Early Childhood Research Quarterly 32 (2015) 60–79

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Do the short-term effects of a high-quality preschool program persist?

Carolyn J. Hill*, William T. Gormley Jr., Shirley Adelstein

Georgetown University, United States

ARTICLE INFO

Article history:
Received 8 May 2013
Received in revised form
16 December 2014
Accepted 29 December 2014
Available online 8 January 2015

Keywords:
Pre-kindergarten
Preschool
Short-term effects
Long-term effects

ABSTRACT

We investigate the persistence of short-term effects of a high-quality school-based pre-kindergarten program in Tulsa, Oklahoma. We analyze third-grade reading and math scores for two cohorts of students eligible to participate in pre-kindergarten in 2000–2001 and 2005–2006, using boosted regression and propensity score matching to select a comparison group of local students who did not participate in the pre-K program. For the early cohort, we find no evidence of persistence of early gains. For the late cohort, we find that early gains persist through third grade in math but not reading, and for boys but not for girls. We discuss possible reasons for the pattern of findings, though our study design does not allow us to identify the causal mechanisms of persistence.

Introduction

State-funded preschool programs have grown dramatically over the past decade. From 2002 to 2012, state-funded pre-K program enrollment doubled, with a total of 40 states now serving more than 1.3 million children. During the same time period, Head Start program enrollment has remained relatively constant, at more than 800,000 children. Approximately 28 percent of all 4-year-olds are now enrolled in state-funded pre-K programs, and an additional 10 percent are enrolled in Head Start (Barnett, Carolan, Squires, & Brown, 2013a). President Obama called for an expansion of high-quality preschool programs to all children in his 2013 State of the Union address, and reiterated his support in the same address a year later.

Given the increasing participation in preschool programs, and the increased policy interest in expanding pre-K opportunities, the longer-term benefits to such early investments are of interest. In this paper, we examine whether the short-term effects of a high-quality state-funded preschool program in Tulsa, Oklahoma, operated at scale, persist over time. For two cohorts of children, we examine whether test score differences between children who participated in public pre-K and a comparable group of local children who did not participate are still evident by the end of third grade. Previous research has shown large short-term gains (at kindergarten entry) in pre-reading, pre-writing, and pre-math skills for three different cohorts of students who participated in Tulsa's school-based pre-K program (Gormley & Gayer, 2005; Gormley, Gayer, Phillips, & Dawson, 2005; Gormley, Phillips, & Gayer, 2008). A statewide study of the Oklahoma pre-K program, using fall 2004 test scores, found statistically significant effects on literacy (the PPVT) but not for math (Wong, Cook, Barnett, & Jung, 2008).

Understanding whether short-term gains of early childhood programs persist, or whether they fade out over time, is of considerable interest to policymakers, educators, and parents who make decisions about funding and placement. Effects of any particular intervention might fade out over time if learning rates stagnate or decline for students who participated in pre-K, if investments are made for students who did not participate that enable them to catch up to participants, or if students forget the material they learn or skills they develop (Jacob, Lefgren, & Sims, 2010). Fade out effects might be observed as well if measurements are not appropriate or robust in a particular year or over time.

Three main factors suggest that effects of a mature, high-quality pre-K program operated at scale are likely to persist. First, the brain's early wiring is critically important to its later development (Shonkoff & Phillips, 2000). Language and other basic skills are more easily acquired at a younger age, and advanced skills can more readily be acquired with a strong foundation in basic skills. In short, "learning begets learning, skill begets skill" (Heckman, 2000, p. 50).

A second, related, reason to expect persistence of effects for a high-quality preschool program is that kindergarten and early elementary teachers may respond to the increasing number of students who have been exposed to preschool education by increasing

* Corresponding author at: McCourt School of Public Policy, Georgetown University, 100 Old North Building, 37th and O Streets NW, Washington, DC 20007, United States. Tel.: +1 202 687 7017.
E-mail address: cjh34@georgetown.edu (C.J. Hill).

http://dx.doi.org/10.1016/j.ecresq.2014.12.005
0885-2006/

the intensity or level of instruction. Kindergarten pedagogy has become more rigorous in recent years (Bassok & Rorem, 2012). Combined with gains in cognitive development, children who participated in preschool may reinforce their learning gains in these more challenging environments. Indeed, Claessens, Engel, and Curran (2014) find higher kindergarten performance on math and reading assessments for all students (regardless of their experiences prior to kindergarten) in classrooms where they were exposed to relatively more advanced content as compared to more basic content.

A third reason to expect persistence is that participation in high-quality pre-K programs is associated with positive social-emotional outcomes, including character building, higher attentiveness, and stronger executive functioning skills (Gormley, Phillips, Newman, Welti, & Adelstein, 2011; Heckman & Kautz, 2013; Weiland & Yoshikawa, 2013). These socio-emotional effects may prove important in elementary school as children define their personal identities, and later on as they choose between high-risk and low-risk behaviors.

Yet at least three factors might explain why pre-K program effects could fade out. First, early elementary teachers may not have adapted the level of instruction to reflect the growing presence of students who experienced high quality pre-K. Kindergarten teachers often emphasize basic counting and shapes when teaching math, even though former pre-K participants (and other children) have already mastered such skills (Engel, Claessens, & Finch, 2013). In such an environment, early gains might dissipate if students do not have opportunities to build on their earlier skills and knowledge.

A second reason that fade-out might occur relates to compensatory investments. Remedial and additional supportive services for lower-performing children can lessen the longer-term gains of early childhood interventions if these supportive services differentially benefit children who did not participate in Tulsa pre-K. While preschool enrollments have been growing in recent years, remedial and supportive services for children have grown as well for children in special education (through the Education for All Handicapped Children Act of 1975 and subsequent legislation) and for children enrolled in after-school services (through the Child Care and Development Block Grant and other programs).

Third, short-term effects of a high-quality preschool program may fade over time if lack of parental investments and other family factors offset or fail to reinforce early gains (Todd & Wolpin, 2003). Poorer outcomes are associated with being in poverty (Duncan & Brooks-Gunn, 1997), living in a single-parent home (McLanahan & Percheski, 2008), having low birthweight (Currie & Hyson, 1999), and having poorly educated parents (Magnuson, 2007). Other factors, such as a rise in the percentage of households where English is not the primary language, make it less likely that educational programs at school will be reinforced at home, perhaps eroding gains in student achievement.

In this paper, we first summarize findings on persistence of effects from other studies of early childhood education programs, in particular those from demonstration projects, Head Start, and preschool programs. Next, we describe the data and methods we use to estimate reading and math test score effects by the end of third grade for two cohorts of children who could have participated in public pre-K in Tulsa, Oklahoma. After presenting the results, we offer further interpretation of our findings.

Demonstration projects

Two random assignment studies in particular have shown the potential of high-quality early childhood education programs to improve long-term outcomes for economically disadvantaged children. Both were targeted programs that focused on a small number of children born in the 1960s and 1970s. The Perry Preschool Study randomly assigned 123 economically disadvantaged black children in Ypsilanti, Michigan either to a two-year high-quality preschool program (the treatment) or to a control group. Researchers tracked both treatment and control group subjects through age 40. Perry participants had higher achievement test scores and homework completion rates in their middle teen years; higher high school graduation rates; higher employment rates in their late twenties, and lower arrest rates and higher earnings at age 40 (Schweinhart et al., 2005). Heckman, Moon, Pinto, Savelyev, and Yavitz (2010) estimate the program's benefit–cost ratio, using a 3 percent discount rate, to be 7.1–12.2 (see Barnett & Masse, 2007, for a higher estimate).

The Carolina Abecedarian Project was a random assignment study of 111 mostly black economically disadvantaged children in Chapel Hill, NC. Born between 1972 and 1977, children in the treatment group received high-quality care from infancy through preschool. In addition to testing in kindergarten and the early elementary school years, follow-ups were conducted at ages 12, 15, 21, and 30. Higher scores on reading and math achievement test scores for treatment group children surfaced early and persisted over time (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001). Treatment group members were more likely to graduate from a four-year college or university, more likely to have full-time jobs, and less likely to be on welfare than control group members (Campbell et al., 2012). On the other hand, no differences in criminal activity, self-reported substance abuse, or adult earnings were found. Barnett and Masse (2007) estimate a benefit–cost ratio for society, using a 3 percent discount rate, to be 2.5. Both the Perry and Abecedarian Projects were targeted, intensive programs that served a small number of children. In contrast, the pre-K program we analyze is offered to all children in a public school setting.

Head Start

Head Start, a federally funded program begun in 1965, provides a comprehensive set of services including education, health, nutrition, and parent involvement to children and families with low incomes. An early, much-criticized study found early fade-out of Head Start's effects, with no differences in third grade (Westinghouse Learning Corporation & Ohio University, 1969). More recently, Currie and colleagues compared children who participated in Head Start with brothers and sisters who did not. White and Hispanic Head Start program participants scored higher than their sibling nonparticipants on standardized verbal and reading achievement tests, both immediately and as their schooling progressed (Currie & Thomas, 1995, 1999; Garces, Thomas, & Currie, 2002). Hispanic participants were much less likely to repeat a grade, and white participants were much more likely to graduate from high school. For black participants, test score effects faded out over time, but arrest rates and booking rates for crime were lower for participants than nonparticipants. Deming (2009) also uses sibling comparisons, finding overall test scores gains at ages 5–6 of about 0.15 standard deviation, fading to 0.05 standard deviation (though not statistically significant) by the early teen years; yet longer-term effects on a scale of early adult outcomes were estimated to be 0.23 standard deviation. Ludwig and Miller (2007) used a regression discontinuity design to compare health outcomes for students who lived in counties where the county just qualified (based on aggregate poverty rates) for grant-writing assistance for Head Start, with those for students in counties that just missed the cutoff. They found substantial reductions in mortality for children aged 5–9 in the Head Start assistance counties compared to those not in the assistance counties. They also found suggestive evidence of positive

effects on high school completion, though not for 8th grade reading or math scores.

A recent random assignment study focused on children who participated in Head Start in the 2002–2003 or 2003–2004 school years as three- or four-year-olds. Short-term impacts were found for a number of measures, including some relatively modest cognitive and social-emotional effects and somewhat stronger effects on access to dental care (Puma, Bell, Cook, Heid, & Lopez, 2005). But most effects had disappeared by the end of kindergarten (Puma, Bell, Cook, & Heid, 2010). By the end of third grade, even fewer effects were evident (Puma et al., 2012). One possible explanation for the apparent Head Start fade-out effect is the high participation in some kind of preschool or child care for children in the control group (unlike the control group children in the Perry or Abecedarian studies). Moreover, 14 percent of 4-year-olds and 18 percent of 3-year-olds in the control group were crossovers who actually enrolled in Head Start (Puma et al., 2012).

Preschool

Until recently, preschools – in general, programs often run by state or local governments that provide educational services to 3- or 4-year-olds – have received relatively less research attention than Head Start. However, several local and state-specific studies and two national studies (in the US and the UK) provide some indication of the longer-term effects for these types of programs.

Chicago's Child-Parent Centers (CPC), established in 1967 with funding from the federal government and run by the Chicago Public Schools, have been studied by Reynolds and colleagues. The school district actively recruited economically disadvantaged children, and parents volunteered to enroll their children in the program, for a treatment group of 989 children who were three or four years old in 1983 or 1984. Using nonexperimental methods to construct a comparison group and estimate program effects, Reynolds found substantial short-term effects on reading and math scores for children who attended preschool, which declined from kindergarten to 3rd grade but then stabilized (A. Reynolds, personal communication, October 22, 2012). Compared to non-participants, CPC participants by age 20 had lower retention in grade rates, lower special education placement rates, higher high school completion rates, and lower arrest rates (Reynolds, Temple, Robertson, & Mann, 2001). At age 28, CPC participants had higher college attendance rates, higher earnings, lower drug and alcohol abuse rates, and lower felony arrest rates (Reynolds, Temple, Ou, Arteaga, & White, 2011). The CPC benefit–cost ratio for society, using a 3 percent discount rate, was estimated to be 7.1 (Reynolds, Temple, Robertson, & Mann, 2002).

A Texas study, exploiting differences in the availability of a targeted state-funded pre-K program within school districts, found evidence of persistent reading effects (0.06 *SD*) and math effects (0.05 *SD*) as of 3rd grade (Andrews, Jargowsky, & Kuhne, 2012). These pre-K participants also were less likely to be retained in grade and less likely to receive special education services. A North Carolina study exploited differences across counties in access to two early childhood programs: Smart Start which aimed to improve county services for all children 0–5 years old, and More at Four which provided high-quality preschool for disadvantaged four-year-olds (Ladd, Muschkin, & Dodge, 2014). The study examined impacts for all children (for participants as well as spillovers for nonparticipants) and in the preferred specification estimated 3rd grade impacts of both Strong Start and More at Four combined (assuming 2009 funding levels) to be 0.20 *SD* for reading and 0.16 *SD* for math; the separate estimated effects of More at Four were 0.13 *SD* and 0.11 *SD*, respectively. A New Jersey study, using multivariate regression, found that the court-ordered Abbott School pre-K program for lower-income school districts produced effects on oral comprehension in 2nd grade of 0.22 *SD* for students who participated in the program for one year, and 0.40 *SD* for those who participated for two years (Frede, Jung, Barnett, & Figueras, 2009). The estimated 2nd grade math effects were 0.24 *SD* and 0.44 *SD* for one- and two-year participants, respectively. By fourth grade, estimated effects of the program for language arts and literacy were not statistically significant for one year of participation and were 0.26 *SD* for two years of participation (Barnett, Jung, Youn, & Frede, 2013b). Corresponding estimated effects on math were 0.17 *SD* and 0.37 *SD* for one- and two-year participants, respectively.

In a study using NAEP data, Fitzpatrick (2008) found some evidence of lasting cognitive benefits from the state's universal pre-K program for Georgia students living in small towns and rural areas. In another study using NAEP data, Cascio and Schanzenbach (2013) found positive impacts on 8th grade math scores for Georgia and Oklahoma, compared to other states, ranging between 0.06 and 0.07 *SD*. In the NAEP studies, it was not possible to control for state-level changes over time such as changes in K-12 spending levels and the size of the ELL population.

In a randomized control trial in Tennessee, Lipsey et al. (2013) found that short-term cognitive and socio-emotional impacts of the Tennessee Voluntary pre-K program had mostly disappeared by the end of 1st grade. Much of this analysis depended on outcomes data that required parental consent; parents of control group children consented at lower rates than parents of treatment group children, and the consent gap was higher for the first of two cohorts.

In Great Britain, the Effective Provision of Pre-school Education Project followed over 3000 children and found substantial long-term effects on cognitive development and other outcomes for children who attended preschool. Controlling for a number of demographic and home environment characteristics, as well as test scores prior to preschool entry, the study found that participation in preschool had positive effects at age 11 on children's English and math test scores and pro-social behavior (Sammons, 2010).

In contrast to the positive findings in Great Britain, Magnuson, Ruhm, and Waldfogel (2007) found evidence of early fade-out effects in their analysis of Early Childhood Longitudinal Study (ECLS) data from across the US. Using propensity score matching and teacher fixed-effects models, they found that pre-K program participation boosted kindergarten test scores in reading and math, but effects diminished considerably by the spring of first grade. Also, negative effects of pre-K participation on social-emotional development worsened from the fall of kindergarten to the spring of first grade (Magnuson et al., 2007). The pre-K programs assessed in this national study almost certainly differed in quality, as measured by teacher education and other inputs, and pre-K program quality has generally improved in recent years. For example, 15 states met all ten of the National Institute for Early Education Research's (NIEER) pre-K quality benchmarks in 2012, as opposed to no states in 2003. Nevertheless, the study suggests that fade-out was pronounced for children who attended pre-K programs throughout the US in 1997–98.

Fade-out in perspective

As Duncan and Magnuson (2013, p. 120) noted, "Most early childhood education studies that have tracked children beyond the end of the program treatment find that effects on test scores fade over time." On the other hand, fade-out takes many different forms, corresponding to different narratives: some pessimistic and others more optimistic.

At one end of the spectrum, short-term cognitive gains might decline quickly over time, leaving virtually no long-term traces of their initial positive impacts. At the other end of the spectrum, short-term cognitive gains might persist unabated over many years. In between, of course, are many other possibilities: diminished but

not extinguished cognitive effects over time, with attendant benefits for program participants; or social-emotional effects that build "soft skills." A growing number of studies suggest that these intermediate possibilities are most likely (Barnett, 2011; Chetty et al., 2011; Heckman et al., 2010), though more empirical research is needed. Also, different narratives may fit different subgroups, even for the same program intervention (Currie & Thomas, 1995, 1999; Heckman et al., 2010).

We cannot resolve these debates by studying one particular program in one particular city for two cohorts of children. Yet our study offers findings from a nationally prominent, contemporary program that has touched the lives of many thousands of students, including students from diverse social classes and diverse racial and ethnic backgrounds.

Method

Programs

Since 1998, Oklahoma has offered universal pre-K education to all eligible children (children who are at least four years old by September 1 of the school year). The state reaches 74 percent of all 4-year-olds, second highest among the 50 states (Barnett, Carolan, & et al., 2013a). Class sizes cannot be greater than 20 children, and child-to-staff ratios cannot exceed 10:1. Lead teachers must have a B.A. and certification in early childhood education; they are paid according to the same salary and benefits schedule as other public school teachers. Curriculum decisions are made at the school or classroom level.

Access to public pre-K in Oklahoma has been universal since 1998, but participation is voluntary. Enrollment statewide in Oklahoma's pre-K program was about five percent in 1997–1998 (the year prior to universal access), almost 40 percent in 1998–1999 (the first year the universal program was offered), just over 50 percent in 2000–2001, and 68 percent in 2005–2006 (the later cohort studied in this paper) (Barnett, Hustedt, Robin, & Schulman, 2005; Gormley & Gayer, 2005). For state pre-K programs in 2001–2002, the year after the early cohort participated in pre-K, the NIEER rated Oklahoma's program first for access for four-year olds, twenty-fifth in spending per child enrolled in state pre-K, and meeting eight out of ten of NIEER's quality standards (Barnett, Robin, Hustedt, & Schulman, 2003). Oklahoma's ratings were the same for 2004–2005, with the exception that spending had dropped to twenty-ninth (Barnett et al., 2005).

Our study focuses on pre-K in the Tulsa, Oklahoma public schools. In the 2000–2001 school year, pre-K programs were offered in 45 of 59 Tulsa Public Schools (TPS) elementary schools; in 2005–2006 they were offered in 50 of 58 TPS elementary schools. Students could attend pre-K programs outside their neighborhood, and busing was often available. Phillips, Gormley, and Lowenstein (2009) have documented levels of instructional support as measured by the CLASS (La Paro, Pianta, Hamre, & Stuhlman, 2002; La Paro, Pianta, & Stuhlman, 2004; Pianta, La Paro, & Hamre, 2008) in the Tulsa pre-K program that exceed those of school-based pre-K programs in other states. Tulsa pre-K teachers also devoted more time to literacy (30%), math (17%), and science (17%) than their counterparts in other states (18%, 9%, and 11% respectively) (Phillips et al., 2009).

Participants

We analyzed two cohorts of children for whom we had available data suitable for analysis. The earlier cohort was enrolled in TPS kindergarten during the 2001–2002 school year (the "early cohort"), and the later cohort was enrolled in TPS kindergarten during the 2006–2007 school year (the "late cohort"). In the year prior to kindergarten, these children either participated in the TPS pre-K program in Tulsa (the treatment group, identified by TPS administrative records described in the next section), or they had some other experience such as participating in a Head Start program, participating in another preschool program, or being cared for at home (the potential comparison group).

We imposed four restrictions to obtain our final analysis sample. Table 1 shows the total number of observations resulting from each restriction, as well as a breakdown by treatment or comparison group. First, using TPS administrative data on birthdates, we removed children from the analysis who were younger than four or older than five years old as of September 1 in 2000 (early cohort) or in 2005 (late cohort), resulting in a sample of children within one year of age. This sample restriction is relatively common and produces a sample with children of similar, and typical, ages during the pre-K year. Second, we excluded children who participated in the four-year old program of the Head Start Community Action Project (CAP) of Tulsa County (identified using CAP administrative data). The CAP Head Start program received state aid under the state's universal pre-K program and has a history of working collaboratively with TPS. Children in CAP Head Start had to meet the additional requirements of the Head Start program, whereas the universal offer of Tulsa pre-K was unrestricted. The CAP Head Start program is unique among Head Start programs, sharing some but not all features of Tulsa's pre-K classroom environments, and differing from pre-K environments and Head Start classrooms in the National Center for Early Development and Learning (NCEDL) Multi-State Study of Pre-Kindergarten and State-Wide Early Education Programs (SWEEP) (Early et al., 2005; Phillips et al., 2009). For these reasons, we view the CAP Head Start program as unique, neither affirmatively "treatment" nor "comparison" group and thus we exclude them from the analysis sample. Third, we excluded children whose records were missing all data. The fourth restriction, resulting from available third grade test score data, is described in the next section.

Reliable data are not available on participation rates by type of care experienced by children remaining in the potential comparison group. However, parent survey responses for the late cohort provide some evidence (though they should be interpreted with caution due to low response rates for this set of questions and potential recall problems of respondents): 44 percent of these children were cared for at home, 26 percent were in a preschool other than TPS pre-K, 13 percent were in a day care center, 9 percent were in a Head Start program other than CAP, and 8 percent were in a home-based day care setting.

Data sources and measures

This study uses data from four sources: (1) administrative data for children enrolled in TPS; (2) administrative data from CAP Head Start, (3) parent survey data from children enrolled in TPS; and (4) administrative data from the Oklahoma Department of Education. TPS staff provided administrative data to the research team for each child enrolled in TPS during an academic year. Information was available on: the child's TPS pre-K program participation, date of birth, race/ethnicity, gender, and school lunch status (free, reduced-price, or full-price lunch eligibility). We used information on Tulsa pre-K program participation to identify our treatment and potential comparison groups. The other characteristics were used as covariates in the matching and estimation models. Administrative data from CAP Head Start were used to identify children who participated in that program during the year prior to TPS kindergarten.

Survey information from sample members' parents constitutes the third data source used in the study. We distributed a paper

Table 1
Sample restrictions and test score data source, by cohort.

Cohort and sample definition	Treatment[a]		Comparison[a]		Total	
	Number dropped	Sample size	Number dropped	Sample size	Number dropped	Sample size
Early cohort						
Original sample – enrolled in kindergarten[b]		1425		2389[c]		3814
Eligible birthday (>01 September 1995 and ≤01 September 1996)		1425	367	2022	367	3447
Not in CAP Head Start[d]		1425	435	1587	435	3012
Have sufficient data		1425	24	1563[c]	24	2988
Have at least one 3rd grade test score (reading or math OPI)	387	1038	602	961	989	1999
# Students in TPS 3rd grade		*761*		*612*		
# Students elsewhere in Oklahoma 3rd grade		*277*		*349*		
Late cohort						
Original sample – enrolled in kindergarten[b]		1594		2520		4114
Eligible birthday (>01 September 2000 and ≤01 September 2001)	29	1565	454	2066	483	3631
Not in CAP Head Start[d]		1565	469	1597	469	3162
Have at least one 3rd grade test score (reading or math OPI)	478	1087	660	937	1138	2024
# Students in TPS 3rd grade		*898*		*712*		
# Students elsewhere in Oklahoma 3rd grade		*189*		*225*		

[a] Treatment and comparison status (that is, whether the child participated in TPS pre-K) was determined using TPS administrative data.
[b] Original sample refers to all students enrolled in TPS Kindergarten during 2001–2002 (for the early cohort); or in 2004–2005 (for the late cohort), identified using TPS administrative data.
[c] Of the 2389 "comparison" observations listed for the early cohort, treatment status is not known for 391 observations: in subsequent sample restrictions, 367 will be dropped based on birth date, and another 24 will be dropped because they are missing nearly all data.
[d] Students in the Community Action Project (CAP) Head Start in Tulsa were identified using administrative data from CAP Head Start.

survey in English and Spanish to the parents of children in both cohorts. For the early cohort, the parent survey was distributed by teachers to parents of all third graders in April and May 2005. In April 2005, third-grade teachers sent home with each child a one-page survey and a letter to his or her parents that explained the purpose of the survey and ensured confidentiality. In May 2005, we asked schools to redistribute the survey. We offered small financial incentives to each school based on the response rate. The one-page survey contained questions on the child's previous preschool experience, siblings, parental marital status, whether the child currently lived with his or her biological father, and the highest level of education attained by each parent. The overall response rate was 61 percent. Ideally, this information would have been collected about the children's families prior to the child's fourth birthday. However, a survey obtaining this information had not been distributed at that point in time. We recognize that some of these measured characteristics from the survey may vary over time, however they are unlikely to have been affected by treatment status.

For the late cohort, the parent survey was administered in August 2006, while children were taking cognitive development tests at school registration. This two-page survey contained the questions above, plus questions about the primary language spoken at home, the child's and parent's place of birth, the availability of internet access at home, and the number of books at home. The overall response rate was approximately 63 percent.

Administrative data from the Oklahoma Department of Education is the fourth data source and the source of the outcome of interest for this study. In particular, we use student scores on the Oklahoma Performance Index (OPI) in third grade. Scores are available separately for math and for reading. The OPI is a scaled score, comparable across years, based on the Oklahoma Core Curriculum Test (OCCT), a criterion-referenced state assessment administered annually in the spring to assess student achievement. The scores factor in the difficulty level of the test and correct for possible guessing. The administration of the OCCT is designed to fulfill No Child Left Behind and state mandates for testing, and math and reading tests are used for federal and state accountability requirements. The third-grade math test had 45 multiple choice questions covering patterns and algebraic reasoning, number sense, number operations and computation, geometry and measurement, and data analysis and probability. The third-grade reading test had 50 multiple choice questions covering vocabulary, comprehension/critical literacy, literature, and research and information (Oklahoma State Department of Education, 2005, 2010). For the early cohort, technical manuals are no longer available that would provide statewide means, standard deviations, and alpha values for the test items. However, this information is available for the late cohort: the statewide third-grade math OPI scaled score ranged from 440 to 990, with a mean of 734.6, standard deviation of 81.8, and alpha of 0.89; and the statewide third-grade reading OPI scaled score ranged from 400 to 990, with a mean of 736.8, standard deviation of 89.4, and alpha of 0.90 (State of Oklahoma Department of Education, 2010).

Ideally, third-grade test score outcomes for all students who participated in TPS pre-K in each of these cohorts could be obtained and compared with those of a fully equivalent group of students who did not participate in TPS pre-K. But sample attrition by third grade occurs due to a number of factors, including students who repeat a grade, move out of state, switch to a home or private school, or take an alternative state assessment. Third grade administrative data from the state Department of Education enabled us to match children who were in TPS in kindergarten with their third grade test scores. If we were restricted to using only third-grade OPI scores for students who remained in TPS, our match rates would have been 46 percent and 44 percent for the early and late cohort, respectively. With the data from the state, (and using last name, first name, and birthdate identifiers), we were able to match an additional 626 and 414 students, resulting in final match rates of 67 percent and 64 percent, respectively (Table 1). This overall rate masks some variation in the match rate by treatment status, which is about 11 percentage points higher for the treatment group than for the comparison group in each cohort: For the early cohort, 73 percent of treatment students were matched with a third grade test, compared with 61 percent of comparison students. For the late cohort, the match rates were 69 percent and 59 percent, respectively. As will be discussed in the limitations section of the paper, the differential match rates could be a concern if comparison group children were more likely to be retained in grade or more likely to be exempt from taking the standard OCCT assessment.

Analysis plan

Ideally we would observe third-grade test scores for each child in two potential conditions: one in which she participated in TPS pre-K, and one in which she did not. We can express the problem as follows: Let $W_i = 1$ for child i who participates in the Tulsa pre-K program, and let $W_i = 0$ for child i who does not participate in the Tulsa pre-K program. Let $Y_i(1)$ refer to the third-grade test score for child i who participated in the pre-K program, and let $Y_i(0)$ refer to the third-grade test score for the same child who did not participate in pre-K. Although both outcomes are possible in theory, in practice we observe only $Y_i(1)$ if $W_i = 1$, and only $Y_i(0)$ if $W_i = 0$ (Holland, 1986; Imbens & Wooldridge, 2009). Because only one of the two potential outcomes for each child is actually realized for each child, we estimate the difference in observed average outcomes for children who did and did not participate in pre-K, taking into account observable characteristics.

Assignment to the treatment (in this case, Tulsa pre-K) or to a control condition through random assignment would ensure that both observable and unobservable characteristics of treatment and control group members are the same, on average. Because pre-K was universally offered in Oklahoma, random assignment was not possible. Under these conditions, an estimate of the program's effect relies on an assumption of unconfoundedness (or conditional independence) of treatment assignment, that is, $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i$ (Imbens & Wooldridge, 2009; Rosenbaum & Rubin, 1983).

A model of the pre-K treatment effect can be obtained by: $Y_{ij} = \beta_0 + \delta_i W_{ij} + \sum_{i=1}^{k} \beta_k X_{ijk} + \varepsilon_{ij}$, where Y_{ij} is the third grade test score for student i in school j, W_{ij} is an indicator for child i defined above to indicate participation in Tulsa's pre-K program, and X_{ijk} are k observable characteristics (described in the next section) for child i. The treatment effect estimate, δ, is unbiased if the model is correct and if all other unmeasured characteristics of students are unrelated both to their participation in pre-K and to their third-grade test scores. This condition is unlikely, however.

Using a regression discontinuity design (RDD), prior research on these same cohorts of children found positive short-term effects on student test scores at kindergarten entry (Gormley & Gayer, 2005; Gormley, Phillips, Adelstein, & Shaw, 2010). In these studies, estimates of the Tulsa pre-K program were obtained by comparing children who had finished Tulsa pre-K (and were just starting kindergarten) with children who were about to begin Tulsa pre-K. The discontinuity was present due to school district policies that limited enrollment based on age as of September 1 of a given year.

For the current analysis, we cannot use RDD to estimate longer-term impacts because both treatment and comparison group children in the short-term RDD studies were eventually exposed to Tulsa pre-K. Instead, we use propensity score matching to identify comparison group members who were not exposed to Tulsa pre-K at all but who were most similar to treatment group children, at least on observable characteristics. While the third-grade outcomes are the primary focus of our paper, we report effects estimated at kindergarten entry using both RDD and propensity score methods from our sample to show their comparability.

We estimate a separate propensity score model for each outcome, for each cohort, and for each subsample. We use a set of covariates associated with both selection into public pre-K and to the third-grade test score, drawn from administrative data (race, gender, age, free/reduced/paid lunch status), as well as from parent surveys (parents' marital status, whether the child's father lived in the home, whether at least one parent was foreign-born, mother's education, language spoken at home, number of books in the home, Internet access at home, and type of preschool/child care when the child was three years old).

We use boosted modeling techniques to estimate the propensity score (McCaffrey, Ridgeway, & Morral, 2004; Schonlau, 2005). Boosted regression can use all available covariates and is not subject to the particular modeling choices made by the analyst (Guo & Fraser, 2010; McCaffrey et al., 2004; Stuart, 2010). These techniques, adapted from the machine learning and statistics literatures, incorporate nonparametric regression or classification trees to find the best model fit in terms of minimizing prediction error. The algorithm iteratively splits the data according to covariate values, applying increasing weights to observations that were not successfully classified in previous rounds. The final boosted model combines results across the iterations to produce a model of pre-K participation. Our boosted model specified five interactions, a training fraction of 0.80, a bagging fraction of 0.5, and a shrinkage parameter of 0.01. These choices are consistent with Schonlau's (2005) recommendations.

We use nearest neighbor one-to-one matching, with replacement (Rosenbaum, 1973). While many algorithms for propensity score analysis exist, there seems to be no consensus on the single best approach in all settings (Caliendo & Kopeinig, 2008; Guo & Fraser, 2010; Stuart, 2010). The choice of matching approach often involves a bias-variance tradeoff. Our approach focuses on achieving covariate balance (Harder, Stuart, & Anthony, 2010), and as shown in the next section, one-to-one matching with replacement produces well-balanced groups overall.

We use OLS on the matched samples to estimate the effects of TPS pre-K, separately for reading and math third-grade OPI scores for the early and late cohort. To account for the fact that some comparison group members were matched more than once, observations were frequency-weighted such that treatment observations received a weight equal to one and comparison observations received a weight equal to the number of times matched; and further to account for the fact that some child observations were used more than one time, we clustered standard errors by child. Regression specifications for the early cohort include covariates for gender, race, lunch status, and age; the later cohort specifications include these covariates plus mother's education, Internet in the home, and presence of father in the home. The covariate sets differ because we had more confidence in the later survey responses than the earlier responses (hence, reliance on administrative data only for the early cohort).

In addition to estimating the four main treatment effects (one for each test in each cohort), we also estimate treatment effects for a number of subgroups based on gender, race, and lunch status. For each outcome, for each cohort, and for each full/subsample analysis, we re-estimate the propensity score model, conducting one-to-one matching with replacement, and assess balance. Our subgroup analyses are "exploratory" in the sense defined by Bloom and Michalopoulos (2010). Thus we do not formally adjust the standard errors to account for multiple hypothesis tests; these subgroup results form the basis for hypotheses to be tested in future studies.

We use multiple imputations to replace missing values of covariates used in our propensity score estimation and regressions predicting program effects (Little & Rubin, 2002; Rubin, 1987). Missing data on covariates from administrative data were rare. As noted earlier, the parent survey response rates were 61 percent and 63 percent for the early and late cohorts, respectively. We imputed values for missing covariates with the Stata *ice* program (Royston, Carlin, & White, 2009) using imputation by chained equations to create 20 imputes – multiple complete data sets based on observed data. Each of these 20 data sets is then analyzed individually, and the results are combined to produce our final parameter estimates and standard errors (Rubin, 1987).

Finally, we conduct sensitivity tests of four types. First, we test whether the results are robust to different covariate controls

Table 2
Full (unmatched) sample characteristics for early cohort by treatment and year.

Characteristic	Early cohort			
	Kindergarten year (2001–2002)		Third-grade year (2004–2005)[a]	
	TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)
From administrative data				
Sample size	1425	1563	1038	961
Age (as of 9/1 of K year)[b]	5.50	5.47	5.53	5.52
Gender (%)				
Male	50.7	50.8	48.6	48.9
Female	49.3	49.2	51.5	51.1
Race/ethnicity (%)				
White, Nonhispanic	37.7	49.6***	37.2	51.7***
Black, Nonhispanic	38.0	20.4***	40.2	19.2***
Hispanic	12.1	16.3***	9.8	15.1***
Native American	10.7	11.9	11.2	12.1
Asian	1.5	1.9	1.6	2.0
Lunch status in kindergarten (%)[c,d]				
Free	57.0	49.5***	53.9	43.2***
Reduced	11.5	7.1***	12.1	7.3***
Paid	31.5	43.4***	33.9	49.5***
From parent surveys				
Mother's education (%)				
No high school degree	16.5	19.3	16.3	19.5
High school graduate	17.1	18.6	17.1	18.7
Some college	50.3	39.9***	50.7	39.7***
College degree	16.1	22.2**	15.9	22.1**
Valid n	*491*	*414*	*485*	*411*
Lives with father (%)				
Yes	57.7	63.2*	57.5	63.2*
No	42.3	36.8*	42.5	36.8*
Valid n	*513*	*440*	*506*	*437*
Mother marital status (%)				
Never married	17.1	12.0**	17.2	11.8**
Married	56.7	61.6	56.9	61.6
Remarried	5.4	5.2	5.5	5.2
Divorced	14.4	14.0	14.2	14.1
Separated	5.0	5.6	4.9	5.7
Widowed	1.4	1.6	1.4	1.6
Valid n	*520*	*443*	*513*	*440*

[a] Third grade analysis sample includes children with at least one OPI score.
[b] Age calculated as of 9/1/2001 for early cohort.
[c] For early cohort, lunch status variable was created from combination of Kindergarten and first grade data. If kindergarten lunch status was missing, First grade lunch status was used.
[d] Early cohort Missingness: 384 observations missing lunch status at kindergarten. 145 observations missing lunch status at third grade.
Difference of means between treatment and comparison groups significant at:
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

in the final OLS model, estimating models (i) without covariates (including only a treatment indicator); (ii) with only covariates from administrative data (race, gender, age, and school lunch status); and (iii) with all available covariates from both administrative and survey data. Second, we test whether the results are robust to a different use of the propensity score by estimating the treatment effects model with propensity score weights. Third, we test whether the results we obtain might be reflecting school characteristics in the 3rd grade instead of Tulsa pre-k participation, by estimating models with school fixed effects. An alternative estimation might include school fixed effects based on treatment school. However, given that comparison group children were not in TPS during their pre-K year (by definition) and were only observed in their kindergarten year, we are not able to estimate a model with pre-K school fixed effects. Fourth, to bound our main findings, we conduct a Rosenbaum bounds test on the estimated pre-K effect (after matching, but without controls for covariates in the estimation). With this exercise, we consider hypothetical values of the correlation between treatment status and some unspecified and unobserved omitted variable, where this unobserved variable is assumed to be perfectly correlated with the test score outcome variable. The question is how strong this unobserved factor would need to be in order for the estimated effect of treatment to no longer obtain statistical significance at the $p < 0.10$ level.

Results

Descriptive statistics

Sample characteristics for the early cohort analysis sample are shown in Table 2, during the kindergarten year and 3rd grade years, separately by treatment group (children who participated in Tulsa

pre-K during 2000–2001) and potential comparison group (children who did not participate in Tulsa pre-K or in CAP Head Start during 2000–2001). Differences of means by treatment status were tested for statistical significance within year.

Overall, children who participated in Tulsa pre-K in the early cohort shared some similarities with those in a potential comparison group, but differed on some observable characteristics. Age and gender balance across groups were roughly equivalent, as expected. In the third grade analysis sample, 40 percent of treatment children were black, compared with 19 percent in the comparison group. The corresponding percentages for white children were 37 and 52 percent, respectively. Hispanic children composed 15 percent of the comparison group in third grade, compared with 10 percent of the pre-K treatment group. Native Americans composed 11–12 percent of the sample. Students receiving a free or reduced-price lunch composed the majority of both the treatment and comparison samples, though just less than half of the comparison group had paid lunch.

Descriptive statistics from measures in the administrative and survey data for the late cohort are shown in Table 3. Gender differences by treatment status in both kindergarten and third grade were evident for this late cohort, as were differences in the proportion of children eligible for a free lunch.

In an effort to eliminate differences in observable characteristics and to construct comparable comparison groups for each cohort, outcome, and subsample combination, we created 20 multiple imputations of missing covariate values, then estimated propensity scores and performed one-to-one matching with replacement, as described above.

Propensity score estimation and balance

Tables 4 and 5 summarize diagnostic statistics from propensity score estimation and matching processes across the 20 imputations for each cohort, outcome, and subsample; boosted model specifications are listed in the table notes, and full results are shown in Appendix A. Table 4 lists separately for each outcome (reading, math), cohort (early, late), and sample (full sample or subgroup): the number of treatment group members, the number of comparison group members (prior to matching), the number of matched comparison group members, the mean propensity score difference between matched pairs in the one-to-one match, and information on the distribution of the propensity scores separately for the treatment and matched comparison groups. For example, the reading outcome in the late cohort full sample had $n = 1068$ treatment group members and $n = 917$ possible comparison group members available to be matched. On average, 476 comparison group members were matched, varying from 459 to 500 across the 20 imputations. On average, the mean difference in propensity scores between matched treatment and comparison group members was only 0.00052 (ranging from 0.00036 to 0.00099). Further, the distributions of propensity scores for treatment and comparison groups overlapped substantially (the area of common support): on average across the 20 imputations, the mean propensity scores was 0.52 for treatment group members and 0.49 for matched comparison group members. The average minimum propensity scores were 0.28 and 0.28, respectively; and the average maximum scores were 0.74 and 0.70 respectively. These statistics showed considerable overlap in the propensity score distributions between treatment and matched comparison group members for this full sample from the late cohort. Similar patterns were seen for the other outcomes, cohorts, and samples.

We relied primarily on the absolute standardized difference (ASD) to assess balance. The ASD was calculated as the absolute difference in means between matched treatment and comparison groups, divided by the square root of the average sample variances for the two groups (Rosenbaum & Rubin, 1985). ASDs less than 0.25 were considered acceptable, with values below 0.10 representing a more stringent standard (Ho, Imai, King, & Stuart, 2007; Rubin, 2001; Stuart, 2010). Mean ASD values across 20 imputations for observable characteristics are shown in Table 5 (full results for each imputation for each cohort, outcome, and subsample are shown in Appendix A). In addition to showing the mean ASD, symbols in the table indicate whether ASDs on that covariate for more than 20 percent of the imputations exceeded the more stringent 0.10 threshold; and whether more than 20 percent exceeded the 0.25 threshold. The ASD < 0.25 criterion was easily met for child characteristics drawn from administrative data (age, gender, free/reduced/paid lunch status, race), and the ASD < 0.10 criterion was met in almost all cases. As noted, we also used covariates from parent surveys to match and assess balance. These characteristics balanced quite well in almost all cases for the full samples in each cohort. In many cases, but not all, when more than 20 percent of the ASDs for a covariate exceeded 0.10, the mean value itself was small in magnitude. When balance could not be achieved, improved balance was not obtained by imposing a caliper; thus we did not impose a caliper, retained all treatment group members, and were able to estimate the ATT. It is possible that unobserved differences between the treatment and comparison groups remain, even though the observed covariates balanced relatively well.

Do children who participated in public pre-K show higher cognitive skills in 3rd grade than comparable children who did not participate in public pre-K?

Estimated effects (and corresponding standard errors) of the TPS pre-K program for reading and math test scores in grade 3 are shown in Table 6, separately by cohort and by method. The first column shows a simple difference of means. For comparison purposes, the second column shows simple OLS results from the full treatment and comparison sample using multiple imputation (20 imputations).

Column 3 shows estimates based on samples constructed using one-to-one propensity score matching with replacement. For the early cohort, we found no statistically significant effects of participation in the TPS pre-K program on third grade test scores, for either reading or math. Similarly for the late cohort we did not find any effects for reading. However, we did find statistically significant results for math: students who participated in Tulsa's pre-K program scored almost 18 points higher ($p < 0.05$) on OPI math than did observationally similar nonparticipants, corresponding to an effect size of 0.18.

To investigate whether the diminished effects that we observed in third grade (compared to earlier research showing strong short-term impacts in kindergarten) were a function of the propensity score methods we use, we estimated an RDD model of the short-term effects in kindergarten, as well as these same short-term effects estimated with the current propensity score matched samples. These results are shown for the early cohort (Fig. 1) and the late cohort (Fig. 2), along with the estimated third grade effects using propensity score methods. For both cohorts, the kindergarten program effect estimates obtained by propensity score matching were smaller than those obtained using RDD. Also for both cohorts, effects were diminishing over time when applying the same estimating technique of propensity score matching.

Note that a higher percentage of treatment group children were tested in kindergarten in each cohort, compared to comparison group children, reflecting the test match difference noted earlier for third grade test scores. For the early cohort in kindergarten, 62 percent of treatment children were tested, compared with 57 percent of comparison children. The corresponding percentages for the late cohort were 81 percent and 70 percent.

Table 3
Full (unmatched) sample characteristics for late cohort by treatment and year.

Characteristic	Late cohort			
	Kindergarten year (2006–07)		Third-grade year (2009–2010)[a]	
	TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	TPS pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)
From administrative data				
Sample size	1565	1597	1087	937
Age (as of 9/1 of K year)[b]	5.50	5.50	5.53	5.55
Gender (%)				
Male	53.0	50.0*	50.2	44.5**
Female	47.0	50.0*	49.8	55.5**
Race/ethnicity (%)[c]				
White, Nonhispanic	32.4	42.7***	33.5	46.5***
Black, Nonhispanic	36.2	24.6***	35.8	23.3***
Hispanic	21.2	20.1	19.6	17.6
Native American	8.9	11.1**	9.3	10.7
Asian	1.4	1.6	1.8	1.9
Lunch status in kindergarten (%)[d]				
Free	65.6	63.4	61.4	56.8**
Reduced	11.8	10.4	13.3	10.7*
Paid	22.5	26.3**	25.4	32.5***
From parent surveys				
Mother's education (%)				
No high school degree	18.8	17.7	15.5	14.2
High school graduate	26.3	24.5	26.7	21.9*
Some college	40.8	39.8	42.4	42.6
College up	14.1	17.9**	15.4	21.3***
Valid n	*900*	*831*	*670*	*535*
Lives with father (%)				
Yes	61.8	57.4**	63.8	59.1*
No	38.2	42.6**	36.2	40.9*
Valid n	*1009*	*928*	*738*	*602*
Mother marital status (%)				
Never Married	24.4	25.1	23.6	23.9
Married	57.9	52.1**	59.7	53.7**
Remarried	2.3	2.5	2.0	2.3
Divorced	8.8	12.9***	8.4	13.3***
Separated	5.3	6.3	4.8	5.5
Widowed	1.3	1.1	1.5	1.3
Valid n	*1019*	*927*	*746*	*602*
Language at home (%)				
English	84.6	85.9	85.3	89.0**
Spanish	13.9	13.3	12.8	10.6
Other	1.5	0.7*	1.9	0.5**
Valid n	*1062*	*975*	*782*	*625*
Internet at home (%)				
Yes	54.2	51.8	57.1	56.1
No	45.8	48.2	42.9	43.9
Valid n	*1017*	*935*	*746*	*604*
Books at home (%)				
0–25	37.6	38.7	34.5	34.7
26–100	33.0	32.0	33.8	32.5
101+	29.4	29.3	31.7	32.8
Valid n	*1021*	*932*	*748*	*603*

[a] Third grade analysis sample includes children with at least one OPI score.
[b] Age calculated as of 9/1/2006 for later cohort.
[c] Late cohort missingness: 35 observations missing race at Kindergarten. 8 observations missing race at third grade.
[d] Late cohort missingness: 34 observations missing lunch status at Kindergarten. 8 observations missing lunch status at third grade.
Notes: Difference of means between treatment and comparison groups significant at
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

For the early cohort, subgroup results (available from the authors) are not presented here because the full-sample effects were not close to being statistically significant. For the late cohort, the estimated math effects on third grade test scores seem to be driven largely by males: while persistent gains by third grade in math were not evident for girls, boys who participated in Tulsa pre-K scored about 21 points higher on the OPI than did boys who did not participate ($p < 0.05$). We found persistent math gains of about 19 points for free-lunch eligible students ($p < 0.05$) and marginally significant gains for paid lunch students ($p < 0.10$), though we are less confident in the latter result due to difficulty finding balance between the treatment and comparison groups. We

Table 4
Propensity score estimation statistics from 20 multiple imputations by outcome, cohort, & subgroup.

Statistic	Reading										
	Early cohort	*Late cohort*									
	Full sample	Full sample	Male	Female	White, Non-Hispanic	Black, Non-Hispanic	Hispanic	Native American	Free lunch	Reduced lunch	Paid lunch
Treatment sample (*n*)	1035	1068	530	538	356	381	211	101	652	143	273
Comparison sample (*n*)	961	917	405	512	427	217	162	93	522	100	296
Matched comparison sample (*n*)											
Minimum	453	459	208	235	171	133	80	38	263	58	123
Mean	478	476	223	252	188	142	89	45	282	63	133
Maximum	507	500	240	268	200	152	98	54	301	69	142
Mean propensity score difference for matched pairs											
Minimum	0.00037	0.00036	0.00060	0.00053	0.00059	0.00090	0.00091	0.00204	0.00052	0.00177	0.00105
Mean	0.00064	0.00052	0.00102	0.00093	0.00135	0.00143	0.00171	0.00340	0.00085	0.00285	0.00148
Maximum	0.00111	0.00099	0.00180	0.00223	0.00282	0.00224	0.00352	0.00507	0.00139	0.00554	0.00221
Distribution of propensity scores											
Treatment Group											
Minimum propensity score											
Minimum	0.193	0.234	0.234	0.250	0.234	0.324	0.286	0.270	0.269	0.270	0.234
Mean	0.248	0.278	0.294	0.281	0.280	0.395	0.341	0.315	0.311	0.340	0.278
Maximum	0.296	0.330	0.343	0.330	0.330	0.444	0.373	0.357	0.359	0.376	0.330
Mean propensity score											
Minimum	0.512	0.514	0.514	0.513	0.469	0.557	0.507	0.485	0.522	0.514	0.490
Mean	0.523	0.524	0.528	0.521	0.476	0.575	0.520	0.500	0.534	0.526	0.501
Maximum	0.537	0.532	0.537	0.530	0.486	0.583	0.532	0.510	0.540	0.540	0.518
Maximum propensity score											
Minimum	0.708	0.665	0.665	0.659	0.606	0.665	0.618	0.615	0.656	0.665	0.659
Mean	0.759	0.735	0.726	0.730	0.680	0.734	0.676	0.662	0.726	0.714	0.725
Maximum	0.823	0.777	0.765	0.777	0.744	0.777	0.738	0.745	0.766	0.765	0.777
Matched Comparison Group											
Minimum propensity score											
Minimum	0.187	0.238	0.222	0.249	0.238	0.329	0.280	0.268	0.267	0.309	0.238
Mean	0.248	0.278	0.294	0.281	0.281	0.401	0.341	0.316	0.310	0.346	0.279
Maximum	0.296	0.330	0.343	0.330	0.330	0.445	0.374	0.357	0.361	0.378	0.330
Mean propensity score											
Minimum	0.473	0.488	0.490	0.481	0.441	0.540	0.489	0.449	0.503	0.495	0.449
Mean	0.478	0.493	0.496	0.491	0.451	0.551	0.497	0.464	0.507	0.505	0.460
Maximum	0.485	0.498	0.506	0.498	0.461	0.558	0.506	0.477	0.512	0.518	0.470
Maximum propensity score											
Minimum	0.686	0.654	0.654	0.651	0.563	0.645	0.584	0.577	0.654	0.614	0.617
Mean	0.720	0.702	0.693	0.690	0.616	0.693	0.639	0.624	0.692	0.665	0.689
Maximum	0.789	0.732	0.732	0.730	0.654	0.732	0.692	0.676	0.727	0.718	0.732

Table 4 (*Continued*)

	Math										
	Early cohort				*Late cohort*						
Statistic	Full Sample	Full Sample	Male	Female	White, Non-Hispanic	Black, Non-Hispanic	Hispanic	Native American	Free lunch	Reduced lunch	Paid lunch
Treatment sample (*n*)	1035	1083	544	539	364	386	213	101	664	143	276
Comparison sample (*n*)	961	932	414	518	431	219	164	100	531	99	303
Matched comparison sample (*n*)											
Minimum	453	462	215	237	175	135	80	40	268	57	123
Mean	478	484	229	254	191	143	90	47	289	62	135
Maximum	507	512	245	269	204	152	99	56	308	68	145
Mean propensity score difference for matched pairs											
Minimum	0.00037	0.00036	0.00059	0.00053	0.00058	0.00090	0.00090	0.00175	0.00051	0.00176	0.00105
Mean	0.00064	0.00052	0.00098	0.00094	0.00133	0.00143	0.00170	0.00323	0.00084	0.00287	0.00146
Maximum	0.00111	0.00099	0.00176	0.00231	0.00277	0.00221	0.00348	0.00471	0.00136	0.00553	0.00219
Distribution of propensity scores											
Treatment Group											
Minimum propensity score											
Minimum	0.193	0.234	0.234	0.250	0.234	0.324	0.286	0.270	0.269	0.270	0.234
Mean	0.250	0.278	0.294	0.281	0.280	0.393	0.340	0.315	0.310	0.340	0.278
Maximum	0.316	0.330	0.343	0.330	0.330	0.444	0.373	0.357	0.359	0.376	0.330
Mean propensity score											
Minimum	0.512	0.514	0.514	0.514	0.468	0.557	0.507	0.485	0.521	0.515	0.490
Mean	0.522	0.524	0.527	0.521	0.476	0.576	0.520	0.500	0.533	0.527	0.501
Maximum	0.537	0.532	0.536	0.530	0.486	0.584	0.532	0.510	0.540	0.541	0.518
Maximum propensity score											
Minimum	0.708	0.665	0.665	0.659	0.606	0.665	0.618	0.615	0.656	0.665	0.659
Mean	0.759	0.735	0.726	0.730	0.680	0.734	0.676	0.662	0.726	0.714	0.725
Maximum	0.823	0.777	0.765	0.777	0.744	0.777	0.738	0.745	0.766	0.765	0.777
Matched Comparison Group											
Minimum propensity score											
Minimum	0.187	0.238	0.222	0.249	0.238	0.329	0.280	0.268	0.267	0.309	0.238
Mean	0.250	0.278	0.294	0.281	0.281	0.399	0.340	0.316	0.310	0.345	0.279
Maximum	0.316	0.330	0.343	0.330	0.330	0.445	0.374	0.357	0.361	0.378	0.330
Mean propensity score											
Minimum	0.473	0.487	0.490	0.481	0.443	0.540	0.488	0.451	0.502	0.498	0.450
Mean	0.478	0.493	0.495	0.491	0.451	0.551	0.497	0.465	0.506	0.506	0.461
Maximum	0.485	0.498	0.505	0.498	0.460	0.558	0.506	0.475	0.511	0.515	0.470
Maximum propensity score											
Minimum	0.678	0.654	0.654	0.651	0.563	0.645	0.584	0.577	0.654	0.614	0.617
Mean	0.719	0.702	0.695	0.690	0.616	0.693	0.640	0.624	0.692	0.665	0.689
Maximum	0.789	0.732	0.732	0.730	0.654	0.732	0.698	0.676	0.727	0.718	0.732

Notes: Cells show statistics across 20 imputations for each sample and outcome. Full results for each imputation for each cohort and outcome are available in Appendix A. Balance statistics for observable characteristics are reported in Table 5.

Table 5
Propensity score balance statistics by cohort, outcome, and subgroup.

Outcome and sample/Subsample	Mean absolute standardized difference (ASD) across 20 imputations									
	Age	Female	Free lunch	Reduced lunch	Paid lunch	White	Black	Hispanic	Native American	Asian
Early cohort (3rd grade in 2004–2005)										
Reading OPI										
Full sample	0.05	0.03	0.07	0.06	0.10[a]	0.08[a]	0.08[a]	0.05	0.03	0.05
Math OPI										
Full sample	0.05	0.00	0.08[a]	0.06	0.10[a]	0.08[a]	0.08[a]	0.05	0.03	0.05
Late cohort (3rd grade in 2009–2010)										
Reading OPI										
Full sample	0.07[a]	0.07	0.04	0.04	0.04	0.04	0.08[a]	0.05	0.06	0.12[a]
Males	0.09[a]	n/a	0.05	0.06	0.04	0.03	0.08[a]	0.05	0.11[a]	0.06
Females	0.06[a]	n/a	0.06	0.08[a]	0.05	0.07	0.10[a]	0.08[a]	0.04	0.20[b]
White, Nonhispanic	0.10	0.06	0.10[a]	0.08[a]	0.09[a]	n/a	n/a	n/a	n/a	n/a
Black, Nonhispanic	0.08[a]	0.10[a]	0.08[a]	0.07[a]	0.10[a]	n/a	n/a	n/a	n/a	n/a
Hispanic	0.07	0.07[a]	0.09[a]	0.09[a]	0.15[a]	n/a	n/a	n/a	n/a	n/a
Native American	0.15[b]	0.18[a]	0.11[a]	0.16[a]	0.12[a]	n/a	n/a	n/a	n/a	n/a
Free lunch	0.10[a]	0.09[a]	n/a	n/a	n/a	0.02	0.05	0.05	0.05	0.08[a]
Reduced lunch	0.07[a]	0.09[a]	n/a	n/a	n/a	0.10[a]	0.11[a]	0.12[a]	0.20[b]	0.18[b]
Paid lunch	0.09[a]	0.11[a]	n/a	n/a	n/a	0.11[a]	0.15[a]	0.11[a]	0.06[a]	0.19[b]
Math OPI										
Full sample	0.08[a]	0.07	0.04	0.04	0.03	0.05	0.08[a]	0.05	0.04	0.12[a]
Males	0.08[a]	n/a	0.05	0.05	0.04	0.03	0.08	0.05	0.09[a]	0.06
Females	0.07[a]	n/a	0.06	0.08[a]	0.05	0.07[a]	0.09[a]	0.08[a]	0.04	0.20[b]
White, Nonhispanic	0.10[a]	0.07	0.09[a]	0.08[a]	0.08[a]	n/a	n/a	n/a	n/a	n/a
Black, Nonhispanic	0.09[a]	0.10	0.08[a]	0.08[a]	0.10[a]	n/a	n/a	n/a	n/a	n/a
Hispanic	0.07	0.07	0.08[a]	0.10[a]	0.12[a]	n/a	n/a	n/a	n/a	n/a
Native American	0.16[b]	0.21	0.12[a]	0.19[b]	0.11[a]	n/a	n/a	n/a	n/a	n/a
Free lunch	0.10[a]	0.10	n/a	n/a	n/a	0.02	0.06	0.06	0.03	0.08[a]
Reduced lunch	0.07	0.09	n/a	n/a	n/a	0.09[a]	0.11[a]	0.11[a]	0.19[b]	0.18[b]
Paid lunch	0.10[a]	0.10	n/a	n/a	n/a	0.13[a]	0.15[a]	0.08[a]	0.06	0.18[b]

Outcome and sample/subsample	Mean absolute standardized difference (ASD) across 20 imputations					
	Parent marital status – married	Parent marital status – divorced	Parent marital status – never married	Parent marital status – remarried	Parent marital status – separated	Parent marital status – widowed
Early cohort (3rd grade in 2004–2005)						
Reading OPI						
Full sample	0.05	0.05	0.06	0.04	0.06	0.04
Math OPI						
Full sample	0.05	0.04	0.05	0.04	0.06	0.05
Late cohort (3rd grade in 2009–2010)						
Reading OPI						
Full sample	0.04	0.05	0.05	0.04	0.04	0.04
Males	0.05	0.06[a]	0.08[a]	0.07	0.05	0.06
Females	0.07[a]	0.05	0.05	0.06	0.06	0.07
White, Nonhispanic	0.07[a]	0.07[a]	0.06	0.07[a]	0.09[a]	0.05
Black, Nonhispanic	0.10[a]	0.11[a]	0.08[a]	0.11[a]	0.09[a]	0.07[a]
Hispanic	0.11[a]	0.09[a]	0.05	0.08[a]	0.10[a]	0.06

Table 5 (*Continued*)

Outcome and sample/subsample	Mean absolute standardized difference (ASD) across 20 imputations					
	Parent marital status – married	Parent marital status – divorced	Parent marital status – never married	Parent marital status – remarried	Parent marital status – separated	Parent marital status – widowed
Native American	0.13[a]	0.18[a]	0.13[a]	0.10[a]	0.16[a]	0.17[a]
Free lunch	0.07	0.06	0.05	0.07[a]	0.05	0.07[a]
Reduced lunch	0.08[a]	0.08[a]	0.13[a]	0.09[a]	0.09[a]	0.20[b]
Paid lunch	0.06[a]	0.04	0.11[a]	0.08[a]	0.10[a]	0.06
Math OPI						
Full sample	0.04	0.05	0.05	0.04	0.04	0.04
Males	0.05	0.06[a]	0.08	0.07	0.05	0.06[a]
Females	0.07[a]	0.05	0.05	0.06	0.07[a]	0.07
White, Nonhispanic	0.08[a]	0.07	0.06	0.08[a]	0.08[a]	0.05
Black, Nonhispanic	0.11[a]	0.11[a]	0.08[a]	0.11[a]	0.09[a]	0.07[a]
Hispanic	0.11[a]	0.09[a]	0.06	0.08[a]	0.10[a]	0.06
Native American	0.12[a]	0.19[b]	0.14[a]	0.12[a]	0.17[a]	0.17[a]
Free lunch	0.08	0.07	0.06	0.06	0.05	0.07
Reduced lunch	0.08[a]	0.09[a]	0.13[a]	0.10[a]	0.11[a]	0.20[b]
Paid lunch	0.06[a]	0.04	0.11[a]	0.07[a]	0.09[a]	0.06

Outcome and sample/subsample	Mean absolute standardized difference (ASD) across 20 imputations								
	Lives with father	Mother educ – less than high school degree	Mother educ – high school degree	Mother educ – some college	Mother educ – college degree	Foreign-born parent	Language spoken at home – English	Language spoken at home – Spanish	Language spoken at home – other
Early cohort (3rd grade in 2004–2005)									
Reading OPI									
Full sample	0.05@		0.04	0.10[a]@					
Math OPI									
Full sample	0.05@		0.04	0.10[a]@					
Late cohort (3rd grade in 2009–2010)									
Reading OPI									
Full sample	0.05	0.04	0.06	0.06	0.04	0.05	0.06	0.05	0.06
Males	0.04	0.05	0.06	0.05	0.08[a]	0.05	0.06[a]	0.07[a]	0.05
Females	0.06	0.06	0.11[a]	0.09[a]	0.06	0.06	0.15[a]	0.13[a]	0.08[a]
White, Nonhispanic	0.08[a]	0.06	0.10[a]	0.06[a]	0.09[a]	0.14[a]	0.13[a]	0.10[a]	0.10[a]
Black, Nonhispanic	0.09[a]	0.07	0.11[a]	0.07[a]	0.09[a]	0.09[a]	0.09[a]	0.09[a]	0.03
Hispanic	0.17[a]	0.09[a]	0.10[a]	0.11[a]	0.09[a]	0.07[a]	0.11[a]	0.10[a]	0.14[a]
Native American	0.10[a]	0.13[a]	0.22[b]	0.16[a]	0.11	0.21[b]	0.10[a]	0.08[a]	0.05[a]
Free lunch	0.06	0.05	0.08[a]	0.06	0.05	0.04	0.06	0.05	0.06
Reduced lunch	0.07[a]	0.10[a]	0.16[a]	0.12[a]	0.12[a]	0.10[a]	0.10[a]	0.08[a]	0.15[a]
Paid lunch	0.06	0.09[a]	0.08[a]	0.09[a]	0.08[a]	0.13[a]	0.18[a]	0.12[a]	0.15[a]
Math OPI									
Full sample	0.05	0.04	0.06	0.06	0.04	0.05	0.06	0.05	0.05
Males	0.05	0.05	0.06	0.05	0.08[a]	0.06	0.07	0.07[a]	0.05
Females	0.07[a]	0.07[a]	0.11[a]	0.09[a]	0.06	0.06	0.15[a]	0.13[a]	0.08
White, Nonhispanic	0.07[a]	0.07[a]	0.11[a]	0.06	0.09[a]	0.14[a]	0.13[a]	0.11[a]	0.10[a]
Black, Nonhispanic	0.09[a]	0.08[a]	0.11[a]	0.08[a]	0.10[a]	0.08[a]	0.08[a]	0.09[a]	0.03
Hispanic	0.17[a]	0.09[a]	0.10[a]	0.11[a]	0.09[a]	0.06	0.11[a]	0.10[a]	0.14[a]
Native American	0.09[a]	0.12[a]	0.21[b]	0.14[a]	0.11[a]	0.16[a]	0.13[a]	0.08[a]	0.08[a]
Free lunch	0.07	0.04	0.08[a]	0.06	0.05	0.05	0.06	0.05	0.06
Reduced lunch	0.07	0.10[a]	0.14[a]	0.12[a]	0.13[a]	0.11[a]	0.09[a]	0.08[a]	0.15[a]
Paid lunch	0.06	0.09[a]	0.08[a]	0.09[a]	0.09[a]	0.12[a]	0.15[a]	0.09[a]	0.14[a]

Outcome and sample/subsample	Mean absolute standardized difference (ASD) across 20 imputations							
	Three-year old care – Center	Three-year old care – Head Start	Three-year old care – Other Home	Three-year old care – NonTPS pre-K	Internet Access	Number of books at home– up to 25	Number of books at home – 26–100	Number of books at home– 100+
Early cohort (3rd grade in 2004–2005)								
Reading OPI								
Full sample								
Math OPI								
Full sample								
Late cohort (3rd grade in 2009–2010)								
Reading OPI								
Full sample	0.05	0.08[a]	0.06	0.06	0.04	0.05	0.06	0.10
Males	0.07[a]	0.09[a]	0.08[a]	0.06	0.05	0.07[a]	0.09[a]	0.14
Females	0.05	0.10[a]	0.07[a]	0.06	0.06	0.05	0.07[a]	0.07
White, Nonhispanic	0.09[a]	0.14[b]	0.13[a]	0.11[a]	0.09[a]	0.13[a]	0.05	0.10
Black, Nonhispanic	0.09[a]	0.08	0.09[a]	0.04	0.12[a]	0.08[a]	0.09[a]	0.09
Hispanic	0.10[a]	0.09[a]	0.07[a]	0.11[a]	0.11[a]	0.11[a]	0.15[a]	0.14
Native American	0.08	0.18[b]	0.15[a]	0.18[b]	0.11[a]	0.08[a]	0.31[b]	0.26
Free lunch	0.08[a]	0.10[a]	0.11[a]	0.06	0.07	0.07[a]	0.06	0.13
Reduced lunch	0.11[a]	0.11[a]	0.09[a]	0.07[a]	0.09[a]	0.10[a]	0.09[a]	0.12
Paid lunch	0.08[a]	0.10[a]	0.09[a]	0.09[a]	0.08[a]	0.11[a]	0.22[b]	0.17
Math OPI								
Full Sample	0.05	0.08[a]	0.06	0.06	0.04	0.05	0.06	0.09
Males	0.07[a]	0.08[a]	0.07	0.06[a]	0.04	0.08[a]	0.09[a]	0.13
Females	0.05	0.11[a]	0.08[a]	0.06	0.06	0.06	0.07[a]	0.07
White, Nonhispanic	0.09[a]	0.13[a]	0.11[a]	0.10[a]	0.09[a]	0.12[a]	0.05	0.10
Black, Nonhispanic	0.08[a]	0.09[a]	0.10[a]	0.05	0.12[a]	0.08[a]	0.09[a]	0.09
Hispanic	0.12[a]	0.09[a]	0.08[a]	0.11[a]	0.10[a]	0.11[a]	0.14[a]	0.13
Native American	0.07	0.20[b]	0.12[a]	0.19[b]	0.11[a]	0.09[a]	0.35[b]	0.29
Free lunch	0.08[a]	0.10[a]	0.11[a]	0.06	0.06	0.08[a]	0.07[a]	0.12
Reduced lunch	0.11[a]	0.12[a]	0.09[a]	0.08[a]	0.09[a]	0.10[a]	0.09[a]	0.11
Paid lunch	0.07[a]	0.10[a]	0.08[a]	0.09[a]	0.07[a]	0.11[a]	0.22[b]	0.18

Notes: Table cells show, for each characteristic in the column, the mean absolute standardized difference (ASD) between the matched treatment and comparison samples. Means are calculated from ASDs for the covariate across 20 imputations. Full details of each imputation are shown in Appendix A, along with boosted model specifications for each model. No superscript following each mean indicates that 4 or fewer ASDs from the 20 imputations exceeded 0.10 and no ASDs exceeded 0.25.

[a] Five or more ASDs exceeded 0.10, but 4 or fewer ASDs exceeded 0.25 (typically 0 or 1 exceeded 0.25).

[b] ASDs in 5 or more imputations exceeded 0.25.

[c] For the early cohort, mother's education was defined someone differently than for the late cohort, with finer gradations at the lower and upper ends of the distribution. As shown in Appendix tables for this cohort, mean ASDs for these categories were < 0.10 and no individual value exceeded 0.10.

Table 6
3rd-Grade reading and math test score impacts of TPS pre-K program: by method, cohort, and subgroup.

Outcome and sample	Difference of means	Std. err.	Regression-adjusted (full comparison sample) with imputed data[a,b,c]	Std. err.	Propensity-score matched[d,e,f]	Std. err.	Propensity-score weighted	Std. err.	School fixed effects (using prop- score matched sample)[h]	Std. err.
Early cohort (3rd grade in 2004–2005)										
Reading OPI										
Full sample	−14.09	3.90***	−2.89	3.73	−3.49	6.71				
Sample size	*1998*		*1998*		*1528*					
Math OPI										
Full sample	−10.20	3.83***	1.78	3.67	1.16	6.40				
Sample size	*1996*		*1996*		*1525*					
Late cohort (3rd grade in 2009–2010)										
Reading OPI										
Full sample	−1.17	4.08	8.41	3.77**	8.60	6.52	8.14	4.00**	3.79	5.56
Sample size	*1985*		*1985*		*1548*		*1985*		*1548*	
Males	0.20	6.06	9.34	5.50*	8.08	9.48	9.51	5.92	1.90	9.52
Females	−1.11	5.47	7.45	5.11	5.16	8.78	6.55	5.27	4.96	8.49
White, Nonhispanic	2.11	5.96	1.71	5.71	−2.36	10.52	1.33	5.92	−6.97	9.65
Black, Nonhispanic	6.45	7.40	8.06	7.30	7.51	11.65[g]	8.17	7.42	2.75	10.88[g]
Hispanic	23.87	8.97***	20.86	8.73**	20.30	14.28[g]	20.20	9.23**	8.08	17.19[g]
Native American	12.26	11.64	14.06	11.22	6.37	18.88[g]	11.48	11.85	26.50	23.49[g]
Free lunch	4.48	5.02	10.76	5.00**	12.48	8.28	10.20	5.20**	8.26	7.84
Reduced lunch	−1.20	10.73	2.58	10.46	−3.54	16.19[g]	−0.99	11.39	−0.11	17.78[g]
Paid lunch	0.93	7.05	6.54	6.72	10.62	12.14[g]	7.97	7.10	−4.58	13.21[g]
Math OPI										
Full sample	6.52	4.37	14.63	4.13***	17.88	7.46**	15.67	4.38***	14.75	6.22**
Sample size	*2015*		*2015*		*1574*		*2015*		*1574*	
Males	8.49	6.41	17.70	5.88***	21.30	9.49**	20.12	6.37**	19.35	9.18**
Females	3.71	5.99	12.06	5.75**	11.53	9.22	11.59	5.88**	9.58	8.78
White, Nonhispanic	10.97	6.61*	9.88	6.54	11.15	12.48	11.08	6.84	7.65	10.91
Black, Nonhispanic	11.06	7.82	11.53	7.73	13.75	12.36[g]	12.52	7.86	9.57	12.38[g]
Hispanic	22.51	9.57**	19.08	9.43**	23.58	16.93	20.41	10.02**	11.43	17.47
Native American	39.95	13.11***	34.72	13.35**	30.74	25.54[g]	33.90	14.53**	61.50	26.47[g]
Free lunch	10.48	5.47*	16.88	5.42***	19.31	8.81**	17.14	5.66***	15.95	8.43*
Reduced lunch	4.34	12.18	8.91	11.73	7.83	17.27[g]	8.81	12.28	7.49	18.23[g]
Paid lunch	11.94	7.69	13.88	7.56*	23.25	13.26*,[g]	16.87	7.99**	8.36	13.49[g]

[a] Regression-adjusted estimates use non-imputed/imputed data, as labeled.
[b] Early cohort regression controls for sex, race, lunch status, and age.
[c] Late cohort regression controls for sex, race, lunch status, age, mother's education, internet in the home, and presence of father in the home.
[d] Propensity-score matched estimates use multiple imputation and boosted regression.
[e] Comparison observations are weighted to account for matching with replacement, such that treatment observations receive a weight equal to 1, and comparison observations receive a weight equal to the number of times they were matched. Also to reflect the use of matching some comparison members multiple times, robust standard errors were clustered by student.
[f] Sample size for propensity-score matched estimates is the sum of treatment observations and matched comparison observations from the first imputation.
[g] Our ability to achieve balance on key covariates varied somewhat across imputations for this subgroup, and thus we have less confidence in this estimate.
[h] Sample size is the sum of treatment observations and matched comparison observations from the first imputation. See note "e" for weighting and s.e. adjustments.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

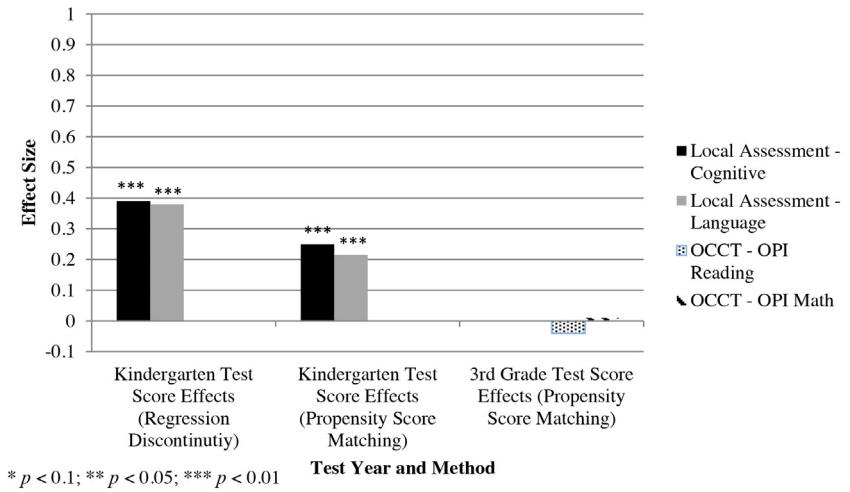

Fig. 1. Early cohort (2001–2002 K) pre-K program effects.

found no statistically significant gains for particular racial/ethnic groups, though adequate balance was not achieved so these findings should be interpreted with caution.

Sensitivity tests

We explored the robustness of our results from the late cohort in four ways. First, we tested whether the results were robust to different covariate controls in the final model, estimating models (i) without covariates (including only a treatment indicator); (ii) with only covariates from administrative data (race, gender, age, and school lunch status); and (iii) with all available covariates from both administrative and survey data. Results were robust to these alternative specifications (available from the authors), with some negligible differences in the treatment effect point estimates following a predictable pattern: the point estimate was largest for the unconditional difference and smallest for the kitchen sink model.

Second, we tested whether the results were sensitive to the choice of one-to-one matching by estimating a model that weights treatment and comparison observations by the propensity score. Column 4 of Table 6 shows these results. Compared with results from the one-to-one matching estimates, standard errors in the propensity-score weighted model tended to be lower (as expected), and the main results from the one-to-one matching held. In a few cases, propensity-score weighted estimates were statistically significant (for example, reading effects for the late cohort full sample, for Hispanic children, and for children receiving a free lunch). Thus the results based on one-to-one propensity score matching were robust to alternative estimation using propensity score weights.

Third, we tested whether the results might instead be reflecting aspects of the school where children were tested in third grade instead of whether children had participated in Tulsa pre-K. These results are reported in Column 5 of Table 6. Results were largely robust to the inclusion of school fixed effects, with point estimates diminishing slightly compared with the one-to-one matching results but with the pattern of statistical significance holding. Thus school composition at time of testing did not seem to be driving our findings.

Fourth, we conducted a Rosenbaum's bounds tests for the later cohort, full sample math result. We found that the result loses statistical significance at Gamma values ranging from 1.15 to 1.20 (Gamma is an odds ratio reflecting the degree to which the omitted factor is assumed to cause the odds of treatment assignment to differ between the treatment and comparison groups). This result suggests some potential fragility of the finding due to an unidentified omitted variable if one exists, however this analysis assumes that the omitted factor perfectly predicts the outcome and increases the odds of treatment by 15–20 percent.

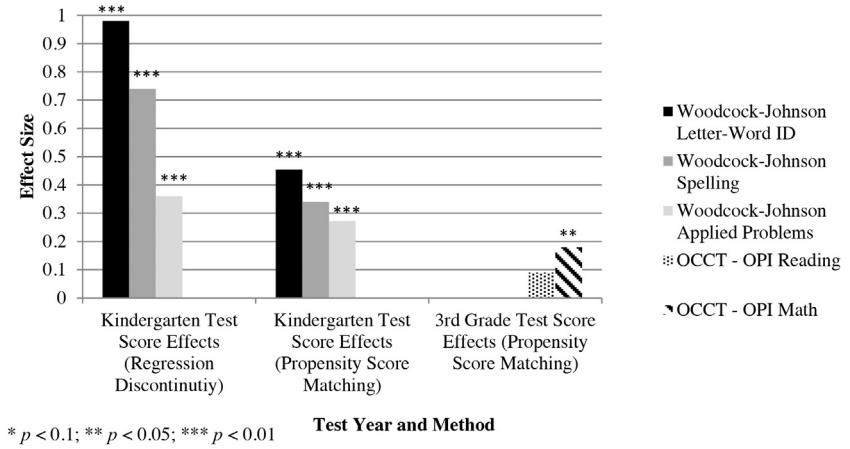

Fig. 2. Late cohort (2006–2007 K) pre-K program effects.

Discussion

For the early cohort, we do not find effects of TPS pre-K participation on third grade test scores in either math or reading. For the late cohort, we find persistent effects on third-grade math test scores of 0.18 *SD* ($p < 0.05$), but no statistically significant effects for third grade reading.

These findings can be interpreted in the context of other studies of preschool programs that have shown varying results in the early elementary years. Though the effect estimates are not directly comparable due to differences in targeted population and estimation techniques, we view our results as broadly consistent with those reported in recent studies.

In particular our results seem most similar to the pattern of results in grades 2 and 4 from one year of Abbott program participation in New Jersey (Barnett et al., 2013b; Frede et al., 2009). Other studies have found persistent effects through third grade in both math and reading: A study of state pre-K in Texas found statistically significant but smaller effects than we found in both math and reading by third grade (Andrews et al., 2012); a study of North Carolina's early childhood initiatives found effects for both reading and math of similar magnitude (these effects were estimated for both participants and spillover effects on nonparticipants) (Ladd et al., 2014); and a study of the Chicago Child-Parent Centers showed declining but still statistically significant effects through third grade (A. Reynolds, personal communication, October 22, 2012). Still other studies have shown effects closer in time to the pre-K experience but fading effects in both math and reading by third grade: the national Head Start study (Puma et al., 2010), and a Tennessee pre-K program (Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013).

The differences in effects we observe across the TPS early and late cohorts may be explained by program maturation, parallel program adjustment, program innovation, or an overall shift toward greater accountability. First, it is possible the program was simply more mature during the later cohort than during the early cohort (when it was only in its third full year of operation). Second, parallel program adjustment or changes in K-3 instruction in Tulsa were possible: Reports from the field in Tulsa suggest growing awareness of school readiness improvements (Shideler & Snow, 2002). K-3 teachers for the early cohort may have focused more on familiar basic skills, as described by Engel et al. (2013), while K-3 teachers for the late cohort may have focused more on new material, as described by Bassok and Rorem (2012). Third, program innovation may explain the differences between the cohorts. TPS launched Tulsa Reads in August 2001 (a system-wide effort to emphasize reading), and Tulsa Counts in August 2003 (a system-wide effort to improve both skills and conceptualization in mathematics). Professional development for both programs was limited, however. TPS also adopted a new mathematics curriculum (Growing with Math) for grades pre-K-5 in spring 2003. This curriculum relied more on manipulative kits that literally enabled younger students to "get a feel" for numbers. Economically disadvantaged children respond better to mathematics problems involving physical objects than to similar problems presented verbally (Clements & Sarama, 2011). Finally, NCLB implementation occurred during the early part of the decade as well, which placed greater emphasis on reading and math test scores, especially during elementary school. Together, these factors may help to explain the positive effects for third graders in 2009–2010 but not in 2004–2005, though given our study design we are not able to rule in or rule out any particular explanation.

One possible explanation for the persistence of early math gains (but not reading gains) in the later Tulsa cohort is that math learning is more confined to the classroom and less subject to change outside the classroom. Whereas many parents promote reading skills, relatively fewer parents devote the same amount of attention to math skills even for younger children (Cannon & Ginsburg, 2008; Ginsberg, Duch, Ertle, & Noble, 2012). If pre-K math instruction creates an advantage for pre-K participants, it is possible that advantage is not affected by family and neighborhood influences in the same ways that reading skills are affected. Also in Tulsa, improving math was a professional development focus for the later cohort which could have resulted in more attention to that part of the curriculum.

Finally, our estimates of the full-sample effects for third-grade math scores in the late cohort were driven primarily by boys. One possible explanation for gender differences in Tulsa could be differences in social-emotional maturity. In Tulsa kindergarten classrooms, boys were more disobedient, more apathetic, less attentive, and marginally more aggressive than girls (supplementary calculations based on Gormley et al., 2010). Another possibility is that Tulsa Counts helped boys in particular to preserve the gains they experienced in pre-K. That intervention, which featured the use of manipulatives, may have been especially appealing to boys (McBride, 2009). Although boys in the comparison group also were exposed to manipulatives in grades K-3, the introduction of this technique at age 4 may have helped to make math more accessible to treatment group boys during a formative period of their development.

Limitations

Ideally, third-grade test score outcomes for all students who participated in TPS pre-K in both the early and late cohorts could be obtained and compared with those of a fully equivalent group of students who did not participate in TPS pre-K. But sample attrition does occur for a number of reasons, notably due to grade retention, taking an alternative form of assessment, or moving out of the district or the state. Matching with data from the state allows us to maintain a considerable number of students in the sample, but full coverage was not achieved and biases may remain that result from sample attrition.

It is possible that differential grade retention by treatment status could affect our findings. We have some evidence from the early cohort to suggest that this may not be a significant factor: Retention rates for TPS pre-K participants ranged from approximately 2–4 percent each year in kindergarten through third grade. These rates are about 1.5–2 percentage points less than retention rates of nonparticipants during the kindergarten and first grade years only; no statistically significant differences in retention rates were evident in second or third grade (these figures are based on supplementary calculations of administrative data by the authors). It is unlikely that these small effects, important though they might be in the short run, could explain the lack of persistence in math and reading scores in the early cohort.

We do not have access to retention data for the later cohort, so we cannot rule out that differential effects on retention rates could affect the observed differences in third-grade test scores for those remaining in the sample. Certainly, it is possible that as the TPS pre-K program has matured, positive impacts (including greater reductions of grade retention) have increased. If students who did not participate in pre-K were more likely to be retained in grade or more likely to be exempt from regular testing than pre-K alumni, then our comparison group, lacking grade-retained students, would be relatively stronger than it otherwise would be, suggesting that our estimated program effects are conservative (i.e., the true program effects are likely larger).

Children on an Individualized Education Plan (IEP) and who met certain other criteria could, if recommended by their IEP team, take the Oklahoma Modified Alternative Assessment Program (OMAAP); or the Oklahoma Alternate Assessment Program (OAAP), which was a portfolio assessment typically administered to children with

more severe disabilities. Scores from these assessments were not comparable to OPI scores from the standard OCCT assessment, and thus were not included in the analysis. We do not have information for the earlier cohort on students taking these alternative assessments. For the later cohort, our supplementary calculations show that among treatment group students with an IEP, 59 percent ($n = 98$) completed the OCCT math assessment while 41 percent ($n = 68$) completed the OMAAP. Among comparison group members with an IEP, 67 percent ($n = 90$) completed the OCCT math assessment while 33 percent ($n = 44$) completed the OMAAP. For reading, among treatment group members with an IEP, 48 percent ($n = 83$) completed the OCCT reading assessment while 52 percent ($n = 89$) completed the OMAAP. Among comparison group members with an IEP, 53 percent ($n = 73$) completed the OCCT while 47 percent ($n = 64$) completed the OMAAP.

It is also possible that intervention effects may appear to fade out due to differences in performance measurement over time, rather than actual convergence of the skills and knowledge of treatment and comparison groups. For example, tests may have ceiling effects (thus masking differences between treatment and comparison groups) or may reflect material that is unrelated to the earlier intervention or to other measures of long-term success. We cannot dismiss these possibilities, especially for the pre-K program's short-term effects on the early cohort, because the test instrument used in kindergarten was not based on a nationally-normed test. In contrast, the nationally-normed Woodcock-Johnson test was used to examine kindergarten outcomes for the late cohort (Woodcock, McGrew, & Mather, 2001).

Finally, our propensity score analysis resulted in samples that balanced relatively well on observable covariates from administrative and survey data. In reporting our results, we have emphasized findings in which we have relatively greater confidence due to observable covariate balance. We cannot, however, rule out the possibility that the treatment and comparison groups differ on unobservable characteristics. This is a concern especially since the offer of public pre-k was universal, yet a considerable number of families chose not to enroll their children in the program. Our analysis procedure eliminates the comparison group members who are least similar to TPS pre-K participants, and three of our robustness checks provide further reassurance in the stability of our main findings. As suggested by the fourth sensitivity test using a Rosenbaum's bounds analysis, however, ultimately we are not able to rule out the possibility that unobservable factors may be driving both participation in the public pre-k program and test outcomes.

Conclusion

Our analysis of a high-quality preschool program finds mixed evidence of its persistence through third grade, as measured by reading and math achievement tests: effects found at kindergarten entry fade out by third grade for an early cohort of students, while short-term gains persist in math but not reading for a later cohort. These findings are consistent with either a program maturation hypothesis (where preschool program implementation improves over time) or a parallel program adjustment hypothesis (where K-3 instruction practices change over time to reflect the growing presence of preschool alumni). They are also consistent with the hypothesis that math gains from preschool participation are less subject to erosion from family and community factors than reading gains, perhaps because math learning is relatively more dependent on what takes place inside the classroom.

To estimate the persistence of preschool program effects, our empirical approach involved multiple imputation to account for missing data, boosted regression to estimate the propensity score, and propensity score matching to construct an observationally similar comparison group. By combining administrative data with original survey data, we were able to utilize a number of covariates, including several variables that captured home characteristics (e.g., mother's education, the primary language spoken at home, internet access at home, and an estimate of books per household). Unlike some studies, we constructed our comparison groups exclusively from local data. Based on examination of matching methods in the job training field, Heckman, Ichimura, and Todd (1997) concluded that geographic similarity and common data sources can eliminate much of the selection bias when constructing a comparison group. If these factors are also important for measuring effects in school settings, then our study is strengthened by the fact that comparison group members are drawn from the same local school district, and by the fact that the same data sources (school administrative data, survey data, and state test score data) were used for both treatment and comparison group members.

In theory, the relationship between short-term and long-term preschool effects could take a number of different forms: persistence, fade-out, sleeper effects, or catch-up effects. Some leading studies (the Perry Preschool project, the Abecedarian project, the Chicago Child-Parent Centers program) have documented some fade-out of cognitive effects during elementary school, followed by some impressive long-term gains (Yoshikawa et al., 2013). Thus, it would be premature to conclude, from diminished 3rd grade test score differentials between treatment group and comparison group children that the long-term benefits of a high-quality pre-K program are unlikely to materialize. In fact, as Duncan and Magnuson (2013) have observed, those early childhood education programs that produce the biggest short-term results also seem to produce the biggest long-term results.

Like some of the most celebrated preschool programs, the Tulsa pre-K program has thus far yielded strong short-term gains. It has also yielded relatively modest gains through grade 3 for the late cohort we examined. This mix of strong initial effects, fade-out, and persistence suggests that the reduction of fade-out should be a high priority for researchers and for public officials. A long list of possible candidates for reducing fade-out includes better alignment of PK-3, longer school days, longer school years, student tutoring programs, after-school programs, parents as teachers programs, better strategies for recruiting and retaining excellent K-12 teachers, innovative professional development strategies, and stronger curricular coordination across grades. Future research is needed to identify the relative effectiveness, and relative benefits and costs, of these and other strategies for sustaining children's gains from high-quality pre-K programs.

Acknowledgements

The authors would like to thank the Tulsa Public Schools and the Oklahoma Department of Education for access to school test data and administrative data. The authors would also like to thank the Foundation for Child Development (Grant # GU-7-08D), the David and Lucile Packard Foundation (Grant # 2008 - 32332), the Spencer Foundation (Grant # 2006 - 00109), the National Academy of Education, and the A.L. Mailman Family Foundation (Grant # 4489-06) for their generous financial assistance. The authors would like to thank anonymous reviewers for many helpful suggestions. Finally, the authors would like to thank Mireya Almazan, Joy Chen, Dan Cullinan, Deanna Ford, Blair Goldstein, Leah Hendey, C.J. Park, Jake Ward, and Catherine Willemin for excellent research assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecresq.2014.12.005.

References

Andrews, R., Jargowsky, P., & Kuhne, K. (2012). *The effects of Texas's pre-kindergarten program on academic performance*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.caldercenter.org/publications/upload/wp-84.pdf

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*, *333*, 975–978. http://dx.doi.org/10.1126/science.1204534

Barnett, W. S., Carolan, M., Squires, J., & Brown, K. (2013). *The state of preschool 2013*. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/sites/nieer/files/yearbook2013.pdf

Barnett, W. S., Jung, K., Youn, M., & Frede, E. (2013). *Abbott preschool program longitudinal effects study: Fifth grade follow-up*. New Brunswick, NJ: National Institute for Education Research, Rutgers University. Retrieved from http://nieer.org/sites/nieer/files/APPLES%205th%20Grade.pdf

Barnett, W. S., Hustedt, J., Robin, K., & Schulman, K. (2005). *The state of preschool 2005*. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/sites/nieer/files/2005yearbook.pdf

Barnett, W. S., & Masse, L. (2007). Comparative benefit–cost analysis of the Abecedarian Program and its policy implications. *Economics of Education Review*, *26*, 113–125. http://dx.doi.org/10.1016/j.econedurev.2005.10.007

Barnett, W. S., Robin, K. B., Hustedt, J. T., & Schulman, K. L. (2003). *The state of preschool 2003*. New Brunswick, NJ: National Institute for Early Education Research. Retrieved from http://www.nieer.org/sites/nieer/files/2003yearbook.pdf

Bassok, D., & Rorem, A. (2012, November). Is kindergarten the new first grade? In *Paper presented at the Annual Meeting of the Association for Public Policy Analysis and Management, Baltimore, MD*.

Bloom, H. S., & Michalopoulos, C. (2010). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups. In *MDRC Working paper*. Retrieved from http://www.mdrc.org/publications/551/full.pdf

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31–72. http://dx.doi.org/10.1111/j.1467-6419.2007.00527.x

Cannon, J., & Ginsburg, H. P. (2008). Doing the math: Maternal beliefs about early mathematics versus language learning. *Early Education & Development*, *19*, 238–260.

Campbell, F., Pungello, E., Miller-Johnson, S., Burchinal, M., & Ramey, C. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology*, *37*, 231–242. http://dx.doi.org/10.1037/0012-1649.37.2.231

Campbell, F., Pungello, E., Burchinal, M., Kainz, K., Pan, Y., Barbarin, O., et al. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian project follow-up. *Developmental Psychology*, *48*, 1033–1044. http://dx.doi.org/10.1037/a0026644

Cascio, E., & Schanzenbach, D. (2013). The impacts of expanding access to high-quality preschool education. In *Brookings Papers on Economic Activity: Fall* (pp. 127–178). Retrieved from http://www.brookings.edu/~/media/Projects/BPEA/Fall%202013/2013b_cascio_preschool_education.pdf

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*, 1593–1660. http://dx.doi.org/10.1093/qje/qjr041

Claessens, A., Engel, M., & Curran, F. C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal*, *51*, 403–434. http://dx.doi.org/10.3102/0002831213513634

Clements, D., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*, 968–970. http://dx.doi.org/10.1126/science.1204537

Currie, J., & Hyson, R. (1999). Is the impact of health shocks cushioned by socioeconomic status? The case of low birthweight. *American Economic Review*, *89*, 245–250. http://dx.doi.org/10.1257/aer.89.2.245

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review*, *85*, 341–364. Retrieved from http://www.jstor.org/stable/2118178

Currie, J., & Thomas, D. (1999). Does Head Start help Hispanic children? *Journal of Public Economics*, *74*, 235–262. http://dx.doi.org/10.1016/S0047-2727(99)00027-4

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal of Applied Economics*, *1*, 111–134. http://dx.doi.org/10.1257/app.1.3.111

Duncan, G., & Brooks-Gunn, J. (Eds.). (1997). *Consequences of growing up poor*. New York, NY: Russell Sage Foundation.

Duncan, G., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109–132. http://dx.doi.org/10.1257/jep.27.2.109

Engel, M., Claessens, A., & Finch, M. (2013). Teaching students what they already know? *Educational Evaluation and Policy Analysis*, *35*, 157–178. http://dx.doi.org/10.3102/0162373712461850

Fitzpatrick, M. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The B.E. Journal of Economic Analysis & Policy*, *8*(1) http://dx.doi.org/10.2202/1935-1682.1897. Article 46

Frede, E., Jung, K., Barnett, W. S., & Figueras, A. (2009). *The APPLES blossom: Abbott preschool program longitudianl effects study: Preliminary results through 2nd grade. Interim report*. New Brunswick, NJ: National Institute for Education Research, Rutgers University. Retrieved from http://nieer.org/pdf/apples_second_grade_results.pdf

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review*, *92*, 999–1012. http://dx.doi.org/10.1257/00028280260344560

Ginsberg, H. P., Duch, H., Ertle, B., & Noble, K. G. (2012). How can parents help their children learn math? In B. H. Wasik (Ed.), *Handbook of family literacy* (pp. 51–65). UK: Routledge.

Gormley, W., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-K program. *Journal of Human Resources*, *40*, 533–558. http://dx.doi.org/10.3368/jhr.XL.3.533

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

Gormley, W., Phillips, D., Adelstein, S., & Shaw, C. (2010). Head Start's comparative advantage: Myth or reality? *Policy Studies Journal*, *38*, 397–418. http://dx.doi.org/10.1111/j.1541-0072.2010.00367

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

Gormley, W., Phillips, D., Newman, K., Welti, K., & Adelstein, S. (2011). Social-emotional effects of early childhood education programs in Tulsa. *Child Development*, *82*, 2095–2109. http://dx.doi.org/10.1111/j.1467-8624.2011.01648.x

Guo, E., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Los Angeles, CA: Sage.

Harder, V. S., Stuart, E. A., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234–249. http://dx.doi.org/10.1037/a0019623

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, *54*, 3–56. http://dx.doi.org/10.1006/reec.1999.0225

Heckman, J. J., Ichimura, I., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654. http://dx.doi.org/10.2307/2971733

Heckman, J., & Kautz, T. (2013). *Fostering and measuring skills: Interventions that improve character and cognition (Working Paper No. 19656)*. National Bureau of Economic Research. http://dx.doi.org/10.3386/w19656

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

Holland, P. T. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960. http://dx.doi.org/10.1080/01621459.1986.10478354

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236. http://dx.doi.org/10.1093/pan/mpl013

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

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*, 915–943. http://dx.doi.org/10.1353/jhr.2010.0029

Ladd, H., Muschkin, C., & Dodge, K. (2014). From birth to school: Early childhood initiatives and third-grade outcomes in North Carolina. *Journal of Policy Analysis and Management*, *33*, 162–187. http://dx.doi.org/10.1002/pam.21734

La Paro, K., Pianta, R., Hamre, B., & Stuhlman, M. (2002). *Classroom Assessment Scoring System (CLASS), pre-K version*. Charlottesville, VA: UVA.

La Paro, K., Pianta, R., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*, 409–426. http://dx.doi.org/10.1086/499760

Lipsey, M., Hofer, K., Dong, N., Farran, D., & Bilbrey, C. (2013). *Evaluation of the Tennessee voluntary pre-kindergarten program*. Nashville: Peabody Research Institute, Vanderbilt University. Retrieved from https://my.vanderbilt.edu/tnprekevaluation/

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

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

Magnuson, K. (2007). Maternal education and children's academic achievement during middle childhood. *Developmental Psychology*, *43*, 1497–1512. http://dx.doi.org/10.1037/0012-1649.43.6.1497

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

McBride, B. (2009). *Teaching to gender differences: Boys will be boys and girls will be girls*. Retrieved from http://www.dedicatedteacher.com/samples/DDTr/ipe3526s.pdf

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425. http://dx.doi.org/10.1037/1082-989X.9.4.403

McLanahan, S., & Percheski, C. (2008). Family structure and the reproduction of inequalities. *Annual Review of Sociology*, *34*, 257–276. http://dx.doi.org/10.1146/annurev.soc.34.040507.134549

Oklahoma State Department of Education. (2005). *Test interpretation manual. Grades 3, 4, 5, 7, & 8. Oklahoma Core Curriculum Tests*.

Oklahoma State Department of Education. (2010). *Test interpretation manual. Grades 3–8. Oklahoma Core Curriculum Tests*. Retrieved from http://ok.gov/sde/sites/ok.gov.sde/files/AA-TIM3-8-10.pdf

Phillips, D., Gormley, W., & Lowenstein, A. (2009). Inside the pre-kindergarten door: Classroom climate and instructional time allocation in Tulsa's pre-K program. *Early Childhood Research Quarterly*, *24*, 213–228. http://dx.doi.org/10.1016/j.ecresq.2009.05.002

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom assessment scoring system manual*. Baltimore, MD: Paul H. Brookes.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study, final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., et al. (2012). *Third grade follow-up to the Head Start impact study final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

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

Reynolds, A., Temple, J., Ou, S., Arteaga, I., & White, B. (2011). School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups. *Science*, *333*, 360–364. http://dx.doi.org/10.1126/science.1203618

Reynolds, A., Temple, J., Robertson, D., & Mann, E. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *Journal of the American Medical Association*, *285*, 2339–2346. http://dx.doi.org/10.1001/jama.286.9.1026

Reynolds, A., Temple, J., Robertson, D., & Mann, E. (2002). Age 21 cost-benefit analysis of the Title I Chicago child-parent centers. *Educational Evaluation and Policy Analysis*, *24*, 267–303. http://dx.doi.org/10.3102/01623737024004267

Rosenbaum, P. R. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*, 159–184. http://dx.doi.org/10.2307/2529684

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. http://dx.doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38. http://dx.doi.org/10.1080/00031305.1985.10479383

Royston, P., Carlin, J. B., & White, I. R. (2009). Multiple imputation for missing values: New features for mim. *The Stata Journal*, *9*, 252–264. Retrieved from http://www.stata-journal.com/

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188. http://dx.doi.org/10.1023/A:1020363010465

Sammons, P. (2010). Do the benefits of pre-school last? Investigating pupil outcomes to the end for key stage 2 (aged 11). In K. Sylva, E. Melhuish, P. Sammons, I. Siraj-Blatchford, & B. Taggart (Eds.), *Early childhood matters: Evidence from the effective pre-school and primary education project* (pp. 114–148). New York, NY: Routledge.

Shideler, R., & Snow, D. (Producers). (2002). *Early childhood development: A Tulsa perspective (video), 16* (Available from Community Service Council of Greater Tulsa, 16 E. 16th St., Tulsa OK.74119).

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a stata plug in. *The Stata Journal*, *5*, 330–354. Retrieved from http://www.stata-journal.com/

Schweinhart, L., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C., & Nores, M. (2005). *Lifetime effects: The HighScope Perry preschool study through age 40*. Ypsilanti, MI: HighScope Educational Research Foundation.

Shonkoff, J., & Phillips, D. (Eds.). (2000). *From neurons to neighborhoods*. Washington, DC: National Academy Press.

State of Oklahoma Department of Education. (2010). *Oklahoma school testing program: Oklahoma Core Curriculum Tests. Grades 3–8 (Technical report, Spring 2010 administration)*. Retrieved from http://www.ok.gov/sde/sites/ok.gov.sde/files/2010Gr3-8.pdf

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21. http://dx.doi.org/10.1214/09-STS313

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, 3–33. http://dx.doi.org/10.1111/1468-0297.00097

Weiland, C., & Yoshikawa, H. (2013). Impacts of prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*, 2112–2130. http://dx.doi.org/10.1111/cdev.12099

Westinghouse Learning Corporation, & Ohio University. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. Athens, OH: Ohio University.

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

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

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., et al. (2013). *Investing in our future: The evidence base on preschool education*. Washington, DC: Society for Research in Child Development and Foundation for Child Development. Retrieved from http://www.srcd.org/2004