CHILD DEVELOPMENT

Child Development, November/December 2013, Volume 84, Number 6, Pages 2112–2130

Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills

Christina Weiland and Hirokazu Yoshikawa
Harvard Graduate School of Education

Publicly funded prekindergarten programs have achieved small-to-large impacts on children's cognitive outcomes. The current study examined the impact of a prekindergarten program that implemented a coaching system and consistent literacy, language, and mathematics curricula on these and other nontargeted, essential components of school readiness, such as executive functioning. Participants included 2,018 four and five-year-old children. Findings indicated that the program had moderate-to-large impacts on children's language, literacy, numeracy and mathematics skills, and small impacts on children's executive functioning and a measure of emotion recognition. Some impacts were considerably larger for some subgroups. For urban public school districts, results inform important programmatic decisions. For policy makers, results confirm that prekindergarten programs can improve educationally vital outcomes for children in meaningful, important ways.

High-quality early childhood education equips children with the cognitive skills required for success in elementary school and beyond. Studies show that intensive preschool interventions can be highly cost effective and have positive impacts into adulthood (Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Reynolds, Temple, White, Ou, & Robertson, 2011). From a developmental science perspective, this makes much sense; children's cognitive skills are malleable at a young age, and thus supporting their early development builds a strong foundation for later educational and intellectual success. Children with higher levels of early vocabulary, reading, mathematics, and executive functioning consistently have greater levels of academic success in elementary and middle school (Duncan et al., 2007; McClelland, Acock, & Morrison, 2006; National Early Literacy Panel, 2008). While the evidence is more mixed for emotional outcomes, both developmental theory and some empirical evidence suggest similar links to later academic outcomes for that domain (Entwisle, Alexander, & Olson, 2005; Pianta & Stuhlman, 2004).

Such findings have helped motivate the recent expansion of state- and locally funded prekindergarten programs in the United States. As of 2010, 40 states had implemented prekindergarten programs, enrolling 27% of the nation's 4-year-olds (Barnett et al., 2010). Evaluations of these programs with the strongest research design to date (regression discontinuity) have confirmed that children enrolled in these programs have higher language, literacy, and mathematics outcomes, on average, at scale (Gormley, Gayer, Phillips, & Dawson, 2005; Gormley, Phillips, & Gayer, 2008; Hustedt, Barnett, Jung, & Goetze, 2009; Hustedt, Barnett, Jung, & Thomas, 2007; Wong, Cook, Barnett, & Jung, 2008). Findings on impacts of public prekindergarten on children's socioemotional skills come from two quasi-experimental (and nonregression discontinuity) studies and findings were mixed (Gormley, Phillips, Newmark, Perper, & Adelstein, 2011; Magnuson, Ruhm, & Waldfogel, 2007).

While overall these results are encouraging, research suggests that many preschool programs struggle to attain good instructional quality (Burchinal, Kainz, & Cai, 2011; Peisner-Feinberg & Burchinal, 1997). Accordingly, there have been many efforts to increase preschool quality, including interventions

This study is funded by the Institute of Education Sciences. Thanks to the Boston Public Schools; Jason Sachs; the BPS Department of Early Childhood; participating coaches, principals, teachers, and children; John Willett; Richard Murnane; Nonie Lesaux; John Papay; and members of the Harvard RD Methodology in Prekindergarten Studies Working Group (particularly Howard Bloom, Jens Ludwig, Doug Miller, Guido Imbens, and Thomas Lemieux). Special thanks to our research assistants Kjersti Ulvestad, Carla Schultz, Julia Hayden, Michael Hurwitz, Hadas Eidelman, Kam Sripada, Ellen Fink, Julia Foodman, Deni Peri, Caitlin Over, and John Goodson.

Correspondence concerning this article should be addressed to Christina Weiland, Harvard Graduate School of Education, 14 Appian Way, Room 704, Cambridge, MA 02139. Electronic mail may be sent to Christina_weiland@mail.harvard.edu.

 0009-3920/2013/8406-0022
DOI: 10.1111/cdev.12099

that use curricula, teacher professional development, or both as quality supports. Many such interventions have shown efficacy when implemented on a small scale or in research demonstration trials. When such interventions are taken to scale, it is widely recognized that achieving positive impacts is more challenging. The intervention's creators, for example, cannot be as heavily involved, and maintaining quality of implementation is more difficult (Shadish, Cook, & Campbell, 2002).

This study, which used data on approximately 2,000 students enrolled in the Boston Public Schools (BPS) public prekindergarten program, represents an intersection of the literature on the effects of public prekindergarten programs and the literature on quality-support interventions in preschool. Regarding the former, as in the strongest of the prekindergarten studies, we used a quasi-experimental regression-discontinuity (RD) design, with the birthday cutoff for entry into the program providing exogenous treatment eligibility, to estimate the effects of public prekindergarten on children's developmental outcomes. Relevant to the quality-support literature, the BPS program combined two features that are prominent in the literature on preschool quality improvement: research-based (mathematics, language, and literacy) curricula, paired with a coaching system for preschool teachers. Curricula were chosen by the district and implemented at scale without involvement of the curriculum developers. The coaching system was developed by the district. Conditions accordingly represent those more typically encountered in public school districts than in research demonstration trials. Although we were not able to identify causally which of these inputs—curricula, coaching, or simply attending prekindergarten—constituted the most "active" ingredients in the intervention, we are nonetheless able to provide domain-specific and policy-relevant information regarding the pedagogical conditions under which impacts were achieved.

Within this context, we examined impacts of the BPS program on children's language, literacy, mathematics, and emotional development, domains that were directly targeted by the district-chosen curricula. One of our mathematics assessments is new to the literature and addresses some of the content limitations of more commonly used preschool mathematics assessments. We also present impacts on executive function (EF) skills, a developmentally important component of school readiness (Blair & Razza, 2007). EF was not directly targeted by the intervention, but theory and empirical work suggest that there may be spillover effects of cognitively focused curricula on this domain. In addition, we collected detailed data on the care type experienced by control-group children. Thus, we were able to specify what the program is being compared to, which is crucial given that the counterfactual for early childhood program attendance has changed substantially since landmark studies of preschool implemented in the 1960s and 1970s (Campbell et al., 2002; Schweinhart, Barnett, & Belfield, 2005). We also tested for statistically significant differences in impacts by gender, free or reduced lunch status, and race or ethnicity. The previous literature suggests that the effects of preschool may differ by these demographic characteristics. Finally, we present evidence that our results are robust to a set of threats to internal validity. Many of these sensitivity analyses—such as robustness of estimates to attrition from and late entry into the prekindergarten program, different start rules by age on certain measures, differences in reactivity to the testing situation in the treatment and control groups, and use of extant data to aid in the interpretation of produced estimates, as only children who took up an offered seat were tested—are new to the RD prekindergarten literature. Carefully examining these threats is important for advancing research methodology in future evaluations.

Short-Term Effects of Prekindergarten

Many previous studies have summarized the literature on the effects of preschool programs on children's developmental outcomes in great detail (Barnett, 1995; Currie, 2001; Gormley et al., 2005; Wong et al., 2008; Yoshikawa, 1995). In brief, prekindergarten appears to have positive, small-to-large effects on children's cognitive development and small effects on children's prosocial and problem behaviors, although the direction of the latter differs by study.

Focusing specifically on the public prekindergarten studies that share this study's research design (RD), researchers have found statistically significant positive impacts on children's mathematics scores in five of seven examined contexts (one city and six states; effect size range = 0.16–0.50) and on children's receptive vocabulary scores in four of eight examined contexts (one city and seven states; effect size range = 0.17–0.36). On assessments not shared across this body of studies, there was evidence of moderate-to-large effects on children's early literacy skills in six of eight examined contexts (Gormley et al., 2005; Gormley et al., 2008; Hustedt et al., 2007; Hustedt et al., 2009; Wong et al., 2008). In addition, in studies of the Tulsa program (the only program in this body of literature to date for which subgroup impacts have been reported), Hispanic children and children raised in poverty, who generally have poorer outcomes than their White and higher income peers, appeared to enjoy greater benefits from enrollment in prekindergarten (Gormley et al., 2005, 2008).

Socioemotional and executive functioning outcomes have not been examined to date in the set of RD studies of the immediate impacts of prekindergarten. However, a recent study that used propensity score methodology found that public prekindergarten produced small reductions in children's timidity and increases in attentiveness (Gormley et al., 2011). A quasi-experimental study found that public prekindergarten increased children's aggression and decreased their self control (Magnuson et al., 2007). However, there were no statistically significant socioemotional effects for children who attended prekindergarten and kindergarten in the same public school.

Curricula and Coaching in Prekindergarten Settings

Curricula. Theory suggests that implementing explicit, intentional curricula in preschool programs may be effective for several reasons. Such curricula may ensure a continuing emphasis on the skills necessary for children's early school success, may help keep children engaged and challenged in the classroom, and may also help maintain classroom quality (Klein & Knitzer, 2006). Empirical evidence supports the effectiveness of some language, literacy, mathematics, EF, and socioemotional curricula on directly targeted child developmental domains (Barnett et al., 2008; Bierman et al., 2008; Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Domitrovich, Cortes, & Greenberg, 2007; Fischel et al., 2007). Effective curricula in prekindergarten may also improve children's outcomes in nontargeted domains. For example, a reading and behavior management curriculum improved children's EF skills (Bierman, Nix, Greenberg, Blair, & Domitrovich, 2008), and a mathematics-focused curriculum improved children's oral language and literacy skills (Sarama, Lange, Clements, & Wolfe, in press).

Similarly, EF may be impacted by exposing children to prekindergarten curricula that have an explicit cognitive focus. There are hypothesized to be three distinct but related components of EF—working memory, inhibitory control, and attention shifting (Blair & Razza, 2007). Each is associated with language and math skills among preschool-aged children (Blair & Razza, 2007; Bull & Scerif, 2001; Diamond, Carlson, & Beck, 2005; Gathercole & Pickering, 2000). From a Vygotskian perspective, improved language may support children's EF skills by enhancing children's outer and then inner speech, which in turn may then improve EFs as children become better able to plan and monitor their behavior (Vygotsky, 1978). Furthermore, early mathematics, language, and literacy tasks all make demands on children's working memory, cognitive flexibility, and inhibitory control (Welsh, Nix, Blair, Bierman, & Nelson, 2010). There is uncertainty about the causal direction of the relation between EF and these cognitive skills, but it is plausible that implementing effective cognitively focused curricula in preschool could improve EF.

Coaching. Coaching is an ongoing professional development model in which an expert (the coach) models instruction, observes teachers' practice, and provides teachers with constructive feedback on their pedagogy (Neuman & Cunningham, 2009). Coaching may or may not involve supporting teachers' implementation of specific curricula (Aikens & Akers, 2011). Coaching can produce gains in preschool classroom quality, teacher instructional practices, and children's cognitive and behavioral development (Aikens & Akers, 2011; Bierman et al., 2008; Neuman & Cunningham, 2009; Raver et al., 2009). Thirteen of 14 studies have found that coaching improves preschool teachers' curriculum implementation (see Aikens & Akers, 2011). Monthly coaching was also part of the professional development model in a randomized controlled trial of Building Blocks, the mathematics curriculum implemented in the current study (Clements & Sarama, 2008). These researchers found large gains in children's mathematics skills at the end of prekindergarten, as well as high levels of curricular fidelity and higher quality mathematics instruction in treatment classrooms.

Subgroup Effects

Do the effects of preschool education differ by sociodemographic factors, such as socioeconomic status, race or ethnicity, or child gender? Large-scale preschool education in the United States emerged from the desire to reduce gaps between the academic performance of children from poor versus better-off homes (Zigler & Styfco, 2010). Nearly all of the literature evaluating the impacts of preschool education on children is based on low-income populations (the median percentage of families in poverty in rigorous preschool evaluations identified in a recent meta-analysis was 91%; Leak et al., 2012). There are some hints in the studies conducted on national data sets that the effects of preschool and center-based care on cognitive outcomes are stronger for lower income families (Brooks-Gunn, Gross, Kraemer, Spiker, & Shapiro, 1992; Currie, 2001). In recent years, there has also been strong interest in whether preschool education might reduce related gaps in cognitive performance by race or ethnicity (Magnuson & Waldfogel, 2005). The national Head Start Impact Study found significantly stronger positive effects of the program on a range of Latino children's developmental outcomes, compared to those of other racial or ethnic groups, in its follow-up to first grade (U.S. Department of Health & Human Services, 2010). The Tulsa prekindergarten study found particularly strong cognitive effects among Latino children (Gormley et al., 2005). Gender has also been of interest as a moderator of preschool impacts. A recent study pooling the Perry, Abecedarian, and Early Training Project data found stronger benefits for girls than boys (Anderson, 2008). However, a meta-analytic study covering a broader range of preschool evaluations did not find this pattern (Kelchen et al., 2012).

In our sample, a substantial proportion of families were not low income, due to the public prekindergarten system not being means tested. We therefore have an opportunity in this study to examine whether the effects of public prekindergarten differ by family socioeconomic background, as well as by race or ethnicity and gender.

In the current study, we address two research questions:

1 What is the impact of the prekindergarten program on children's early mathematics, language, literacy, EF, and emotional development?
2 Do some child subgroups (as defined by family income, race or ethnicity, or child gender) benefit statistically significantly more from the prekindergarten program than others?

Method

Intervention

Setting. In 2008–2009, the BPS 4-year-old prekindergarten program served approximately 2,045 children in 69 elementary schools. Any child within the city of Boston who turned 4 by September 1 could apply for the program; unlike many public prekindergarten programs in other districts and states (Barnett et al., 2010), children's access was not limited by their family income or other restrictions. There is no perfect metric to determine how many of the city's 4-year-olds are enrolled in the BPS prekindergarten program. One metric relies on the U.S. Census's 2010 estimate of the percentage of children under age 5 in Boston (U.S. Census Bureau, 2012). Based on those numbers, about 34% of the city's 4-year-olds were enrolled in the BPS prekindergarten program in 2008–2009. A second estimate is based on the number of children who ultimately enroll in BPS kindergarten. In 2009–2010, among children enrolled in kindergarten, 43% of those children had attended prekindergarten in BPS in the previous year (excluding those in special education-only classrooms, as these children would have been served by the district even in the absence of the prekindergarten program due to federal requirements).

Treatment condition. Children who attended the program in the treatment year (2008–2009) received a year of free full-day prekindergarten in an urban public school setting. The evaluation year was the 2nd year of full implementation of the literacy and language curriculum Opening the World of Learning (OWL; Schickedanz & Dickinson, 2005) and the mathematics curriculum Building Blocks (Clements & Sarama, 2007a). The theory of change in BPS was that implementing explicit, intentional, and uniform curricula across classrooms with professional development supports would improve and maintain the quality of support provided to teachers and optimize resource allocation (e.g., through the streamlining of teacher training; Sachs & Weiland, 2010). In a fidelity study conducted the year treatment children were enrolled in prekindergarten, coaches trained on fidelity measures for each curriculum reported that they were implemented with moderately high fidelity (Weiland, Eidelman, & Yoshikawa, 2012).

Curricula background and implementation. The OWL curriculum targets children's early language and literacy skills and includes a social-skills component embedded in each unit, in which teachers discuss socioemotional issues with children and integrate emotion-related vocabulary words. The Building Blocks curriculum targets early mathematics skills, particularly (a) number and simple arithmetic and (b) geometry, measurement, and spatial sense. Three mathematical themes—patterns, data, and sorting and sequencing—are woven into these two main areas. In addition, many activities are intentionally child directed, with children making up their own problems or creating their own geometric designs (Clements & Sarama, 2007a). Its pedagogical approach has a heavy focus on language, as it requires children to explain their mathematical reasoning verbally. Neither curriculum targets children's EF skills directly.

OWL and Building Blocks have shown positive effects on children's outcomes in other studies (Ashe, Reed, Dickinson, Morse, & Wilson, 2009; Clements & Sarama, 2007b; Clements et al., 2011). However, the evidence base for Building Blocks is stronger than that for OWL. Children in eight programs that implemented OWL showed consistently positive effects in studies that used pre–post designs with no control group (Wilson, Morse, & Dickinson, 2009). However, a recent randomized controlled trial in Head Start centers (Dickinson, Freiberg, & Barnes, 2011; Dickinson et al., 2011) found no impacts of OWL on children's language and literacy outcomes at the end of preschool, and some negative effects at the end of kindergarten and the end of first grade. However, these latter results are somewhat difficult to interpret, as the fidelity of implementation in the treatment groups was relatively low and control classrooms had partially implemented OWL. Teachers were also on average better educated in the eight programs that showed positive effects than in the RCT (65% vs. 17% with a bachelor's degree [BA], respectively).

Teacher qualifications and professional development supports. All BPS prekindergarten teachers are subject to the same educational requirements and pay scale as K-12 teachers. All prekindergarten teachers must have at least a BA and must obtain a masters degree within 5 years. Placing BPS within the national context, in 2010, 27 of 40 states required a BA for teachers in state-funded prekindergarten programs (Barnett et al., 2010). During the treatment year, 78% of program teachers held masters degrees and 75% had at least 5 years of teaching experience. Prekindergarten teachers received a variety of supports in the year prior to our evaluation and in the evaluation year itself, including curriculum-specific training and weekly to biweekly on-site support from an experienced early childhood coach trained in both curricula. In the 1st year of implementation, teachers were offered 2 days of curricular training in Building Blocks and 5 days in OWL. During the school year, teachers were offered 4 days of training in Building Blocks and 2 days of training in OWL. In the 2nd year of implementation, all teachers new to the prekindergarten program were offered 5 days of curricular training before the start of the school year and 6 days of training during the school year. For more on teacher background characteristics, see online supporting information Appendix S4, Table S1.

Coaching sessions were tailored to address the individual needs of each teacher in implementing the curricula and managing the classroom. All early childhood coaches held masters degrees. On average, early childhood coaches had themselves taught previously in early childhood classrooms for 8.8 years (range = 2–20 years, SD = 4.9 years) and had worked as a district early childhood coach an average of 3.3 years (range = 0.5–7 years, SD = 2.2 years).

Sample

In fall 2009, children in the BPS prekindergarten program and all children who attended the program in the previous year were eligible for the study. Children in special-education-only classrooms were excluded due to concerns about the appropriateness of the assessment battery for children who were not mainstreamed. For a child to participate in the study, the principal, classroom teacher, and parent (or guardian) of the child all had to consent to participate. In fall 2009, all eligible principals and teachers were invited to participate. Of 79 elementary schools with eligible children, 12 principals declined to participate (15%). Approximately 93% of eligible teachers in participating schools agreed to participate in child-level data collection in fall 2009 (N = 250 out 270), an average of 3.7 teachers per participating school. Participating schools and teachers were representative of district schools and teachers (see online supporting information Appendix S3).

We translated parent consent forms into five languages and forwarded them to the child's home up to three times. Within participating classrooms in the 67 participating schools, 69% of 2,938 eligible children returned consent forms, for a total sample size of 2,018. This represents 54% of eligible children enrolled in the district in fall 2009. Compared to nonparticipants on 14 characteristics, study participants were more likely to live in the east attendance zone (44% vs. 35%; one of three attendance zones; $p < .001$), less likely to live in the north attendance zone (28% vs. 35%; $p < .001$), more likely to have special needs (9% vs. 6%; $p < .01$), more likely to be White (18% vs. 15%; $p < .01$), more likely to be Asian (11% vs. 9%; $p < .05$), and less likely to be Hispanic (41% vs. 46%; $p < .01$). Participating children were nested in 238 classrooms (the difference between this figure and the 250 consented teachers is due to 7 classrooms having two teachers and 5 teachers agreeing to participate, but with very few students eligible for the study and none who ultimately returned consent forms). The number of participating children per classroom ranged from 1 to 22 (average of 8.5, $SD = 5.2$).

The final sample of 2,018 is racially, linguistically, and socioeconomically diverse. Forty-one percent of the children were Hispanic, 26% were Black, 18% were White, 11% were Asian, and 3% were of mixed, or other, race. Fifty percent of the sample spoke only English, 28% spoke Spanish, and 22% spoke a language other than English or Spanish. Sixteen languages were represented in the "other" category; within this category, the most commonly spoken languages were Vietnamese (30%), Haitian (12%), and Cape Verdean Creole (8%). Approximately 69% of sampled children were eligible for free or reduced lunch.

Child Assessment Procedures

Children were tested by study-trained child assessors. These assessors had to establish target reliability on the full battery of tests and show good rapport and child management skills in both simulated and real testing situations. All assessors were college educated and approximately one third held masters degrees. On average, the complete battery of nine tests took 45–50 min to administer. Children were tested in a single session if possible, with the session divided into smaller segments if the child showed signs of fatigue. We randomized the order of tests to limit the possibility of biasing results systematically due to child fatigue. The assessors visited classrooms in fall 2009, as close to the start of the school year as teacher and school schedules and study staffing would allow. Assessors were first allowed into schools 2 weeks after the start of school (end of September). Approximately 33% of the data were collected by the end of October, 88% collected by the end of November, and 98% collected by the end of December. Children were assessed in English.

Outcome Measures

Receptive vocabulary. Children's receptive vocabulary was measured using the Peabody Picture Vocabulary Test III (PPVT–III; Dunn & Dunn, 1997), a nationally normed measure that has been used widely in diverse samples of young children (U.S. Department of Health and Human Services, 2010). The test has excellent split-half and test–retest reliability estimates, as well as strong qualitative and quantitative validity properties (Dunn & Dunn, 1997). It requires children to choose (verbally or nonverbally) which of four pictures best represents a stimulus word. In our analysis, as in other prekindergarten RD studies (Hustedt et al., 2007; Hustedt et al., 2009; Wong et al., 2008), we used the raw score total as our outcome measure.

Prereading and reading skills. The Woodcock–Johnson Letter-Word Identification subscale (Woodcock, McGrew, & Mather, 2001) is a nationally normed, widely used measure (Gormley et al., 2005; Peisner-Feinberg et al., 2001). Children are asked to identify and pronounce isolated letters and entire words fluently. According to the developers, the estimated test–retest reliability of the Letter-Word subscale for 2- to 7-year-olds is 0.96. Consistent with other prekindergarten RD studies (Gormley et al., 2005; Gormley et al., 2008), we used the raw score total as an outcome in our analysis.

Numeracy and early math. The Woodcock–Johnson Applied Problems subscale (Woodcock et al., 2001) is a numeracy and early mathematics measure that requires children to perform relatively simple calculations to analyze and solve arithmetic problems. Its estimated test–retest reliability for 2- to 7-year-old children is 0.90 (Woodcock et al., 2001) and it has been used widely with diverse populations of young children (Gormley et al., 2005; Peisner-Feinberg et al., 2001; Wong et al., 2008). In our analysis, as in other prekindergarten RD studies (Gormley et al., 2005; Gormley et al., 2008; Hustedt et al., 2007; Hustedt et al., 2009; Wong et al., 2008), we used the raw score total as an outcome.

The Applied Problems subtest does not measure geometric and spatial capacities and researchers have raised some concerns regarding the test's comprehensiveness, appropriateness, and sensitivity in use with young children (Clements, Sarama, & Liu, 2008). Therefore, we also assessed children's mathematics skills using a subset of 19 items from the Research-Based Elementary Mathematics Assessment (REMA; Clements et al., 2008), as this measure assesses a wider range of early numeracy, geometry, and spatial skills. We used Rasch modeling and other psychometric analysis to assess the shortened REMA's psychometric properties and confirmed that it was a valid measure of children's early mathematics skills (Weiland et al., 2012). In all analyses, we used the Rasch-estimated child ability scores as the outcome.

EF skills. Our battery of tests included assessments that tapped three principal dimensions of EF:

working memory, cognitive inhibitory control, and attention shifting. Forward Digit Span and Backward Digit Span (FDS and BDS, respectively; Gathercole & Pickering, 2000; Wechsler, 1986) tapped different components of working memory. BDS measures the central executive component, while FDS measures phonological loop. In both tasks, the assessor reads aloud a string of numbers to the test child, with approximately a 1-s pause between digits. The child then either has to repeat back exactly what the assessor said (in FDS) or reverse the string of numbers (in BDS). Before items are administered, the child must pass a practice trial, demonstrating that he or she understands the directions of the task. FDS is scored from 1 to 6, while BDS is scored from 1 to 5. The score represents the child's digit span memory (i.e., a 2 represents a digit span memory of two digits).

For attention shifting, we used the Dimensional Change Card Sort (DCCS) and a subset of items from the Task Orientation Questionnaire (TOQ; Smith-Donald, Raver, Hayes, & Richardson, 2007). In the DCCS (Frye, Zelazo, & Palfai, 1995), children were shown target cards that differed along dimensions of color and shape (e.g., red and blue, rabbits and boats). Children learned to sort the cards according to one dimension (shape or color) and then were asked to sort the cards on the other dimension. After practice trials to confirm that children understood the rules, the assessor administered up to 10 trials on the DCCS. After 6 trials, if a child had missed more than 1 trial, the testing was discontinued. If the child had missed only 1 or 0 trials, the assessor continued until Trial 10. The final DCCS total score was the number of trials (out of 10) in which the child managed to shift attention from the prior criterion and sort the cards according to the new criterion correctly.

The full TOQ assesses the child's emotional state and capacity to sustain focus on a set of tasks during a testing session. After administering the child assessment battery, assessors rated each child on 13 items reflecting his or her capacity to sustain attention to the tasks, demonstrate self-regulation, and engage actively to achieve a goal. Each item was rated on a 4-point scale, with clear behavioral descriptors provided for each point on the scale. Using the full sample of children, we conducted a confirmatory factor analysis on the full set of TOQ items and confirmed the presence of three distinct constructs—positive emotion, attention shifting, and impulse control. The fit of the factor model was good (comparative fit index [CFI] = .976, root mean square error of approximation [RMSEA] = .058, standardized root mean square residual [SRMR] = .048). The four items that measured attention shifting included "Pays attention to instructions and demonstration," "Careful, interested in accuracy," "Sustains concentration—willing to try repetitive tasks," and "Cooperates, complies with tester's requests." In our analyses, we used a unit-weighted average of responses to these four items as our attention-shifting outcome.

To assess children's cognitive inhibitory control, we used Pencil Tapping (Diamond & Taylor, 1996). The child was asked to tap twice if the evaluator tapped once and tap once if the evaluator tapped twice. Assessors first administered a set of practice trials to ensure that children understood the rules of the task. Children who passed the practice were then administered 16 total trials. The task measures children's cognitive inhibitory control and, to a lesser degree, working memory and fine motor activity (Bierman, Nix, et al., 2008). Scores recorded the correct number of trials out of 16 that children achieved. Because of concern that tapping a pencil could prove difficult for preschoolers and might conflate cognitive inhibitory control with fine motor skills, we substituted larger plastic kitchen spoons for pencils in this task.

Emotional development. Our chosen emotional development outcomes are all derived from either direct testing or assessor ratings of children. Commonly used measures of children's behavior in preschool often rely on parent and teacher reports. However, parents and teachers may have different expectations of children based on whether they are entering preschool versus kindergarten, a problem discussed in Gormley et al.'s (2011) evaluation of the Tulsa prekindergarten program's impacts on children's socioemotional outcomes. Because our RD design compares preschool children with kindergarten children across an age cutoff, intervention effects on outcomes measured by parent and teacher reports could have been confounded with differences in reporters' expectations based on the child's age.

We used three measures of emotional development: the Emotion Recognition Questionnaire (ERQ; Ribordy, Camras, Stefani, & Spaccarelli, 1988), TOQ Positive Emotion, and TOQ Impulse Control (Smith-Donald et al., 2007). The ERQ assesses children's ability to identify emotions. In the ERQ, children listened to 16 stories that described characters in different situations and were shown a picture corresponding to the situation. They were then asked to identify the character's feeling by pointing to pictures of happy, mad, sad, or scared faces. The faces shown matched the gender of the child (i.e., boys were shown boy faces and girls were shown girl faces). Children received 2 points for identifying the correct emotion, 1 point if they misidentified the emotion but identified the valence correctly, and 0 points if they identified neither emotion nor valence correctly, for a maximum score of 32. Before administering the test, the assessor first established that the child could identify the happy, mad, sad, or scared faces correctly. The ERQ has been used with children in Head Start and has demonstrated sensitivity to intervention effects (Bierman et al., 2008).

The confirmatory factor analysis described previously on the TOQ identified three items for positive emotion: "alert and interactive; is not withdrawn," "shows pleasure in accomplishment and active task mastery," and "confident"; and three items for impulse control: "can wait during and between tasks," "remains in seat appropriately during test," and "modulates and regulates arousal level in self." In our analyses, scores on our Positive Emotion and Impulse Control outcomes were unit-weighted averages of children's responses to the position emotion and impulse control factors, respectively.

Predictors

Forcing variable. Using district administrative records, we constructed a continuous predictor to measure how many days from the cutoff the child's birthdate fell, centered on September 1. This predictor was the "forcing variable" in our RD analysis—the clear cutpoint that is the exogenous determinant of children's eligibility for treatment (Lee & Lemieux, 2010). Positive integer values indicated that the child was born before September 1 and negative, after. A value of 0 indicated that the child was born on September 1.

Treatment variable. We also created a dichotomous variable that recorded whether children were in the treatment group (set equal to 1, when centered child age was 0 or greater) or the control group (set equal to 0, when centered child age was less than 0).

Covariates and Descriptive Characteristics

Administrative data. From district administrative records, we obtained information on children's race or ethnicity, home language, free and reduced lunch status, gender, and special needs status. We used a vector of dichotomous indicators to represent child race or ethnicity, each coded 1 when the child was from the particular racial or ethnic group, 0 otherwise. Racial or ethnic groups were Asian, Black, Hispanic, Other, and White. Similarly, we used a vector of dichotomous indicators to represent children's home language (English, Spanish, or Other), each coded 1 when the requisite language was the child's home language, 0 otherwise. We also constructed dichotomous indicators to represent child free and reduced lunch status, gender, and special needs status, each coded 1 if the child fell into a demographic category and 0 otherwise. These covariates have been shown to predict children's early cognitive and educational outcomes in other studies, and there is a consensus in the early childhood education literature that these should be controlled in impact analyses (Clements et al., 2011; Wong et al., 2008).

Preprogram Child-Care Types

We were also able to obtain parent-reported information on the primary type of child care that children experienced before entering the 4-year-old district prekindergarten program. When registering their children for prekindergarten, parents were asked about the child's last child-care experience, including the name of the provider, and were asked to choose one from the following care types: Head Start, private preschool, public preschool, licensed family day care, family day care, and other or none. Because parents often disagreed about program type for the same program name, we cleaned and recoded these data extensively, confirming the type for each named program so that codes are consistent across children. We verified the program type via extensive web searches and through lists of programs and types obtained from the Massachusetts Department of Early Education and Care, the Boston Early Education Quality Improvement Project, and the National Association for the Education of Young Children. Information was often unavailable regarding whether a family day-care provider was licensed and parents frequently disagreed regarding the same provider's licensing status. Thus, we collapsed licensed family day care and family day care into one category in our analysis. Other or none almost always refers to relative care, such as parental care or care by an immediate relative.

Data Analytic Strategy

Impacts: Basic framework. For the impact estimates, we capitalized on the exogenous variation in program receipt created by the use of the district's age cutoff to determine children's entry into the program. The RD approach is useful when there is a clear cutpoint on a "forcing variable," such as child age, that is the exogenous determinant of children's eligibility for treatment. On one side of the cutoff, participants are assigned to a particular treatment, whereas on the other side of the cutoff, they are not (Imbens & Lemieux, 2008; Shadish et al., 2002; Thistlethwaite & Campbell, 1960; Trochim, 1984). In our case, children must have turned 4 years old on or before September 1, 2008 to attend the prekindergarten program (the treatment) in the 2008–2009 school year (Year 1). Any differences in average school-readiness outcomes in fall 2009 (the beginning of the 2009–2010 school year, or Year 2) between children who fell just to one side, or the other, of the cutoff, provided unbiased estimates of the causal impact of the program for children of this age. Under the standard RD design, we capitalize on the data of children remote from the birthday cutoff to estimate the treatment effect for those target children whose birthdays fell in the immediate vicinity of September 1, on one side or the other. As is common in RD studies, our results only generalize to students right at the cutoff.

Interpretation of the impact estimates. A standard application of the RD methodology, provided all assumptions are met, provides an unbiased estimate of the average effect of assignment to the treatment condition (vs. control) for participants immediately on either side of the cutoff (Bloom, 2012; Murnane & Willett, 2010). This estimate is known as the intent-to-treat (ITT) estimate as it summarizes the average difference between participants who were assigned to the treatment and control conditions, whether they end up taking up their assigned place in either the treatment or the control group. In our study, however, the only children tested are those who actually showed up in the schools at the point of testing (fall 2009). As such, the treatment estimate is not a classic ITT estimate. It also does not meet the definition of a treatment-on-the-treated (TOT) estimate, or the effect of the intervention on those who actually took up the treatment, as TOT estimates are derived from ITT estimates (Angrist & Pischke, 2008). As such, estimates produced by our study and by previous prekindergarten RD with age cutoff studies are neither pure ITT nor pure TOT estimates. Previous such studies have left this problem unresolved (Gormley et al., 2005; Wong et al., 2008).

We took several steps to address this problem (for details concerning our strategies and results, see online supporting information Appendix S1). In brief, we contend that our RD estimates are definitionally ITT estimates with potential selection bias. However, simulations and analysis using administrative data suggest that the magnitude of our estimates is closer to TOT than ITT. As such, we interpret them as representing effects for those who enrolled in the program. Later in this article, and more fully in the online supporting information Appendices S1, S2, and S4, we provide evidence that detected effects are robust to a multitude of sensitivity analyses.

Adjusting for attrition and late enrollment. To adjust for children who were missing outcome data due to attrition or late enrollment, we used propensity score weighting. Using administrative records from enrollment applications, we identified students who participated in the prekindergarten program in Year 1 but attrited from the district by time of testing (Year 2; $N = 209$). We also identified control-group children who were not included in our tested sample because they either attrited before testing ($N = 63$) or enrolled after the testing period ($N = 54$). Previous such studies have not accounted for these additional groups of children. Adjusting for them is key, given that they technically should be included in our analysis of those who took up the program. Because we had administrative data for these attriter and late-entry children, we were able to adjust for observed differences between our child assessment (impact) sample of 2,018 and the larger sample including them. Illustrating the importance of this adjustment, in Table S2 in online supporting information Appendix S4, we present descriptive statistics on the demographics of both the tested sample and the attriter and late-entry sample. As shown in the table, there are statistically significant differences between the two samples on 6 of 14 examined demographic characteristics.

To conduct the required adjustments, our propensity score model was as follows:

$$\mathrm{PS}_{ijk} = \Pr(\text{child\_tested} = 1 | \sum \mathrm{X}_{ijk}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \mathrm{X}_{ijk})}} \tag{1}$$

where PS is the probability that the *i*th student, in the *j*th classroom of the *k*th school would be tested, conditional on X, a vector of student-level covariates (race or ethnicity, gender, home zone, language, and siblings). We fitted this model, obtained predicted values of these propensity that a child would be tested, and then inverted these propensities to obtain an inverse probability weight (IPW) that we could use in our subsequent RD analysis to counteract selection into testing (Imbens & Wooldridge, 2009; Murnane & Willett, 2010). Conceptually, our IPW approach upweights children whose entry into the tested or untested condition was not predicted well by the selection model in Equation 1 and for whom we then assume that the endogenous contribution of self-selection plays less of a role in the determination of the RD estimate.

RD impact approach. We incorporated the IP weights into our RD analyses using weighted least squares regression, in the sample of tested children who did possess values on the empirical outcomes. Our impact equation was as follows:

$$\begin{aligned}\text{OUTCOME}_{ijk} = \beta_0 + \beta_1\text{TREAT}_{ijk} + \beta_2\text{CAGE}_{ijk} \\ + \beta_3\text{TREAT}_{ijk} * \text{CAGE}_{ijk} + \beta_4\text{Y}_k + \varepsilon_{ijk}\end{aligned} \quad (2)$$

where OUTCOME is a generic representation of the child-level test score, TREAT is a dichotomous indicator of treatment or control-group status, CAGE is the child's age centered on the September 1 cutoff, Y is a vector of school fixed effects, and ε is a student-level error term. We estimated robust standard errors to account for the clustering of children within classrooms. We did not include student demographics in Equation 2, as they had already been accounted for through the IPW.

Our analytical strategy and robustness checks for our RD analyses were informed by Lee and Lemieux (2010) and *What Works Clearinghouse* guidelines (Schochet et al., 2010). We first conducted a graphical analysis, displaying and smoothing the relation between the outcome child age on either side of the cutoff, by superimposing a fitted linear regression line and a smoothed, locally weighted nonparametric regression line on a scatter plot of the raw data. These empirical plots suggested the functional form of the outcome and forcing variable relation and revealed whether there was indeed a discontinuity in the average value of the outcome between the groups assigned to the treatment and control conditions, at the cutoff. Second, because specifying the correct functional form of the relation between outcome and the forcing variable is one of the chief challenges in RD analysis (Imbens & Lemieux, 2008; Ludwig & Miller, 2007), when we specified a linear relation between the two variables, we did so within a window, or bandwidth, on either side of the age cutoff, within which one might reasonably argue that the functional form of the outcome and forcing variable relation was "locally" linear. This approach is a flexible method that allows for the inclusion of covariates, and gives equal weight to all observations that fall into a local bandwidth (Imbens & Lemieux, 2008). This approach also has better boundary properties than other standard nonparametric smoothing strategies (Hahn, Todd, & Van der Klaauw, 2001). A nearly identical version of the method was used to estimate successfully the impacts of Head Start on child mortality rates and educational attainment, in another RD-designed evaluation (Ludwig & Miller, 2007).

Third, as a check on the specification of our local linear regression models, we also fitted a series of additional models in which we replaced the linear specification of the outcome and forcing variable relation with polynomial specifications and interaction terms of the necessary order between the treatment and forcing variables. We compared fit statistics across models and overspecified the models as a robustness check. Although less efficient than when models are underspecified, overspecification yields less biased estimates (Trochim, 1984) and has been used as a strategy in other early childhood RD designs (Gormley et al., 2005; Wong et al., 2008).

As a fourth step, we examined the sensitivity of our results to choice of bandwidth (Lee & Lemieux, 2010). Within selected bandwidths, we reestimated the IP weights from Equation 1, using the sample of observations corresponding to that bandwidth. To provide easy comparisons with other RD prekindergarten studies (Gormley et al., 2005; Wong et al., 2008), we adopted a bandwidth of 6 months on either side of the age cutoff and fitted our different specifications of the RD model (Equation 2) to data within this window. We also employed the cross-validation procedure of Lee and Lemieux (2010) and Imbens and Lemieux (2008) to estimate an "optimal" bandwidth, by minimizing the mean squared error of prediction at the cutoff. Within each bandwidth choice, we repeated the modeling steps outlined above and obtained additional estimates of the treatment effects.

Subgroup analysis. We extended our basic approach to estimate treatment effects for selected subgroups. The subgroups of interest included those defined by race or ethnicity (Black, Latino, White, and Asian), free and reduced lunch status, and gender. Due to the paucity of data for the Other race or ethnicity group, we did not fit models that included this subgroup. Our primary model for estimating these subgroup effects was as follows:

$$\begin{aligned}\text{OUTCOME}_{ijk} = {} & \beta_0 + \beta_1\text{TREAT}_{ijk} + \beta_2\text{CAGE}_{ijk} \\ & + \beta_3\text{TREAT}_{ijk} * \text{CAGE}_{ijk} \\ & + \beta_4\text{SUBGROUP}_{ijk} \\ & + \beta_5\text{TREAT}_{ijk} * \text{SUBGROUP}_{ijk} \\ & + \beta_6\text{SUBGROUP}_{ijk} * \text{CAGE}_{ijk} \\ & + \beta_7\text{TREAT}_{ijk} * \text{SUBGROUP}_{ijk} \\ & * \text{CAGE}_{ijk} + \beta_8 Y_k + \varepsilon_{ij},\end{aligned} \tag{3}$$

where ε is a student-level error term. In this model, we represent the different sets of subgroups with a generic predictor, SUBGROUP. The predictors whose associated slope parameters represent the treatment effects for the different subgroups are as follows: (a) the dichotomous predictor SUBGROUP, indicating membership in a subgroup of interest; (b) the interaction term TREAT*SUBGROUP; (c) the interaction term SUBGROUP*CAGE; and (d) the three-way interaction term SUBGROUP*CAGE *TREAT. We also tested whether it was necessary to include higher order quadratic and cubic terms, adding in the necessary higher order terms for SUBGROUP*CAGE and TREAT*SUBGROUP*CAGE. In each analysis, we included IPW as previously explained to adjust for children who were not tested because of attrition or late enrollment. Equation 3, like Equation 2, does not include a vector of other student characteristics, as they were accounted for through the IPW. Also, for a given subgroup model, the IPW does not include the subgroup characteristic of interest. This is because including the subgroup in the weight prohibits us from including a fixed effect for the subgroup of interest (it would "double count" the subgroup effect). We reported here only those subgroup effects that are robust across bandwidth (see Figures 1 and 2). Results including all statistically significant subgroup effects across all bandwidths are available upon request.

In fitting all our regression models, we used the method of multiple imputation (with 50 imputations) to account for missing data, following Graham (2009). In Table 1, we present summary statistics on the child outcomes, including the percent missing for each outcome.

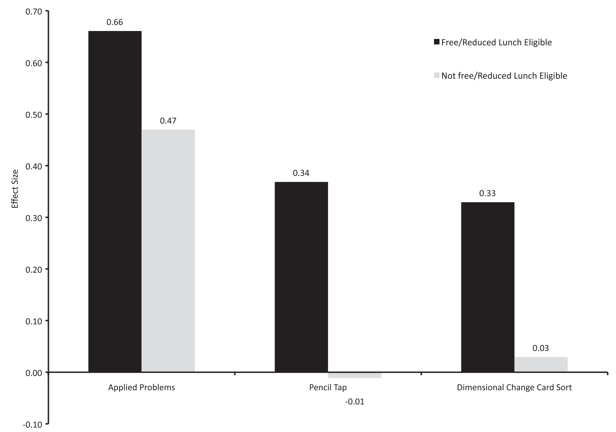

Figure 1. Estimated effect sizes of the prekindergarten program on selected outcomes, by children's free or reduced lunch status. Effect sizes were estimated from fitted regression-discontinuity models within a bandwidth of 365 days on either side of the age cutoff and with a linear relation specified between the achievement outcomes and age.

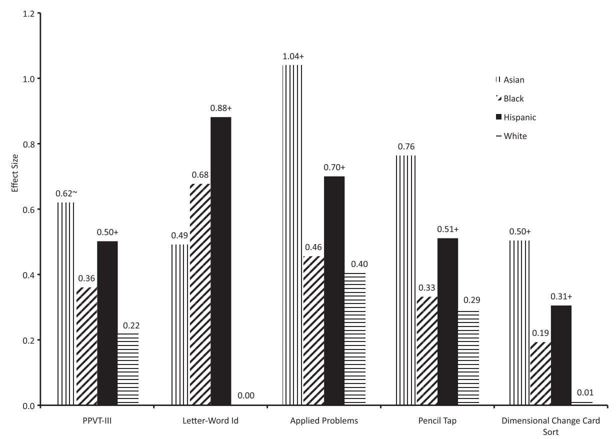

Figure 2. Estimated effect sizes of the prekindergarten program on selected outcomes, by children's race or ethnicity. + denotes that the effect for the racial or ethnic group was larger than that for White children ($p < .05$), and the effect was robust to bandwidth and functional form. ~ denotes that the effect for a racial or ethnic group was larger than that for Whites with a bandwidth of 365 days but that the effect was not robust to bandwidth and functional form. Effect sizes were estimated from fitted regression-discontinuity models within a bandwidth of 365 days on either side of the age cutoff and with a linear relation specified between the achievement outcomes and age. Subgroup effects were estimated from statistically significant interactions between race or ethnicity and treatment status ($p < .05$). Other statistical interactions between race or ethnicity, distance from the age cutoff, and treatment status were included as needed.

Results

Descriptive Statistics on Control-Group Care Types

Parents of children in the control group reported the following care types in the year in which their children were too young to enter the BPS program: Head Start (16%), public centers (12%), private centers (29%), nonrelative home-based care (10%), and relative care (33%). Two thirds of control children thus experienced some kind of nonrelative care in 2008–2009 and 57%, center care or preschool.

Main Impacts

Participation in the prekindergarten program led to statistically significant improvements in mathematics, literacy, and language skills (Table 2). Effect sizes were as follows: 0.45 for receptive vocabulary (PPVT), 0.62 for early reading (Letter–Word Identification), 0.58 for numeracy (Applied Problems), and 0.49 for numeracy and geometry (REMA Short). We also found statistically significant, positive impacts on most measures of EF and on one measure of emotional development (Tables 3 and 4). Effect sizes were 0.23 for working memory (both FDS and BDS), 0.20 for inhibitory control (Pencil Tap), 0.27 for attention shifting (DCCS), and 0.18 for emotion recognition (Emotion Recognition Questionnaire). Results for outcomes from the TOQ—attention shifting, positive emotion, and impulse control—were positive in sign but were not statistically significant. Effect sizes were very similar in models with and without the IPW correction for attrition and late entry (online supporting information Appendix S4, Table S3).

Subgroup Impacts

We also found that some subgroups of children benefited more from the program than did others. For instance, children who were eligible for free or reduced lunch benefited statistically significantly more than those who were ineligible on numeracy (Applied Problems), inhibitory control (Pencil Tap), and attention shifting (DCCS; see Figure 1). For numeracy, effect sizes for both groups were in the moderate-to-large range (0.66 and 0.47, respectively). For inhibitory control and attention shifting, the benefits of the treatment accrued nearly entirely to the children who were free or reduced lunch eligible, with a very small or zero effect at the cutoff for the children who were not free or reduced lunch eligible. For all other outcomes, impacts did not vary by free- and reduced lunch status.

In Figure 2, we display our estimates of effect size by children's race or ethnicity. Impacts were statistically significantly larger for Hispanic children than for White children on 8 of 12 assessments. These differential effects were robust to sensitivity analyses for five assessments: PPVT, Letter-Word ID, Applied Problems, Pencil Tap, and DCCS outcomes (measures across nearly the full range of domains assessed). Effects for Asian children were statistically significantly larger than those for White children on 8 of 12 assessments, but the estimated differences were robust to sensitivity analyses for only the Applied Problems and DCCS outcomes, in part due to the small size of the Asian sample. Effects for Black children were statistically significantly larger than those for White children on 3 of 12 assessments, but these differences were not robust to sensitivity analysis. All outcomes for which there were statistically significant race or ethnicity effects that were robust across bandwidth and functional form also passed general linear hypothesis (GLH) tests. That is, we found that the joint effect of the relevant subgroup characteristics multiplied by the treatment variable was not zero (e.g., F statistic $p < .10$). The exception was the

Table 1
Sample Means (Standard Deviations) for Selected Child Outcomes (N = 2,018)

	Full sample	Born before cutoff; attended prekindergarten in 2008–2009	Born after cutoff; attended prekindergarten in 2009–2010	% missing total
PPVT–III	58.26 (21.84)	69.16 (17.65)	48.08 (20.44)	5.40
W–J Letter-Word Id	12.44 (7.18)	15.99 (7.03)	9.18 (5.59)	3.87
W–J Applied Problems	13.74 (5.30)	16.54 (4.35)	11.16 (4.75)	3.87
REMA Short Form	−0.08 (1.31)	0.62 (1.12)	−0.73 (1.13)	4.36
Pencil Tap	10.77 (6.00)	12.94 (4.56)	8.69 (6.47)	6.94
Dimension Change Card Sort	6.64 (4.26)	8.01 (3.46)	5.37 (4.54)	4.61
Backward Digit Span	1.53 (0.79)	1.78 (0.87)	1.29 (0.62)	9.56
Forward Digit Span	4.15 (1.28)	4.46 (1.18)	3.86 (1.31)	5.60
TOQ Attention	3.47 (0.66)	3.61 (0.57)	3.34 (0.71)	5.15
TOQ Positive Emotion	3.24 (0.56)	3.34 (0.52)	3.15 (0.59)	5.20
TOQ Impulse Control	3.62 (0.61)	3.70 (0.56)	3.54 (0.64)	5.05
Emotion Recognition Questionnaire	25.80 (5.08)	27.52 (3.24)	24.20 (5.90)	5.70

Note. PPVT = Peabody Picture Vocabulary Test; W–J Letter-Word Id = Woodcock–Johnson Letter-Word Identification; W–J Applied Problems = Woodcock–Johnson Applied Problems; REMA = Research-Based Early Mathematics Assessment; TOQ = Task Orientation Questionnaire.

Table 2
Estimated Treatment Impact (Standard Errors) on Language, Literacy, and Mathematics Outcomes, for Samples of Children Within Selected Bandwidths Around the Age Cutoff on the Forcing Variable

	PPVT–III		W–J Letter-Word ID		W–J Applied Problems		Research-Based Early Mathematics Assessment Short Form		
BW (in days)	**365 +**	180	**365 +**	180	**365**	180	**365**	180	111 +
Treatment	**9.00*** (1.81)**	7.85** (2.60)	**3.45*** (0.55)**	2.61** (0.78)	**2.81*** (0.46)**	2.59*** (0.62)	**0.57*** (0.12)**	0.49** (0.15)	0.37* (0.19)
Effect size	**0.44**	0.38	**0.62**	0.47	**0.59**	0.55	**0.50**	0.43	0.33
Functional form of hypothesized outcome and child-age relation	**Linear**	Linear	**Linear + int.**	Linear	**Linear**	Linear	**Linear**	Linear + int.	Linear + int.
N	**2,018**	969	**2,018**	969	**2,018**	969	**2,018**	969	627

Note. All fitted regression models include the fixed effects of schools and standard errors are corrected for the clustering of children within classrooms. For all outcomes, we fitted regression models using only samples of observations that fell within 365 and 180 days of the cutoff. We also fit models in samples of children that fell within the optimal bandwidth (BW) determined via the cross-validation procedure (+ denotes the optimal bandwidth). For outcomes where the optimal bandwidth was 365 or 180 days, we fitted two models. Within each analysis, we modeled the outcome as a linear, quadratic, and cubic function of the forcing variable, and we also fit models that included interactions between the child-age variable and the treatment indicator. Preferred models are listed in bold. Effect sizes are expressed in terms of the standard deviation of the control group. PPVT = Peabody Picture Vocabulary Test; W–J Letter-Word Id = Woodcock–Johnson Letter-Word Identification; W–J Applied Problems = Woodcock–Johnson Applied Problems.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3
Estimated Treatment Impact (Standard Errors) on Executive Functioning Outcomes, for Samples of Children Within Selected Bandwidths Around the Age Cutoff on the Forcing Variable

	Pencil Tap			Backward Digit Span			Forward Digit Span		Dimensional Change Card Sort			TOQ Attention		
BW (in days)	**365**	180	287+	**365**	180	221+	**365**	180+	**365**	180	300+	**365**	180	147+
Treatment	**1.39***	1.33†	1.49**	**0.15***	0.16†	0.19*	**0.31****	0.46*	**1.25****	1.34*	1.21**	**0.08**	0.05	0.06
	(0.54)	(0.79)	(0.57)	**(0.07)**	(0.10)	(0.08)	**(0.12)**	(0.18)	**(0.40)**	(0.54)	(0.43)	**(0.07)**	(0.09)	(0.10)
Effect size	**0.21**	0.21	0.23	**0.24**	0.26	0.31	**0.24**	0.35	**0.28**	0.30	0.27	**0.11**	0.07	0.08
Functional form of hypothesized outcome and child-age relation	**Linear + int.**	Linear	Linear + int.	**Linear**	Linear	Linear	**Linear**	Linear	**Linear**	Linear	Linear	**Linear**	Linear	Linear
N	**2,018**	969	1,439	**2,018**	969	1,199	**2,018**	969	**2,018**	969	1,610	**2,018**	969	799

Note. All fitted regression models include the fixed effects of schools and standard errors are corrected for the clustering of children within classrooms. For all outcomes, we fitted regression models using only samples of observations that fell within 365 and 180 days of the cutoff. We also fit models in samples of children that fell within the optimal bandwidth (BW) determined via the cross-validation procedure (+ denotes the optimal bandwidth). For outcomes where the optimal bandwidth was 365 or 180 days, we fitted two models. Within each analysis, we modeled the outcome as a linear, quadratic, and cubic function of the forcing variable and we also fit models that included interactions between the child-age variable and the treatment indicator. Preferred models are listed in bold. Effect sizes are expressed in terms of the standard deviation of the control group. TOQ = Task Orientation Questionnaire.
†$p < .10$. *$p < .05$. **$p < .01$.

Table 4
Estimated Treatment Impact (Standard Errors) on Emotional Development Outcomes, for Samples of Children Within Selected Bandwidths Around the Age Cutoff on the Forcing Variable

	Emotion Recognition Questionnaire			TOQ Positive Emotion			TOQ Impulse Control		
BW (in days)	**365**	180	293+	**365**	180	332+	365	180	**129+**
Treatment	**1.12* (0.50)**	$1.22^{\dagger}$ (0.70)	0.84 (0.58)	**0.02 (0.05)**	0.01 (0.07)	0.08 (0.06)	0.05 (0.11)	0.09 (0.08)	**0.13 (0.09)**
Effect size	**0.19**	0.21	0.14	**0.03**	0.02	0.01	0.07	0.14	**0.20**
Functional form of hypothesized outcome and child-age relation	**Linear**	Linear	Linear	**Linear + int.**	Linear + int.	Linear + int.	Cubic + int.	Linear	**Linear**
N	**2,018**	969	1,582	**2,018**	969	1,795	2,018	969	**724**

Note. All fitted regression models include the fixed effects of schools and standard errors are corrected for the clustering of children within classrooms. For all outcomes, we fitted regression models using only samples of observations that fell within 365 and 180 days of the cutoff. We also fit models in samples of children that fell within the optimal bandwidth (BW) determined via the cross-validation procedure (+ denotes the optimal bandwidth). For outcomes where the optimal bandwidth was to 365 or 180 days, we fitted two models. Within each analysis, we modeled the outcome as a linear, quadratic, and cubic function the forcing variable and we also fit models that included interactions between the child-age variable and the treatment indicator. Preferred models are listed in bold. Effect sizes are expressed in terms of the standard deviation of the control group. TOQ = Task Orientation Questionnaire.
$^{\dagger}p < .10$. $^{*}p < .05$.

effect of Letter-Word Id for Hispanics: In a GLH test, we could not reject the null hypothesis that the joint effect of the interactions between the race or ethnicity variables and the treatment indicator was zero, $F(3) = 1.86$, $p = .14$. We found no differences in impacts of the program by gender.

Robustness Checks

We followed best practices as described in the RD literature and conducted extensive sensitivity analyses to confirm the robustness of our findings (Imbens & Lemieux, 2008; Lee & Lemieux, 2010). Threats to the internal validity of our results included: (1) treatment misallocation at the cutoff; (2) nonsmooth or discontinuous variation in observed and unobserved student characteristics around the cutoff; (3) discontinuities in the outcomes at points other than the cutoff; (4) incorrect specification of the functional form of the relation between outcome and forcing variable; (5) sensitivity of results to the choice of bandwidth around the age cutoff; (6) inflated estimates of treatment effect due to treatment-group children being more familiar with, and comfortable in, testing situations than control-group children; (7) the accumulation of Type I error as a result of multiple tests being conducted; (8) sensitivity of results to use of different start rules on the PPVT–III; and (9) sensitivity of results due to use of raw scores rather than IRT-based *W* scores on the Woodcock–Johnson Letter-Word Identification and Applied Problems subscales. Threats 1 to 5 and Threats 8 and 9 could result in either an over- or underestimation of the true impact of the treatment, whereas Threat 6 could lead to an overestimate of the true impact and Threat 7 could lead to an overstatement of the statistical significance of our findings. We examined each of these threats in turn and found no evidence that suggested any threats to the internal validity of our identifying assumptions (see online supporting information Appendix S2 for details).

Discussion

We found that a prekindergarten program that combined evidence-based curricula with trained BA- and masters-level teachers and coaching support produced positive effects on multiple domains of school readiness. We detected substantial and statistically significant effects of the prekindergarten program on educational outcomes both in domains that were targeted directly by the prekindergarten curriculum—literacy, language, mathematics, and emotional development—and in a related but nontargeted domain (EF).

Language, literacy, and mathematics impacts were in the moderate-to-large range (effect sizes 0.45–0.62), whereas EF impacts were in the small range (0.20–0.27). From a developmental perspective, the small positive impacts on children's EF dimensions—working memory, inhibitory control, and attention shifting—are particularly interesting. Small impacts on EF are consistent with the "spillover" hypothesis described earlier in this article; that is, mathematics, language, and literacy curricula that are cognitively focused may also improve other cognitive developmental domains like EF, even without directly targeting them. For example, evidence suggests that mathematics skills such as number composition and decomposition are quite closely related to working memory (Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007). Furthermore, preschool numeracy and geometry activities make demands on children's ability to shift attention appropriately among problem elements, and to inhibit automatic or prepotent responding to only one aspect of a given problem (Welsh et al., 2010). Language skills such as expressive and receptive vocabulary are associated with better performance on inhibitory control and attention shifting among young children (Fuhs & Day, 2011). The curricula implemented in Boston aimed to enhance these particular mathematics, language, and literacy skills and therefore may have led to simultaneous impacts on EF dimensions. The possible mathematics-EF spillover is particularly promising, given that the optimal approach for promoting EF skills in prekindergarten is unknown and given that early mathematics skills are a robust predictor of later academic achievement in both math and reading (Duncan et al., 2007).

Although we cannot pinpoint specific active ingredients that led to detected effects, we believe the combination of curricula and coaching, implemented with majority masters-level teachers, likely played a major role. The OWL and Building Blocks curricula have shown promising results to date in other studies (Ashe et al., 2009; Clements & Sarama, 2007b; Clements et al., 2011) and we found that teachers implemented them moderately well. Furthermore, it is possible that implementing both a mathematics curriculum and a language and literacy curriculum created a synergistic effect, as both evidence and theory suggest that stronger literacy and language skills can support children's learning of mathematics skills, and vice versa (Duncan et al., 2007; Harrison, McLeod, Berthelsen, & Walker, 2009; Wagner, Venezky, & Street, 1999).

The mix of children from lower and higher income families in the BPS prekindergarten program may also have contributed to the detected impacts. Boston and Tulsa are the only public prekindergarten contexts examined to date in which applications were not restricted by family income requirements, and both achieved particularly strong results. Among older students, having higher achieving peers from higher income families can affect individual children's achievement, particularly for lower ability students or those from poorer backgrounds (Zimmer & Toma, 2000). The positive effects of having higher ability peers also occur among preschoolers (Henry & Rickman, 2007). Across the 40 states with prekindergarten programs, only 8 did not have requirements prioritizing lower income families (Barnett et al., 2010).

The counterfactual care options in Boston are worth considering as a potential alternative explanation of detected effects. Strong results in Boston could have been a function of lower quality alternative care in the control group. Approximately two thirds of control-group children were enrolled in nonrelative care and nearly half were enrolled in center care, proportions that roughly mirror national trends (Haskins & Barnett, 2010). Making this alternative explanation unlikely, relative to other states, child-care regulations in Massachusetts are among the most stringent in the nation (National Association of Child Care Resource & Referral Agencies, 2011).

In terms of subgroups, we found that impacts on most outcome measures were not statistically significantly different when comparing children from more affluent versus less affluent households. Likewise, focusing on results that were robust to bandwidth and functional form, effects for Hispanic and Asian children were not statistically significantly higher than those of White children for the majority of outcomes. Our findings run counter to some studies that suggest that the positive benefits of preschool accrue mostly or entirely to poorer and minority children (see Currie, 2001). As in the Tulsa prekindergarten program (Gormley et al., 2005), more affluent and White children also benefited from the BPS prekindergarten program.

Nonetheless, findings for Hispanic children versus their White peers should be highlighted, as we found the largest number of statistically significant effects for Hispanics (5 of 12 measured, encompassing all examined cognitive domains). A limitation of our study is that children were tested in English only. However, our findings align with those from the Head Start Impact Study (U.S.

Department of Health and Human Services, 2010) and from the Tulsa prekindergarten evaluation (Gormley et al., 2005), which also found larger impacts on cognitive outcomes for Hispanic children. Evidence suggests that Hispanic children may be particularly likely to benefit from high-quality, supportive instructional contexts (Han, 2008). Furthermore, the rates of growth of children from lower income Spanish-speaking homes can surpass that of native-born children in both word reading and oral language skills (Mancilla-Martinez & Lesaux, 2011). Nationally, Hispanic children are underrepresented in preschool programs and their enrollment rates in recent years have even declined (Fuller & Kim, 2011). In Boston, among Hispanic children entering regular education kindergarten in fall 2009, 39% had experienced the BPS prekindergarten in the previous year, compared to 42% of Blacks, 51% of Whites, and 58% of Asians. Policy-level efforts to increase the enrollment of Hispanic children in prekindergarten programs may be particularly beneficial from both developmental and cost–benefit perspectives.

Ultimately, our study cannot unpack the causal mechanisms behind the detected effects. Our results concern the effects of the combination of these particular prekindergarten curricula and coaching, in the context of Boston's prekindergarten teaching workforce, on children's developmental outcomes. Identifying the causal active ingredients should be a priority in future research on the impact of prekindergarten programs. Likewise, due to the RD design, our results generalize only to students at the cutoff. Future research should prioritize using other research designs, such as randomized controlled trials, to inform the degree to which impacts in our study and similar studies generalize to those farther away from the cutoff. An additional limitation of our study is that children were tested in English due to concerns about the psychometric validity of combining scores from the English and Spanish versions of the same measure (e.g., the PPVT and its Spanish-language counterpart, the Test de Vocabulario en Imagenes Peabody use different norming populations, as well as different stop rules).

Despite these limitations, our results provide further evidence on the benefits of public prekindergarten programs for children. In particular, the combination of evidence-based curricula and coaching supports implemented at scale in the context of Boston's public schooling system brought about educationally and statistically significant improvements in multiple domains of school readiness. As such, the results contribute to the literatures on preschool quality improvement as well as public prekindergarten evaluations.

References

Aikens, N., & Akers, L. (2011). *Background review of existing literature on coaching*. Washington, DC: Mathematica Policy Research.

Anderson, M. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association*, *103*, 1481–1495. doi:10.1198/016214508000000841

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

Ashe, M. K., Reed, S., Dickinson, D. K., Morse, A. B., & Wilson, S. J. (2009). Opening the World of Learning: Features, effectiveness, and implementation strategies. *Early Childhood Services*, *3*, 179–191.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children*, *5*, 25–50. doi:10.2307/1602366

Barnett, W. S., Epstein, D. J., Carolan, M. E., Fitzgerald, J., Ackerman, D. J., & Friedman, A. H. (2010). *The state of preschool 2010*. National Institute for Early Education Research. Retrieved February 28, 2013, from http://nieer.org/yearbook/

Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., et al. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, *23*, 299–313. doi:10.1016/j.ecresq.2008.03.001

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., et al. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, *179*, 1802–1817. doi:10.1111/j.1467-8624.2008.01227.x

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, *20*, 821–843. doi:10.1017/S0954579408000394

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663. doi:10.1111/j.1467-8624.2007.01019.x

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*, 43–82. doi:10.1080/19345747.2011.578707

Brooks-Gunn, J., Gross, R. T., Kraemer, H. C., Spiker, D., & Shapiro, S. (1992). Enhancing the cognitive outcomes of low birth weight, premature infants: For whom is the intervention most effective? *Pediatrics*, *89*, 1209–1215.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*, 273–293.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–31). Baltimore, MD: Brookes.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science*, *6*, 42–57. doi:10.1207/S1532480XADS0601_05

Clements, D. H., & Sarama, J. (2007a). *SRA real math, PreK-building blocks*. Columbus, OH: SRA/McGraw-Hill.

Clements, D. H., & Sarama, J. (2007b). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*, 136–163.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*, 443–494. doi:10.3102/0002831207312908

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Math Assessment. *Educational Psychology*, *28*, 457–482. doi:10.1080/01443410701777272

Clements, D. H., Sarama, J. H., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *4*, 127–166.

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives*, *15*, 213–238. doi:10.1257/jep. 15.2.213

Diamond, A., Carlson, S. M., & Beck, D. M. (2005). Preschool children's performance in task switching on the Dimensional Change Card Sort Task: Separating the dimensions aids the ability to switch. *Developmental Neuropsychology*, *28*, 689–729. doi:10.1207/s15326942dn2802_7

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "do as I say, not as I do." *Developmental Psychobiology*, *29*, 315–344. doi:10.1002/(SICI)1098-2302(199605)29:4%3c315:AID-DEV2%3e3.0.CO;2-T

Dickinson, D. K., Freiberg, J. B., & Barnes, E. (2011). Why are so few interventions really effective? A call for fine-grained research methodology. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. III, pp. 337–357). New York: Guilford Press.

Dickinson, D. K., Kaiser, A., Roberts, M., Hofer, K. G., Darrow, C. L., & Griffenhagen, J. B. (2011). *The effects of two language focused preschool curricula on children's achievement through first grade*. Paper presented at the Society for Research in Educational Effectiveness, Washington, DC.

Domitrovich, C. E., Cortes, R., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the Preschool PATHS Program. *Journal of Primary Prevention*, *28*, 67–91. doi:10.1007/s10935-007-0081-0

Duncan, G. J., Claessens, A., Huston, A. C., Pagani, L. S., Engel, M., Sexton, H., et al. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Bloomington, MN: Pearson Assessments.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2005). First grade and educational attainment by age 22: A new story. *American Journal of Sociology*, *110*, 1458–1502. doi:10.1086/428444

Fischel, J. E., Bracken, S. S., Fuchs-Eisenberg, A., Spira, E. G., Katz, S., & Shaller, G. (2007). Evaluation of curricular approaches to enhance preschool early literacy skills. *Journal of Literacy Research*, *39*, 471–501.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*, 483–527. doi:10.1016/0885-2014(95)90024-1

Fuhs, M. E., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at Head Start. *Developmental Psychology*, *47*, 404–416. doi:10.1037/a0021065

Fuller, B., & Kim, A. Y. (2011). *Latino access to preschool stalls after earlier gains*. Berkeley, CA: Institute of Human Development at the University of California. Retrieved February 28, 2013, from http://ihd.berkeley.edu/Latino%20preschool%20decline%20-%20NOLA-NJLC-Brief-2011-FINAL.pdf

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Education Psychology*, *70*, 177–194. doi:10.1348/000709900158047

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, *4*, 1343–1359. doi:10.1111/j.1467-8624.2007.01069.x

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*, 872–884. doi:10.1037/0012-1649.41.6.872

Gormley, W. T., Phillips, D. A., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*, *320*, 1723–1724. doi:10.1126/science.1156019

Gormley, W. T., Phillips, D. A., Newmark, K., Perper, K., & Adelstein, S. (2011). Social-emotional effects of early childhood education programs in Tulsa. Center for Research on Children in the United States. Retrieved February 28, 2013, from http://www.crocus.georgetown.edu/reports/CROCUSworkingpaper15.pdf

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

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

Han, W. J. (2008). The academic trajectories of children of immigrants and their school environments. *Developmental Psychology*, *44*, 1572–1590. doi:10.1037/a0013886

Harrison, L. J., McLeod, S., Berthelsen, D., & Walker, S. (2009). Literacy, numeracy, and learning in school-aged children identified as having speech and language impairment in early childhood. *International Journal of Speech-Language Pathology*, *11*, 392–403. doi:10.1080/17549500903093749

Haskins, R., & Barnett, W. S. (2010). *Investing in young children: New directions in federal preschool and early childhood policy*. Washington, DC: Center on Children and Families at Brookings and the National Institute for Early Education Research.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*, 114–128. doi:10.1016/j.jpubeco.2009.11.001

Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review*, *26*, 100–112. doi:10.1016/j.econedurev.2005.09.006

Hustedt, J. T., Barnett, W. S., Jung, K., & Goetze, L. D. (2009). *The New Mexico PreK evaluation: Results from the initial four years of a new state preschool initiative—Final report*. National Institute for Early Education Research. Retrieved September 13, 2010, from http://nieer.org/pdf/new-mexico-initial-4-years.pdf

Hustedt, J. T., Barnett, W. S., Jung, K., & Thomas, J. (2007). *The effects of the Arkansas Better Chance Program on young children's school readiness*. State of Arkansas. Retrieved September 9, 2010, from http://www.arkansas.gov/childcare/abc/pdf/longreport.pdf

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635. doi:10.1016/j.jeconom.2007.05.001

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*, 5–86. doi:10.1257/jel.47.1.5

Kelchen, R., Magnuson, K. A., Duncan, G. J., Schindler, H. S., Shager, H., & Yoshikawa, H. (2012). *Do the effects of early childhood education programs differ by gender? A meta-analysis*. Manuscript under review.

Klein, L., & Knitzer, J. (2006). *Effective preschool curricula and teaching strategies*. National Center for Children in Poverty. Retrieved February 28, 2013, from http://www.nccp.org/publications/pdf/text_668.pdf

Leak, J., Duncan, G., Li, W., Magnuson, K., Schindler, H., & Yoshikawa, H. (2012). *Is timing everything? How early childhood education program cognitive and achievement impacts vary by starting age, program duration and time since the end of the program*. Manuscript submitted for publication.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355. doi:org/10.1257/jel.48.2.281

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*, 159–208. doi:10.1162/qjec.122.1.159

Magnuson, K., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*, 33–51. doi:10.1016/j.econedurev.2005.09.008

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children*, *15*, 169–196. doi:10.1353/foc.2005.0005

Mancilla-Martinez, J., & Lesaux, N. K. (2011). Early home language use and later vocabulary development. *Journal of Educational Psychology*, *103*, 535–546. doi:10.1037/a0023655

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly*, *21*, 471–490. doi:10.1016/j.ecresq.2006.09.003

Murnane, R., & Willett, J. (2010). *Method matters: Improving causal inference in educational research*. New York: Oxford University Press.

National Association of Child Care Resource & Referral Agencies. (2011). We can do better: NACCRRA's ranking of state child care center regulation and oversight. Retrieved February 28, 2013, from http://www.naccrra.org/about-child-care/statechildcarelicening/we-can-do-better-state-child-care-center-licensing

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy practices. *American Educational Research Journal*, *46*, 532–566. doi:10.3102/0002831208328088

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between childcare experiences and concurrent development. The cost, quality and outcomes study. *Merrill-Palmer Quarterly*, *43*, 451–477.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social development trajectories through second grade. *Child Development*, *72*, 1534–1553. doi:10.1111/1467-8624.00364

Pianta, R. C., & Stuhlman, M. W. (2004). Teacher-child relationships and children's success in the first years of school. *School Psychology Review*, *33*, 444–458.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*, 302–316. doi:10.1037/a0015302

Reynolds, A. J., Temple, J. A., White, B., Ou, S., & Robertson, D. L. (2011). Age-26 cost benefit analysis of the Child-Parent Center Early Education Program. *Child Development*, *82*, 379–404. doi:10.1111/j.1467-8624.2010.01563.x

Ribordy, S. C., Camras, L. A., Stefani, R., & Spaccarelli, S. (1988). Vignettes for emotion recognition research and affect education programs with children. *Journal of Clinical Child Psychology*, *17*, 322–325. doi:10.1207/s15374424jccp1704_4

Sachs, J., & Weiland, C. (2010). Boston's rapid expansion of public school-based preschool: Promoting quality, lessons learned. *Young Children*, *65*, 74–77.

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. (in press). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*, 489–502.

Schickedanz, J., & Dickinson, D. (2005). *Opening the world of learning*. Iowa City, IA: Pearson.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., et al. (2010). *Standards for regression discontinuity designs*. Institute of Education Sciences. Retrieved February 28, 2013, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_rd.pdf

Schweinhart, L. J., Barnett, W. S., & Belfield, C. R. (2005). *Lifetime effects: The High/Scope Perry Preschool Study through age 40*. Ypsilanti, MI: High/Scope Press.

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

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, *22*, 173–187. doi:10.1016/j.ecresq.2007.01.002

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the *ex post facto* experiment. *Journal of Education Psychology*, *51*, 309–317. doi:10.1037/h0044319

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

U.S. Census Bureau. (2012). *State and county quick facts*. U.S. Census Bureau. Retrieved October 24, 2012, from http://quickfacts.census.gov/qfd/states/25/2507000.html.

U.S. Department of Health and Human Services. (2010). *Head Start Impact Study: Final report*. Washington, DC: Administration for Children and Families, Office of Planning, Research and Evaluation.

Vygotsky, L. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Wagner, D. A., Venezky, R. L., & Street, B. V. (1999). *Literacy: An international handbook*. Boulder, CO: Westview Press.

Wechsler, D. (1986). *Wechsler Intelligence Scale for Children–Revised*. New York: Psychological Corporation.

Weiland, C., Eidelman, H., & Yoshikawa, H. (2012). *Fidelity of implementation in an at-scale prekindergarten program and links to children's cognitive outcomes*. Manuscript in preparation.

Weiland, C., Wolfe, C., Hurwitz, M., Yoshikawa, H., Clements, D., & Sarama, J. (2012). Early mathematics assessment: Validation of a preschool mathematics screening tool. *Journal of Educational Psychology*, *32*, 311–333. doi:10.1080/01443410.2011.654190

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, *102*, 43–53. doi:10.1037/a0016738

Wilson, S. J., Morse, A. B., & Dickinson, D. K. (2009). *Examining the effectiveness of OWL as used in ERF Projects: Final report of results from the OWL Consortium Project*. Nashville, TN: Vanderbilt University Center for Evaluation Research & Methodology.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state prekindergarten programs. *Journal of Policy Analysis and Management*, *27*, 122–154. doi:10.1002/pam.20310

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement*. Itasca, IL: Riverside.

Yoshikawa, H. (1995). Long-term effects of early childhood programs on social outcomes and delinquency. *The Future of Children*, *5*, 51–75. doi:10.2307/1602367

Zigler, E., & Styfco, S. J. (2010). *The hidden history of Head Start*. New York: Oxford University Press.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management*, *19*, 75–92. doi:10.1002/(SICI)1520-6688(200024)19:1%3c75:AID-PAM5%3e3.0.CO;2-W

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Interpreting the RD Estimates.

Appendix S2. Addressing Threats to Validity and Robustness Checks.

Appendix S3. Comparison of Participating and Nonparticipating Schools and Teachers.

Appendix S4. Additional Supporting Tables and Figures.

Appendix S5. References.