlinearities, the bandwidth around the age cutoff, and other issues.

Effect sizes varied by outcome domain and by state within outcome domain. Estimated effects on emergent literacy were almost uniformly large, often 1 standard deviation or more, the size of the achievement gaps by race and income (Friedman-Krauss et al., 2016). Estimated effects on math were moderate. Estimated effects on language were smallest and were less than 0.10 standard deviations for some states. Estimates for language were particularly sensitive to differences in analytical methods, and we have strongest confidence that language effects were positive for Michigan, New Jersey, and Oklahoma.

Although our results are consistent with past findings, the pattern of effects is troubling. Estimated effects at kindergarten entry were significantly larger for simple literacy skills than for the other two outcomes. The literacy estimates could be biased upward by parental efforts to prepare children for kindergarten, but this pattern across outcomes has been observed in studies not subject to that threat (Manship et al., 2015; Puma et al., 2010). Gains in unconstrained knowledge and skills (like language and math) may provide a stronger basis for sustained long-term gains in achievement compared with gains in such constrained skills as letter knowledge (Bailey et al., 2017). Language has been found to be particularly predictive of literacy success beyond Grade 3 (Snow & Matthews, 2016).

The pattern of findings suggests that state preK programs may need to increase attention to unconstrained skills. Estimated effects on language in our study of large-scale public programs are much smaller than for the well-known small-scale programs found to have long-term academic effects. For some states, these estimated effects are quite small, similar to the national Head Start Impact Study that found few lasting effects on achievement. This suggests that state preK programs should ensure that supports for learning and teaching, including their curriculum and professional development, deepen and enrich preschool education. Devoting particular attention to language could be promising, without neglecting other domains, such as social and emotional development, that we did not measure. Additionally, language can be considered a measure of conceptual knowledge, which argues for a content-rich curricular approach.

Explaining variation in effect sizes among the states is complicated by the number of dimensions on which state preK can vary. Program features, population characteristics, and the counterfactual condition vary among states. Each state's effects depend not only on the contributions of their state-funded preK program to learning and development but also on the care and education arrangements and other supports available to the comparison group that vary by state (Barnett & Friedman-Krauss, 2016; Blau & Currie, 2006). Possibly because of the large number of potential explanations, we find no clear explanations for the sources of differences in state preK outcomes in the current study.

New Jersey's preK was the only one of the eight programs in which the vast majority of children entered at age 3 (2 years rather than 1 year before kindergarten). The New Jersey estimates essentially represent the impact of 1 year of preK at age 4, contingent on participation at age 3. Estimated effects on language and math were roughly average, but the estimated effect on literacy was significantly below average. Possibly this difference reflects a ceiling issue, but such programs may need to carefully attend to building on the literacy skills of 4-year-olds who have already attended a year of preK.

Extensive sensitivity analyses were conducted recognizing the methodology's limitations, key assumptions, and recent challenges to typical implementation of the approach. Our results were fairly robust across different analyses, but graphical analyses did not always support the same conclusions as parametric analyses. Also, results sometimes varied by functional form or bandwidth. Most conflicts occurred for language outcomes for which estimated effects were modest at best. The best remedies for this problem are likely to be larger sample sizes and focusing data collection on children close to the birthdate cutoff to the extent possible.

Recently identified limitations of the age-cutoff RD deserve serious attention but do not seem to warrant abandoning the approach. Some estimated effects were essentially zero, whereas uniformly positive results would be expected if positive findings were primarily due to design flaws. However, caution should be exercised in comparing our estimates with those from other methodological approaches. The counterfactual condition may differ from that in studies using same age comparison groups, and we found that some of the other limitations of our approach mattered. Ceiling effects for literacy at age 5 could bias estimates downward. Also, some estimates became more consistent across methods when we adjusted for differential attrition between treatment and comparison groups. Improved measures, procedures that reduce attrition or equate groups for attrition, and more detailed data on family background, home learning, and classroom experiences for both groups would improve the validity of estimates.

Our findings also have broader implications for state preK monitoring and evaluation. First, state preK program

effectiveness cannot simply be assumed but should be measured regularly. Estimates of important outcomes are near zero for some states. Second, states should not rely on assessment of narrow literacy skills alone as an indicator of preK's broader effectiveness. Every state now has comprehensive standards for preschool learning and development (Barnett et al., 2017). Monitoring and evaluation of preK program effectiveness should be correspondingly broad, to the extent possible, including language, mathematics, and other measures predictive of long-term achievement and school success. The broader literature would support assessment of executive function and social and emotional development, as well (Phillips et al., 2017).

Other methodologies will continue to be important for evaluation of state preK programs. RCTs offer advantages in interpretation and statistical power (Bloom, 2005). However, RCTs have their own limitations, including the potential to create compensatory rivalry and practical obstacles to obtaining a control group for universal programs (McCambridge, Kypri, & Elbourne, 2014). The preschool education experiences of the control group may not represent what would happen if state preK did not exist, increasing demand for other services. RCTs and other strong alternatives to the age-cutoff RD are needed for longitudinal follow-up, as this RD cannot be used for longitudinal studies. In all of these studies, more attention should be paid to carefully measuring both the treatment and the counterfactual to provide a better understanding of exactly what produces effects.

Overall, our study adds to the evidence that public preK can improve learning and development for both disadvantaged and general populations, at least in the short term, but it also raises concerns about variability in effectiveness across outcome domains and states. Although these results cannot be extrapolated to all state preK programs, they are based on a diverse sample of state preK programs over a fairly long time. We suggest that states devote increased attention to frequent, more rigorous, and broader evaluations of their preK programs following our specific recommendations above. We hope this study will be viewed as providing one set of evaluation results to which others will add using similar, but improved, methods.

References

- Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*, 7–39.
- Barnett, W. S., & Frede, E. C. (2017). Long-term effects of a system of high-quality universal preschool education in the United States. In H.-P. Blossfeld, N. Kulic, J. Skopek, & M. Triventi (Eds.), *Childcare, early education and social inequality: An international perspective* (pp. 152–172). Cheltenham, UK: Edward Elgar.
- Barnett, W. S., & Friedman-Krauss, A. H. (2016). *State(s) of Head Start*. New Brunswick, NJ: National Institute for Early Education Research.
- Barnett, W. S., Friedman-Krauss, A. H., Weisenfeld, G. G., Horowitz, M., Kasmin, R., & Squires, J. H. (2017). *The state of preschool 2016: State preschool yearbook*. New Brunswick, NJ: 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, pp. 43–59). Beverly Hills, CA: Sage.
- Blau, D., & Currie, J. (2006). Pre-school, day care, and after-school care: Who's minding the kids? In E. Hanushek, & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 1163–1277). New York, NY: North-Holland.
- Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. New York, NY: Russell Sage Foundation.
- Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*, 43–82.
- Buckles, K. S., & Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics, 95*, 711–724.
- Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust non-parametric confidence intervals for regression-discontinuity designs. *Econometrica, 82*, 2295–2326.
- Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*, 579–620.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*, 596–610.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724–750.
- Duncan, G. J., & Magnuson, K. A. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*, 109–132.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.
- Dunn, L., Lugo, D., Padilla, E., & Dunn, L. (1986). *Test de Vocabulario en Imágenes Peabody*. Circle Pines, MN: American Guidance Service.
- Friedman-Krauss, A., Barnett, W. S., & Nores, M. (2016). *How much can high-quality universal pre-K reduce achievement gaps?* Washington, DC: Center for American Progress/New Brunswick, NJ: National Institute on Early Education Research. Retrieved from <https://cdn.americanprogress.org/wp-content/uploads/2016/04/01115656/NIEERAchievementGaps-report.pdf>
- Gilliam, W. S., & Ripple, C. H. (2004). What can be learned from state-funded prekindergarten initiatives? A data-based approach to the Head Start devolution debate. In E. Zigler, & S. Styfco (Eds.), *The Head Start debates* (pp. 477–497). Baltimore, MD: Brookes.
- Gilliam, W. S., & Zigler, E. F. (2001). A critical meta-analysis of all impact evaluations of all state-funded preschool from

- 1977 to 1998: Implications for policy, service delivery, and program evaluation. *Early Childhood Research Quarterly*, 15, 441–473.
- Gormley, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*, 320, 1723–1724.
- Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, 184, 1265–1269.
- Hill, C. J., Gormley, W. T., Jr., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly*, 32, 60–79.
- Hustedt, J., Jung, K., Barnett, W. S., & Williams, T. (2015). Kindergarten readiness impacts of the Arkansas Better Chance state prekindergarten initiative. *Elementary School Journal*, 116, 198–216.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48, 281–355.
- Lipsey, M. W., Farran, D. C., Bilbrey, C., Hofer, K. G., & Dong, N. (2011). *Initial results of the Tennessee Voluntary Pre-K program*. Nashville, TN: Vanderbilt University, Peabody Research Institute.
- Lipsey, M. W., Farran, D. C., & Hofer, K. G. (2015). *A randomized control trial of a statewide voluntary prekindergarten program on children's skills and behaviors through third grade*. Nashville, TN: Vanderbilt University, Peabody Research Institute.
- Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten Program: End of pre-K results from the randomized control design*. Nashville, TN: Vanderbilt University, Peabody Research Institute.
- Lipsey, M., Weiland, C., Yoshikawa, H., Wilson, S., & Hofer, K. (2015). Estimating preschool effects using the age cutoff regression-discontinuity design. *Educational Evaluation and Policy Analysis*, 37, 296–313.
- Lonigan, C. J., Wagner, R., Torgesen, J., & Rashotte, C. (2007). *Test of Preschool Early Literacy*. Austin, TX: PRO-ED.
- Lonigan, C., Wagner, R., Torgeson, J., & Rashotte, C. (2002). *Preschool Comprehensive Test of Phonological and Print Processing*. Tallahassee: Department of Psychology, Florida State University.
- Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does pre-kindergarten improve school preparation and performance? *Economics of Education Review*, 26, 33–51.
- Manship, K., Quick, H., Holod, A., Mills, N., Ogut, B., Chernoff, J., . . . González, R. (2015). *Impact of California's transitional kindergarten program, 2013–14*. San Mateo, CA: American Institutes for Research. Retrieved from <http://www.air.org/sites/default/files/downloads/report/Impact-of-Californias-Transitional-Kindergarten-Program-Dec-15.pdf>.
- McCambridge, J., Kypri, K., & Elbourne, D. (2014). In randomization we trust? There are overlooked problems in experimenting with people in behavioral intervention trials. *Journal of Clinical Epidemiology*, 67, 247–253.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142, 698–714.
- Peisner-Feinberg, E. S., & Schaaf, J. M. (2011). *Effects of the North Carolina More at Four prekindergarten program on children's school readiness skills*. Chapel Hill: University of North Carolina, FPG Child Development Institute.
- Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., . . . Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects, a consensus statement*. Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2017/04/consensus-statement_final.pdf
- Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., . . . Ciarico, J. (2010). *Head Start Impact Study: Final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children*, 26, 57–74.
- Snyder, T. D., de Brey, C., & Dillow, S. A. (2016). *Table 207.10: Number of 3- to 5-year-olds not yet enrolled in kindergarten and percentage participating in home literacy activities with a family member, by type and frequency of activity and selected child and family characteristics: 2001, 2007, and 2012. Digest of Education Statistics 2014 (NCES 2016-006)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Strategic Research Group. (2008). *Assessing the effectiveness of Tennessee's pre-kindergarten program: Annual report 2008–2009*. Columbus, OH: Strategic Research Group. Retrieved from <http://www.comptroller.tn.gov/repository/re/srgannualreport08-09.pdf>
- van der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, 22, 219–245.
- Weiland, C., McCoy, D. C., Grace, E., & Park, S. O. (2017). Natural window of opportunity? Low-income parents' responses to their children's impending kindergarten entry. *AERA Open*, 3, 1–15.
- Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84, 2112–2130.
- Wong, V., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, 27, 122–154.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement*. Itasca, IL: Riverside.
- Woodcock, R. W., & Muñoz, A. F. (1990). *Bateria Woodcock-Muñoz Pruebas de Aprovechamiento-Revisados*. Itasca, IL: Riverside.
- Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, 26, 21–35.

Authors

W. STEVEN BARNETT is a senior codirector of the National Institute for Early Education Research, Graduate School of Education, Rutgers University. His research interests include early childhood care and education policy and the economics of education.

KWANGHEE JUNG is an associate director for data management and statistics at the National Institute for Early Education Research. Her research interests include the effects of childcare and early education on children's learning and development, and family risk factors and academic achievement of young children in ethnic-minority families.

ALLISON FRIEDMAN-KRAUSS is an assistant research professor at the National Institute for Early Education Research. Her research interests include unpacking impacts of early education interventions, early education quality, and the cognitive and social-emotional development of low-income children.

ELLEN C. FREDE is a senior codirector of the National Institute for Early Education Research. Her research interests include child and classroom assessment, curriculum and professional development systems, and early learning and development particularly for dual language learners.

MILAGROS NORES is a codirector for research at the National Institute for Early Education Research. Her research interests include early childhood development, data-driven policy development, evaluation design, economics, cultural diversity, and English language learning.

JASON T. HUSTEDT is an assistant professor in the Department of Human Development and Family Studies at the University of Delaware. His research focuses on the impacts of state-funded pre-kindergarten initiatives on young children, federal and state early childhood policy, and preschoolers' and toddlers' interactions with their parents and peers.

CAROLLEE HOWES is a professor of education at University of California, Los Angeles. Her research interests include social development, children's experiences in childcare and other preschool settings, the concurrent and long-term outcomes from those experiences, and efforts to improve the quality of young children's experiences.

MARIJATA DANIEL-ECHOLS is the director of the Center for Health Equity Practice at Michigan Public Health Institute. Her research interests include using the intersection of research, policy, and practice to counter systemic inequalities and promote social justice.