DO THE SHORT-TERM EFFECTS OF A STRONG PRESCHOOL PROGRAM PERSIST?

Carolyn J. Hill

William T. Gormley, Jr.

Shirley Adelstein

Center for Research on Children in the U.S. Working Paper

Georgetown University

May 2012

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 Foundation for Child Development, the David and Lucile Packard Foundation, the Spencer Foundation, the National Academy of Education, and the A.L. Mailman Family Foundation for their generous financial assistance; Mireya Almazan, Joy Chen, Dan Cullinan, Deanna Ford, Blair Goldstein, Leah Hendey, C.J. Park, Jake Ward, and Catherine Willemin for excellent research assistance; Tim Bartik, Nora Gordon, Brian Jacob, and Rebecca Ryan for helpful comments on earlier drafts. All errors remain our own.

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 2000-01 and 2005-06, 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.

JEL No. I2, I21, I28

State-funded preschool programs have grown dramatically in recent years. From 2002 to 2010, state-funded pre-K program enrollment nearly doubled, with a total of 40 states now serving almost 1.3 million children. During the same time period, Head Start program enrollment has remained relatively constant, at just over 900,000 children. Approximately 27 percent of all 4-year-olds are now enrolled in state-funded pre-K programs, and an additional 11 percent are enrolled in Head Start (Barnett et al. 2010).

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. Our current analysis builds on previous work showing 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 and Gayer 2005; Gormley et al. 2005; Gormley, Phillips, and Gayer 2008; see also Wong et al. 2008).[1] For two cohorts, we examine whether test score differences between children who participated in pre-K and a local comparable group of children who did not participate are still evident by the end of third grade.

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 must make funding and placement decisions. 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, and Sims 2010, 916).

Two 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 and Phillips 2000). Language and other basic skills are more easily acquired at a younger age, and more advanced skills can be more readily acquired with a strong foundation in basic skills. In short, ―learning begets learning, skill begets skill" (Heckman 2000: 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 ratcheting up their instruction. Combined with gains in cognitive development, children who participated in preschool may reinforce their learning gains in such an environment.

On the other hand, pre-K program effects might fade out, for at least three reasons. First, early elementary teachers may not have actually adapted the level of instruction to reflect the growing presence of students who experienced preschool. In such an environment, early gains might dissipate if students do not have opportunities to build on their earlier skills and knowledge.

Second, remedial and additional supportive services for lower-performing children provide compensatory investments that can lessen the longer-term gains of early childhood interventions if these interventions differentially benefit comparison group children. 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 disadvantaged children enrolled in after-school services (through the Child Care and Development Block Grant and other programs).

Third, short-term effects of a strong preschool program may fade over time if lack of parental investments and other family factors offset or fail to reinforce early gains (Todd and Wolpin 2003). Poorer outcomes are associated with being in poverty (Duncan and Brooks-Gunn 1997), living in a single-parent home (McLanahan and Sandefur 1994), having low birthweight (Currie and 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 participated in pre-K. After presenting the results, we offer further interpretation of our findings. In brief, we find no evidence of persistence by third grade on reading and math scores for the early cohort. For the late cohort, who participated in a more mature pre-K program, we find that short-term gains persist in math but not reading, and that gains persist for boys but not for girls.

BACKGROUND

Prior research has estimated longer-term effects for a range of preschool types, including targeted demonstration projects, Head Start, and other preschool programs.

Demonstration Projects

Two random assignment studies have shown the potential of high-quality early childhood education programs to improve long-term outcomes for 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 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 beginning in the early 1960s, through age 40. Perry participants had higher achievement test scores and homework completion rates in their middle teen years; more high school graduation rates; higher employment rates in their late 20s, and lower arrest rates and higher earnings at age 40 (Schweinhart et al. 2005). The benefit-cost ratio for the program is 7.1 to 12.2 (Heckman et al. 2010; also see Barnett and Massey 2007), driven largely by reduced criminal behavior for a small number of men.

The Carolina Abecedarian Project was a random assignment study of 111 mostly black disadvantaged children in Chapel Hill, NC. Born between 1972 and 1977, the treatment children 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, and 21. Higher scores on reading and math achievement test scores for treatment group children surfaced early and persisted over time (Campbell et al. 2001). Treatment group members were more likely to attend a four-year college or university, more likely to have skilled jobs, less likely to describe themselves as regular smokers, and less likely to become parents as teenagers (Campbell and Ramey 2010). No differences in criminal activity were found. The school-age interventions produced short-term positive impacts, but only for those who participated in preschool beforehand, and no lasting impacts. Barnett and Massey (2007) estimate a benefit-cost ratio of 2.5:1, using a 3 percent discount rate. 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

A federally-funded program begun in 1965, Head Start provides a comprehensive set of services (including education, health, nutrition, and parent involvement) to low-income children. An early, much-criticized study found early fade-out of Head Start's effects, with no differences in third grade (Westinghouse Learning Corporation 1969). More recently, Currie and colleagues used sibling comparisons, contrasting children who participated in Head Start with brothers and sisters who did not. They found that white and Hispanic Head Start program participants scored higher than their sibling nonparticipants on standardized verbal and reading achievement tests, both immediately and later on (Currie and Thomas, 1995, 1999; Garces et al. 2002). For black participants, test score effects faded out over time, but arrest rates and booking rates for crime were lower for participant than nonparticipants. Black males who participated in Head Start were also more likely to have graduated from high school. No evidence of long-term improvements in math scores was found for Head Start program participants (Garces et al. 2002). 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 in counties that just qualified (based on aggregate poverty rates) for Head Start, with those for students in counties that just missed the cutoff. They found substantial differences in mortality for program participants and non-participants, ages 5-9. They also found

suggestive evidence of positive effects on high school completion, though not 8th grade reading or math scores.

A recent random assignment study analyzes children who participated in Head Start in the 2002-03 or 2003-04 school years, as three- or four-year-olds. Short-term impacts were found for a number of measures, including some relatively modest cognitive effects and more substantial health effects (Puma et al., 2005). But by the end of first grade, most of the initial positive effects had disappeared. In fact, most effects had disappeared by the end of kindergarten (Puma et al. 2010). One possible explanation 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).

Preschool

Until recently, preschools—in general, programs often run by state or local governments that provide educational services to three- or four-year-olds—have received relatively less research attention than Head Start. Three recent studies 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 needy 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 (including propensity score methods) to construct a comparison group and estimate program effects, Reynolds (1994: 800) found substantial short-term effects on reading and math scores for children who attended preschool, declining from kindergarten to 3rd grade, disappearing altogether, then reappearing two to three

years later. Compared to non-participants, CPC participants by age 20 had lower retention in grade rates, and lower special education placement rates, higher high school completion rates, and lower arrest rates (Reynolds et al. 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 et al. 2011). The benefit-cost ratio for the Child-Parent Centers was 7.14 overall and 3.85 to the general public (Reynolds et al. 2002: 286, using a 3 percent discount rate).

The Effective Provision of Pre-school Education Project, following over 3,000 children in Great Britain, also found substantial long-term effects on cognitive development and other outcomes for children who attend 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: 126).

In contrast to the positive findings in Chicago and 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 U.S. 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 that 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: 45). The pre-K programs assessed in this national study almost certainly differed in quality, and pre-K program quality has generally improved in recent years. Nevertheless, the study suggests that fade-out was pronounced for children who attended pre-K programs throughout the U.S. in 1997-98.

DATA

We analyze the longer-term effects of Oklahoma's universal pre-K program for students in the Tulsa, Oklahoma school district. Since 1998 Oklahoma has offered universal pre-K education to all eligible children (i.e., those children who are at least four years old by September 1 of the school year).[2] The state reaches more than two-thirds of all four-year-olds, a higher percentage than any other state (Barnett et al. 2010). 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 be certified in early childhood education; they are paid according to the same salary and benefits schedule as other public school teachers.

While access to public pre-K in Oklahoma has been universal since 1998, participation is voluntary. Parents may choose to care for their children at home or enroll their four-year olds in private preschool, Head Start (if eligible), or with other childcare providers or caregivers. Estimates indicate that enrollment statewide in Oklahoma's pre-K program was about 5 percent in 1997-98 (the year prior to universal access), growing to almost 40 percent in 1998-99 (the first year the universal program was offered), to just over 50 percent in 2000-01 and to 68 percent in 2005-06 (the cohorts studied in the current paper) (Gormley and Gayer 2005; Barnett et al. 2005)

We use administrative data and parent survey data from the Tulsa Public Schools (TPS), and administrative data from the Oklahoma Department of Education. These data were collected for two cohorts of children. 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").[3] Both cohorts can be divided into three groups: children who attended the TPS pre-K program the previous year; children who attended

the Community Action Project (CAP) of Tulsa County Head Start program the previous year; and children who attended neither TPS pre-K nor CAP Head Start the previous year but may have experienced other child care or preschool arrangements. Our current analysis uses children in the first and last of these groups (removing children from the CAP Head Start program from the analysis), and we further limit the sample to children who were between four and five years old as of September 1 in 2000 (for the early cohort) or in 2005 (for the late cohort) (Appendix Table B1 shows sample sizes resulting from each of these sample restrictions).

Administrative records from TPS provided information on each student's pre-K program participation, date of birth, race/ethnicity, gender, and school lunch status (free, reduced-price, or full-price lunch). Sample characteristics for the early cohort analysis sample are shown in Table 1, during the kindergarten year and 3rd grade years, separately by treatment group (children who participated in Tulsa pre-K during 2000-01) and comparison group (children who did not participate in Tulsa pre-K or in CAP Head Start during 2000-01). Differences of means (or proportions) by treatment status are tested for statistical significance within year.

Average age and gender balance across groups are roughly equivalent, as expected. The percentage of black students in the treatment group was greater than in the comparison group, and the percentage of white students was less. For example, 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 constituted a greater percentage of the comparison sample (for example, 15 percent in the 3rd grade year, compared with 10 percent of the pre-K treatment group in that year). Native Americans constituted 11 to 12 percent of the sample. Students receiving a free or reduced-price lunch constituted the majority of both the treatment and comparison samples, though just less

than half of the comparison group had paid lunch. Overall, children who participated in Tulsa pre-K in the early cohort share some similarities with those in a potential comparison group, but differ on some observable characteristics.

[Table 1 about here]

Administrative data for the late cohort are shown in Table 2. Gender differences by treatment status in both kindergarten and third grade are evident for this late cohort, as are differences in the proportion of children eligible for a free lunch.

[Table 2 about here]

To supplement the administrative data, we distributed a 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.[4] The one-page survey contained questions on the child's previous preschool experience, siblings, parental marital status, whether currently living with his or her biological father, and the highest level of education attained by both parents.[5] 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 additional questions, including 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 61 percent for the early cohort and 63 percent for the late cohort. Appendix Tables B2 and B3 show descriptive statistics prior to matching for the survey variables for both cohorts.

Third grade test scores, provided by the Oklahoma State Department of Education, serve as the outcomes of interest for both cohorts. These test scores are based on the Oklahoma Core Curriculum Test (OCCT), a criterion-referenced state assessment administered annually in the

spring to assess student achievement. 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 current analysis is based on the standard assessment in both math and reading. In addition to scores for individual sub-sections, tested students received an overall performance score for each subject, the Oklahoma Performance Index (OPI). OPI scores are reported on a scale from 400-990; they factor in the difficulty level of the test and correct for possible guessing.

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 Department of Education enabled us to match children for whom we had kindergarten test scores with their third grade test scores. If we were restricted to using only third-grade OPI scores for students who remained in the Tulsa system, our match rates would have been 46 percent and 54 percent for the early and late cohort, respectively. With the data from the state, we were able to match an additional 622 and 414 students, resulting in final match rates of 67 percent and 64 percent, respectively (Appendix Table B1). 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.[6]

We use multiple imputation for missing values of covariates used in our propensity score models and regressions predicting program effects (Rubin 1987; Little and Rubin 2002). In particular, we used the Stata *ice* program (Royston, Carlin, and White 2009) using imputation by

chained equations to create 20 imputes – multiple complete data sets based on observed data. Each of these twenty data sets is then analyzed individually, and the results are combined to produce our final parameter estimates and standard errors (Rubin 1987).

EMPIRICAL APPROACH

Ideally we would observe third-grade test scores for a child in two potential states: 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 potentially observable, in practice we observe only $Y_i(1)$ if W_i=1, and only $Y_i(0)$ if W_i=0 (Holland 1986; Imbens and Wooldridge 2009). Because only one of the two potential outcomes for each child is actually realized for each child, we estimate the average effect of treatment on the treated as 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 characteristics of treatment and control group members are the same, on average. Because pre-K was universally offered in Oklahoma,[7] random assignment was not possible. Under these conditions, an estimate of the program's effect relies on an assumption of unconfoundedness (or conditional independence) in treatment assignment, that is: $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i$ (Rosenbaum and Rubin 1983; Imbens and Wooldridge 2009).

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 of δ, is unbiased 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.

We use propensity score matching to identify comparison group members who are most similar to treatment group children. We use boosted modeling techniques to estimate the propensity score (Schonlau 2005; McCaffrey, Ridgeway, and Morral 2004). 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.[8] We use a rich set of covariates in these models for each cohort, with characteristics drawn from administrative data (race, gender, age, free/reduced lunch status), as well as characteristics collected in parent surveys. Boosted regression can use all available covariates and is not subject to the particular modeling choices made by the analyst (McCaffrey et al. 2004; Guo and Fraser 2010; Stuart 2010).

We use nearest neighbor one-to-one matching, with replacement (Rosenbaum1973). Across the 20 imputations, the total sample sizes for the treatment group (n=1,083) and unmatched comparison group (n=932) are the same because we did not impose a caliper and did

not trim based on common support. When balance could not be achieved without a caliper, imposing a caliper did not result in improved balance. The matched comparison sample size varies somewhat across imputations, ranging from 462 to 512, with a mean of 484. The mean difference in propensity scores between matched treatment and comparison group members is quite small (0.00052 across the 20 imputations), and values of the propensity score for the matched sample at different points in the distribution overlapped substantially.

We assess balance by relying primarily on the absolute standardized difference (ASD), calculated as the absolute difference in means between matched treatment and comparison groups, divided by the average of the square root of the average sample variances for the two groups (Rosenbaum and Rubin 1985). ASDs less than 0.25 are considered acceptable, with values below 0.10 representing a more stringent standard (Ho et al. 2007; Rubin 2001; Stuart 2010). Full results for each imputation, as well as the percentage reduction in bias associated with matching, are available from the authors. 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. While observable covariate balance is achieved to a considerable degree, it is possible that unobserved differences between the treatment and comparison groups remain.

We use OLS on the matched samples to estimate the effects of TPS pre-K, separately for the 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. Standard errors were clustered 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 (and for the late cohort, based on social-emotional scores in kindergarten). We do not formally adjust the standard errors to account for multiple tests, as we consider the subgroup results to be exploratory rather than confirmatory (Bloom and Michalopoulos 2010).

RESULTS

Estimated effects of the TPS pre-K program for reading and math test scores in grade 3 are shown in Table 3, separately by cohort and by method. The first column shows a simple difference of means, with the corresponding standard error of the difference. The second column shows OLS results from the full treatment and comparison sample, dropping observations with any item nonresponse. The third column shows OLS results from the full treatment and comparison sample but using multiple imputation (with 20 imputes); and the fourth column shows results from propensity score one-to-one nearest neighbor matching, based on a dataset constructed with multiple imputation described earlier. In our analysis below, we focus on the propensity score estimates in the last column.

[Table 3 about here]

For the early cohort we find no statistically significant effects of participation in the TPS pre-K program on third grade test scores, for either reading or math. In contrast, for the late cohort we do not find any effects for reading, but do 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.

Prior research on these same cohorts of children found positive short-term effects (at kindergarten entry) on student test scores (Gormley and Gayer 2005; Gormley et al. 2010), estimated with a RDD design. To demonstrate that the diminished effects we observe by third grade are unlikely to be solely a function of the propensity score methods used, Figure 1 shows a comparison of (a) short-term effects (kindergarten test scores) estimated with RDD; (b) short-term effects estimated with propensity score matching; and (c) 3rd grade effects estimated with propensity score matching, for the early cohort. In Figure 2 we present the same comparisons for the late cohort.[9] Kindergarten program effect estimates obtained by propensity score matching are more conservative than those obtained using RDD. For both cohorts, it is also clear that effects are diminishing over time when applying the same estimating technique of propensity score matching.

[Figures 1 and 2 about here]

Subgroup results for the early cohort (available from the authors) are not shown because the full-sample effects were not close to being statistically significant. The estimated math effects on third grade test scores for the late cohort seem to be driven largely by males: while persistent gains by third grade in math are not evident for girls,[10] boys who participated in Tulsa pre-K score 21.3 points higher on the OPI than do boys who did not participate ($p<0.05$). We find persistent math gains of approximately the same magnitude for free-lunch eligible students.

We find no statistically significant gains for particular racial/ethnic groups, but we do find marginally significant effects of pre-K participation on third grade math scores for black males (ES=0.31, p=0.105).[11]

We conducted a number of sensitivity tests. First, we tested whether the results were robust to different covariate controls in the final model, estimating models (i) without covariates (including just 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, 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 estimated the results using weights based on the propensity score instead of one-to-one matching with replacement. The results were robust to this alternative estimation method, and in a few cases were even marginally stronger and statistically significant where they had not been for the matching methods. Finally, the results were robust to the inclusion of school fixed effects. (Sensitivity test results are available from the authors.)

DISCUSSION

In this section, we discuss some possible explanations for our findings of persistent effects on third-grade test scores for the second cohort but not the first, for math but not for reading, and for boys but not for girls. With the data available, however, we are not able to isolate the effects of any particular explanation.

Differences by Cohort

Program maturation may account for our findings of persistent effects of the TPS pre-K program for the late cohort (enrolled in pre-K in 2005-06) but not for the early cohort (enrolled in 2000-01). For the early cohort, Oklahoma's universal pre-K program was only in its third full year of operation. TPS and other school systems had to implement the pre-K program relatively quickly, making many choices about hiring, curriculum, and professional development. Even if all these choices were sensible, such a large shift in policy takes time to implement effectively. Thus, one difference between the two cohorts may be that the pre-K program was more mature and running more smoothly in 2005-06 than it was in 2000-01.

Parallel program adjustment may also explain the differences across cohorts. For short-term learning gains to be sustained, changes are likely needed in K-3 instruction – what is taught, when it is taught, and how it is taught. If K-3 teachers do not adjust curriculum and teaching strategies to account for the growing presence of children who experienced preschool, then comparison group children are likely to catch up while learning gains of treatment group children are likely to stall. Reports from the field in Tulsa suggest growing awareness of school readiness improvements, which were well-publicized in Tulsa and which many teachers reported noticing themselves (Community Service Council of Greater Tulsa 2002). But Oklahoma ranks 44th in state spending per capita on elementary and secondary education (U.S. Census Bureau 2009: Table 253), and its fourth graders routinely score below the national average on standardized tests (NCES 2010a, 2010b), suggesting a relatively weak elementary school program overall.

Program innovation provides a third possible explanation for differences in persistence across cohorts. Between 2001 and 2006, TPS launched two new programs that had the potential to improve math and reading instruction for pre-K and beyond. In August 2001, TPS launched Tulsa Reads – a system-wide effort to make ~~–e~~very adult in the (school) building a reading

teacher." During the first year of operation, TPS devoted seven full days of professional development almost exclusively to reading . In August 2003, TPS launched Tulsa Counts – a system-wide effort to improve both skills and conceptualization in mathematics. One feature of Tulsa Counts was to focus on a manageable set of critical tasks and to move away from math instruction that was perceived to be ~~"a~~ mile wide and an inch deep" (Tulsa Public Schools 2003: 2).

Yet the Tulsa Reads and Tulsa Counts did not transform early elementary education in TPS. Major budget cuts during the second year of operation for Tulsa Reads resulted in professional development being shut down almost completely. In 2003-04, however, TPS hired literacy coaches for the lowest performing schools in reading . Tulsa Counts faced problems as well, because of relatively few professional development days to support it. Even so, a shift in mathematics pedagogy more generally was occurring in Tulsa as it adopted a new mathematics curriculum (Growing with Math) for grades pre-K-5. This curriculum relied more on manipulative kits like tubs of counting bears that literally enabled younger students to ~~"g~~et a feel" for numbers. Disadvantaged children in particular respond better to mathematics problems involving physical objects than to similar problems presented verbally (Clements and Sarama 2011).

To sum up, TPS became more intentional in its reading and its math instruction between 2001 and 2006. TPS focused teachers' attention, iteratively, on reading and then math. Also, TPS made reading and math instructional improvements a priority across all grades. All of this, of course, coincided with the implementation of NCLB, 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 2009-10 but not in 2004-05.

Differences by Subject

Our findings show persistence of effects by third grade for math but not for reading. Yet the initial learning gains (at kindergarten entry) for the 2005-06 pre-K cohort were substantially greater for pre-reading and pre-writing skills than for pre-math skills (Gormley, Phillips, and Gayer 2008), suggesting potential for a similar pattern by third grade. Further, a common assertion by TPS teachers in the early grades is that teaching reading is easier for them than teaching math. If so, one might expect reading gains to be more easily sustained because the presumptive quality of K-3 reading instruction was superior to that of K-3 math. On the other hand, if math instruction was weaker, perhaps K-3 students who did not enroll in pre-K had fewer opportunities to catch up with their peers.

Perhaps the most convincing explanation for the persistence of early math gains, as opposed to early reading gains, is that math learning is more self-contained, more confined to the classroom, and less subject to either improvement or deterioration outside the classroom. Children's verbal skills depend a great deal on their parents, their peers, and their neighbors, who can either enhance or damage those skills. In contrast, children's math skills depend much more on what goes on inside the classroom. Whereas many parents read to their children, encourage reading, arrange play dates with other children who have above average verbal skills, and generally promote reading skills, few parents devote the same amount of energy and concern to math skills. Thus, if pre-K math instruction creates an advantage for pre-K participants, that advantage is less likely to be eroded by family and neighborhood influences than the greater initial advantage in reading skills that pre-K participants exhibited.

It is possible that the timing of Tulsa Reads and Tulsa Counts initiatives could influence the patterns of persistence. The math initiative was undertaken two years before the 2005-06 pre-

K cohort experienced pre-K. In contrast, the reading initiative was undertaken four years before that time. Conceivably, the more recent emphasis on math meant that school officials generally and teachers in particular were more energized about math than about reading: improving math was the latest fad, which could have resulted in more attention to that part of the curriculum.

Differences by Gender

The full-sample positive findings on persistent cognitive effects for the late cohort are being driven primarily by boys. The medium-term effects of the TPS pre-K program for boys are statistically significant for math. For reading, subgroup results depend somewhat on the estimation method: using one-to-one matching with replacement, gains are not persistent for either gender. But using propensity score weighting (which results in lower standard errors compared to the one-to-one matching estimation), results are marginally significant for boys ($p=0.108$) with an effect size of 0.10.

These results are somewhat surprising because, in the short run, effects of the TPS pre-K program on applied problems test scores were actually somewhat higher for girls than for boys. Yet, by third grade, the boys who participated in TPS pre-K are doing better than other boys, while the girls who participated in TPS pre-K are roughly comparable to other girls.

These results differ from findings for the Perry Preschool Study, in which Anderson (2008) found that girls, not boys, were benefiting from the Perry program at age 10, as measured by their test scores. Of course, sample members were not a racially diverse cross-section of students, and the intervention took place two generations before the Tulsa intervention. Nevertheless, in our study, it was the black males, not the black females, whose third grade test scores (in math) were higher than comparable students if they participated in pre-K.

One possible explanation for gender differences in Tulsa could be differences in social-emotional maturity. From an early age, boys are less focused and more easily distracted than girls. In Tulsa kindergarten classrooms, we found statistically significant differences between boys and girls across multiple dimensions.[12] Boys were more disobedient, more apathetic, and marginally more aggressive than girls. They were also less attentive. In short, boys are more likely to exhibit behavior problems which can interfere with classroom learning. A strong pre-K program may keep boys on track and give them an edge as they move through the early elementary grades.

Another possibility is that the specific math intervention selected by TPS in 2003 helped boys in particular to preserve the gains they experienced in pre-K. That intervention, which featured the use of manipulatives (physical objects to be used for basic math exercises) may have been especially appealing to boys, whose brains are more geared for "spatial mechanical" functioning (McBride 2009) Although comparison group boys also were exposed to manipulatives in grades K-3, the introduction of this technique at age 4 may have helped to make math more enjoyable and more accessible to treatment group boys during a formative period of child development.

It is possible that intervention effects may appear to fade out due to differences in performance measurement over time, rather than because the skills of the treatment and control groups are actually converging. For example, tests may have ceiling effects (thus masking differences between treatment and control groups) or may reflect material that is unrelated to the earlier intervention or to other measures of long-term success. While we cannot dismiss these possibilities, they seem more applicable to the early cohort than to the late cohort: the

kindergarten results for the late cohort were based on a nationally-normed test (the Woodcock-Johnson Achievement tests), while results for the early cohort were not.

CONCLUSION

Our analysis of a strong 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, 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 a comparison group. By combining administrative data with original survey data, we were able to utilize a wide variety of covariates, including several variables that capture 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 that the same data sources (school administrative data, survey data, and state test score data) are used for both treatment and comparison group members.

To place our findings in perspective, it should be noted that preschool program effects on school achievement often diminish over time (Puma et al. 2010; Magnuson et al. 2007; Schweinhart et al. 2005; Campbell et al. 2001), sometimes disappearing altogether by grade 3 or sooner (Reynolds 1994; Reynolds and Temple 1995; Magnuson et al. 2007; Puma et al. 2010). Yet some of these same studies have shown subsequent positive effects on high school completion, college attendance, and a wide variety of social outcomes for teenagers and adults. Whether this turns out to be true of the Tulsa pre-K program remains to be seen. We do, however, find that Tulsa's more mature pre-K program has short-term effects on student math achievement that remain discernible through third grade, driven primarily by boys.

REFERENCES

Anderson, Michael. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association* 103, pp. 1481-95.

Barnett, W. Steven, Epstein, Dale, Carolan, Megan, Fitzgerald, Jen, Ackerman, Debra, and Friedman, Allison. 2010. "The State of Preschool 2010." New Brunswick, N.J.: National Institute for Early Education Research.

Barnett, W. Steven and Massey, Leonard. 2007. "Comparative Benefit-Cost Analysis of the Abecedarian Program and its Policy Implications," *Economics of Education Review* 26, pp. 113-25.

Barnett, W. Steven, Hustedt, Jason, Robin, Kenneth, and Schulman, Karen. 2005. "The State of Preschool 2005." New Brunswick, N.J.: National Institute for Early Education Research.

Bloom, Howard S. and Charles Michalopoulos. 2010. "When is the Story in the Subgroups? Strategies for Interpreting and Reporting Intervention Effects for Subgroups." MDRC Working Paper. November. http://www.mdrc.org/publications/551/full.pdf

Campbell, Frances, Pungello, Elizabeth, Miller-Johnson, Shari, Burchinal, Margaret, and Ramey, Craig. 2001. "The Development of Cognitive and Academic Abilities: Growth Curves from an Early Childhood Educational Experiment," *Developmental Psychology* 37:2, pp. 231-42.

Campbell, Frances and Ramey, Craig. 2010. "Carolina Abecedarian Project." In Arthur Reynolds, Arthur Rolnick, Michelle Englund, and Judy Temple, eds., *Childhood*

Programs and Practices in the First Decade of Life (New York: Cambridge University Press), pp. 76-98;.

Clements, Douglas and Sarama, Julie. 2011. "Early Childhood Mathematics Intervention," *Science* 333 (August 19), pp. 968-70.

Community Service Council of Greater Tulsa. 2002. "Early Childhood Development." Tulsa: November 20, Video-tape.

Currie, Janet and Hyson, Rosemary. 1999. "Is the Impact of Health Shocks Cushioned by Socioeconomic Status? The Case of Low Birthweight" *American Economic Review* 89: 2, pp. 245-50.

Currie, Janet and Thomas, Duncan. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85 (June), pp. 341-64.

Currie, Janet and Thomas, Duncan. 1999. "Does Head Start Help Hispanic Children?" *Journal of Public Economics* 74 (November), pp. 235-62.

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111-134.

Duncan, Greg and Brooks-Gunn, Jeanne, eds. 1997. *Consequences of Growing Up Poor*. New York: Russell Sage Foundation.

Garces, Eliana, Thomas, Duncan, and Currie, Janet. 2002. "Longer-Term Effects of Head Start," *American Economic Review* 92 (September), pp. 999-1012.

Gormley, William and Gayer, Ted. 2005. "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program," *Journal of Human Resources* 40, pp. 533-58.

Gormley, William, Gayer, Ted, Phillips, Deborah, and Dawson, Brittany. 2005. ―The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41, pp. 872-84.

Gormley, William, Phillips, Deborah, Adelstein, Shirley, and Shaw, Catherine. 2010. ―Head Start's Comparative Advantage: Myth or Reality?" *Policy Studies Journal* (August), pp. 397-418.

Gormley, William, Phillips, Deborah, and Gayer, Ted. 2008."Preschool Programs Can Boost School Readiness," *Science* 320 (June), pp. 1723-24.

Guo, Ehenyang, and Mark W. Fraser. 2010. *Propensity Score Analysis: Statistical Methods and Applications* (Los Angeles: Sage Publications).

Heckman, James. 2000. ―Policies to Foster Human Capital," *Research in Economics* 54, pp. 3-56.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. ―Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605–54.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. ―A New Cost-Benefit and Rate of Return Analysis for the Perry Preschool Program: A Summary." In Arthur Reynolds et al., eds., *Childhood Programs and Practices in the First Decade of Life* (New York: Cambridge University Press), pp. 366-80.

Holland, Paul T. 1986. ―Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945-960.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. ―Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15: 199-236.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5-86.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45(4):915-943.

Little, Roderick J.A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data.* 2nd Ed. Hoboken, N.J.: John Wiley & Sons, Inc.

Ludwig, Jens and Miller, Douglas. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics* 122 (February), pp. 159-208.

Magnuson, Katherine, Ruhm, Christopher, and Waldfogel, Jane. 2007. "Does Prekindergarten Improve School Preparation and Performance?" *Economics of Education Review* 26, pp. 33-51.

Magnuson, Katherine. 2007. "Maternal Education and Children's Academic Achievement During Middle Childhood," *Developmental Psychology* 43: 6, pp. 1497-512.

McBride, Bill. 2009. "Teaching to Gender Differences: Boys Will Be Boys and Girls Will Be Girls," http://www.dedicatedteacher.com/samples/DDTr/ipe3526s.pdf.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9(4): 403-425.

McLanahan, Sara and Sandefur, Gary. 1994. *Growing Up with a Single Parent*. Cambridge, Mass.: Harvard University Press.

National Center for Education Statistics. 2010a. "The Nation's Report Card: Mathematics 2009," http://nces.ed.gov/nationsreportcard/pdf/stt2009/2010454OK4.pdf.

National Center for Education Statistics. 2010b. "The Nation's Report Card: Reading 2009," http://nces.ed.gov/nationsreportcard/pdf/stt2009/2010460OK4.pdf.

Phillips, Deborah, Gormley, William, and Lowenstein, Amy. 2009. "Inside the Pre-K Door: Classroom Climate and Instructional Time Allocation in Tulsa's Pre-K Program," *Early Childhood Research Quarterly 24,* pp. 213-28.

Puma, Michael, Bell, Stephen, Cook, Ronna, and Heid, Camilla. 2010. "Head Start Impact Study, Final Report." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families (January).

Puma, Michael, Bell, Stephen, Cook, Ronna, Heid, Camilla, and Lopez, Michael. 2005. "Head Start Impact Study: First Year Findings." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families.

Reynolds, Arthur. 1994. "Effects of a Preschool Plus Follow-On Intervention for Children at Risk," *Developmental Psychology* 30: 6, pp. 787-804.

Reynolds, Arthur, Temple, Judy, Ou, Suh-Ruu, Arteaga, Irma, and White, Barry. 2011. "School-Based Early Childhood Education and Age-28 Well-Being: Effects by Timing, Dosage, and Subgroups," *Science* 333 (July 25), pp. 360-64.

Reynolds, Arthur, Temple, Judy, Robertson, Dylan, and Mann, Emily. 2002. "Age 21 Cost-Benefit Analysis of the Title I Chicago Child-Parent Centers," *Educational Evaluation and Policy Analysis* 24: 4 (Winter), pp. 267-303.

Reynolds, Arthur, Temple, Judy, Robertson, Dylan, and Mann, Emily. 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: 18 (May 9), pp. 2339-46.

Rosenbaum, Paul R. 1973. ―Matching to Remove Bias in Observational Studies.' Biometrics. 29: 159-184.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. ―The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. ―Constructing a Control Group Using Multivariate matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39:33-38.

Royston, Patrick, John B. Carlin, and Ian R. White. 2009. ―Multiple Imputation for Missing Values: New Features for mim." *The Stata Journal* 9: 252-64.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Sammons, Pam. 2010. ―Do the Benefits of Pre-school Last?" In Kathy Sylva, Edward Melhuish, Pam Sammons, Siraj-Blatchford, Iram, and Taggart, Brenda, eds., *Early Childhood Matters: Evidence from the Effective Pre-school and Primary Education Project*. New York: Routledge, pp. 114-48.

Schonlau, Martin. 2005. ―Boosted Regression (Boosting): An Introductory Tutorial and a Stata Plugin." *The Stata Journal* 5(3): 330-354.

Schweinhart, Lawrence, Montie, Jeanne, Xiang, Zonping, Barnett, W. Steven, Belfield, Clive, and Nores, Milagros. 2005. ―Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40." Ypsilanti, Mich.: High/Scope Educational Research Foundation.

Shonkoff, Jack and Phillips, Deborah, eds. 2000. *From Neurons to Neighborhoods*. Washington, D.C.: National Academy Press.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1): 1-21.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113(485): F3-F33.

Tulsa Public Schools. 2003. "Tulsa Counts! Tulsa Public Schools Mathematics Initiative." Tulsa: June.

U.S. Census Bureau. 2009. "Statistical Abstract of the United States." Washington, D.C.: U.S. Government Printing Office.

Westinghouse Learning Corporation (and Ohio University). 1969. "The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development." Athens: Ohio University, June 12.

Wong, Vivian, Cook, Thomas, Barnett, W. Steven, and Jung, Kwanghee . 2008. "An Effectiveness-Based Evaluation of Five State Pre-kindergarten Programs," *Journal of Policy Analysis and Management* 27, pp. 122-54.

TABLE 1
Full (Unmatched) Sample Characteristics for Early Cohort by Treatment and Year

Characteristic	*Kindergarten Year (2001-02)* TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)		*Third-Grade Year (2004-05)*[a] TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	
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	***
Sample Size	1425	1563		1038	961	

NOTES:

Difference of means between treatment and comparison groups significant at * $p < 0.1$;

** $p < 0.05$; *** $p < 0.01$

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.

TABLE 2
Full (Unmatched) Sample Characteristics for Late Cohort by Treatment and Year

Characteristic	*Kindergarten Year (2006-07)* TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)		*Third-Grade Year (2009-10)*[a] TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	
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	***
Sample Size	1565	1597		1087	937	

NOTES:

Difference of means between treatment and comparison groups significant at * $p < 0.1$;

** $p < 0.05$; *** $p < 0.01$

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 lunch status at Third Grade.

d. Late Cohort Missingness: 34 observations missing lunch status at Kindergarten. 8 observations missing lunch status at Third Grade.

TABLE 3
Estimated 3rd-Grade Test Score Impacts of TPS Pre-K Program, by Method, Cohort, and Subgroup

	Difference of Means	Std. Err.		Regression-Adjusted (Full Comparison Sample) with Non-Imputed Data[a, b, c]	Std. Err.		Regression-Adjusted (Full Comparison Sample) with Imputed Data[a, b, c]	Std. Err.		Propensity-Score Matched[d, e, f]	Std. Err.	
EARLY COHORT (3rd grade in 2004-05)												
Reading OPI												
Full Sample	-14.09	3.90	***	-3.21	3.87		-2.89	3.73		-3.49	6.71	
Sample Size	*1998*			*1853*			*1998*			*1528*		
Math OPI												
Full Sample	-10.20	3.83	***	0.15	3.83		1.78	3.67		1.16	6.40	
Sample Size	*1996*			*1852*			*1996*			*1525*		
LATE COHORT (3rd grade in 2009-10)												
Reading OPI												
Full Sample	-1.17	4.08		7.87	4.75	*	8.41	3.77	**	8.60	6.52	
Sample Size	*1985*			*1162*			*1985*			*1548*		
Males	0.20	6.06		11.61	6.80	*	9.34	5.50	*	8.08	9.48	
Females	-1.11	5.47		4.40	6.67		7.45	5.11		5.16	8.78	
White, Nonhispanic	2.11	5.96		7.62	6.63		1.71	5.71		-2.36	10.52	
Black, Nonhispanic	6.45	7.40		-4.25	10.71		8.06	7.30		7.51	11.65	[g]
Hispanic	23.87	8.97	***	15.73	10.93		20.86	8.73	**	20.30	14.28	[g]
Native American	12.26	11.64		21.47	17.50		14.06	11.22		6.37	18.88	[g]
Free Lunch	4.48	5.02		12.36	6.91	*	10.76	5.00	**	12.48	8.28	
Reduced Lunch	-1.20	10.73		2.86	13.56		2.58	10.46		-3.54	16.19	[g]
Paid Lunch	0.93	7.05		5.69	7.63		6.54	6.72		10.62	12.14	[g]

Math OPI													
Full Sample	6.52	4.37		15.31	5.54	***	14.63	4.13	***	17.88	7.46	**	
Sample Size	*2015*			*1179*			*2015*			*1574*			
Males	8.49	6.41		22.47	7.78	***	17.70	5.88	***	21.30	9.49	**	
Females	3.71	5.99		8.43	7.89		12.06	5.75	**	11.53	9.22		
White, Nonhispanic	10.97	6.61	*	16.29	8.02	**	9.88	6.54		11.15	12.48		
Black, Nonhispanic	11.06	7.82		1.88	11.23		11.53	7.73		13.75	12.36	[g]	
Hispanic	22.51	9.57	**	10.50	13.48		19.08	9.43	**	23.58	16.93		
Native American	39.95	13.11	***	42.14	20.04	**	34.72	13.35	**	30.74	25.54	[g]	
Free Lunch	10.48	5.47	*	19.59	8.23	**	16.88	5.42	***	19.31	8.81	**	
Reduced Lunch	4.34	12.18		1.70	15.17		8.91	11.73		7.83	17.27	[g]	
Paid Lunch	11.94	7.69		16.07	8.73	*	13.88	7.56	*	23.25	13.26	*	[g]

NOTES:

Statistical significance indicated as * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

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. Robust

standard errors were used.

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.

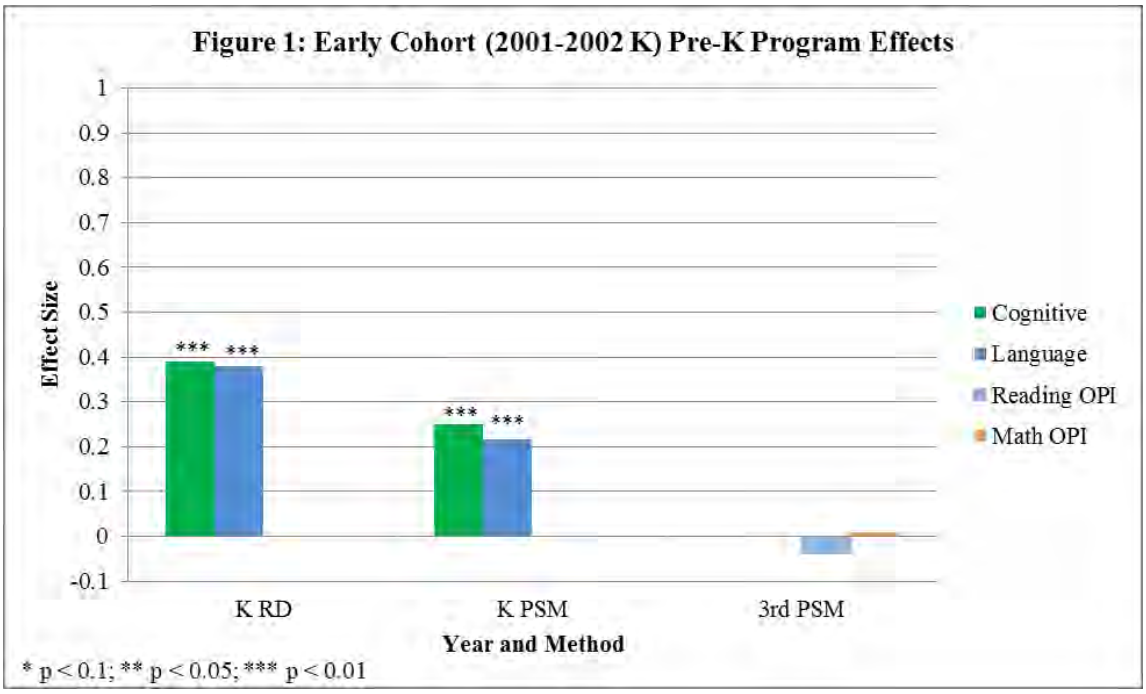
Figure 1: Early Cohort (2001-2002 K) Pre-K Program Effects
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
Effect Size
*** ***
*** ***
K RD
K PSM
3rd PSM
Year and Method
Cognitive
Language
Reading OPI
Math OPI
* p < 0.1; ** p < 0.05; *** p < 0.01

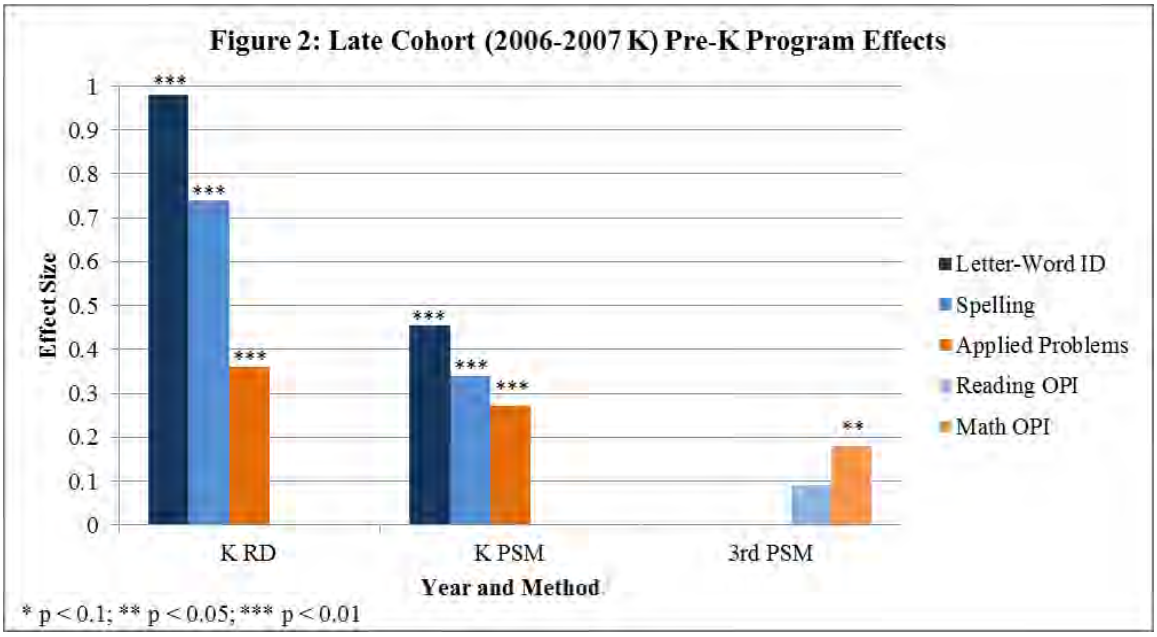
Figure 2: Late Cohort (2006-2007 K) Pre-K Program Effects
Effect Size
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
***
***
***
***
***
***
**
K RD
K PSM
3rd PSM
Year and Method
Letter-Word ID
Spelling
Applied Problems
Reading OPI
Math OPI
* p < 0.1; ** p < 0.05; *** p < 0.01

TABLE B1
Sample Restrictions and Test Score Data Source, By Cohort

Cohort and Restriction	Number Dropped	Sample Size
EARLY COHORT		
Original Sample		3,814
Eligible Birthday (>01sep1995 & <=01sep1996)	367	3,447
Not in Head Start	435	3,012
Have sufficient data (24 obs missing nearly all variables)	24	2,988
Have at least one 3rd grade test score (Reading or Math OPI)	989	**1,999**
# students in TPS 3rd grade: 1,373		
# students elsewhere in Oklahoma 3rd grade: 626		
LATE COHORT		
Original Sample		4,114
Eligible Birthday (>01sep2000 & <=01sep2001)	483	3,631
Not in Head Start	469	3,162
Have at least one 3rd grade test score (Reading or Math OPI)	1,138	**2,024**
# students in TPS 3rd grade: 1,610		
# students elsewhere in Oklahoma 3rd grade: 414		

NOTES:

a. Of the final sample of 1,999, there were 1,038 treatment observations and 961 comparisons.

b. Of the final sample of 2,024, there were 1,087 treatment observations and 937 comparisons.

TABLE B2
Full (Unmatched) Sample Characteristics for Early Cohort by Treatment and Year (Parent Survey Data)

Characteristic	*Kindergarten Year (2001-02)* TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)		*Third-Grade Year (2004-05)*[a] TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	
Mother's Education (%)						
No HS	16.5	19.3		16.3	19.5	
HS Grad	17.1	18.6		17.1	18.7	
Some college	50.3	39.9	***	50.7	39.7	***
College up	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*	
Sample Size	1425	1563		1038	961	

NOTES:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

a. 3rd grade analysis sample includes children with at least one OPI score.

TABLE B3
Full (Unmatched) Sample Characteristics for Late Cohort by Treatment and Year (Parent Survey Data)

Characteristic	*Kindergarten Year (2006-07)* TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)		*Third-Grade Year (2009-10)*[a] TPS Pre-K (Treatment Group)	No TPSPK or HS (Comparison Group)	
Mother's Education (%)						
No HS	18.8	17.7		15.5	14.2	
HS Grad	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*	
Sample Size	1565	1597		1087	937	

NOTES:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

a. 3rd grade analysis sample includes children with at least one OPI score.

NOTES

[1] We have documented levels of instructional support and emotional support in the Tulsa pre-K program that exceed those of school-based pre-K programs in other states (Phillips et al. 2009).

[2] Between 1990 and 1998, Oklahoma funded pre-K education for children who also were eligible for Head Start, and allowed ineligible children to enroll for a fee when space was available.

[3] In the 2000-01 school year, pre-K programs were offered in 45 of 59 TPS elementary schools; in 2005-06 they were offered in 50 of 58 TPS elementary schools. Students were free to attend pre-K programs outside their neighborhood, and busing was often available.

[4] In April 2005, third-grade teachers sent home with each child a one-page survey (in both English and Spanish) and a letter to his/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.

[5] Ideally, this information would have been collected about the children's families prior to their fourth birthday; however, this was not feasible. The type of information collected is not likely to change much during the period examined, and it is not likely to be affected by the presence or absence of treatment (the child's enrollment in the TPS pre-K program).

[6] For the early cohort, 73 percent of treatment students were matched with a third grade test, compared with 62 percent of comparison students. For the late cohort, the match rates were 70 percent and 59 percent, respectively.

[7] A regression discontinuity design (RDD) was used in previous pre-K research by Gormley & Gayer (2005), Wong et al. (2008), and Gormley et al. (2008). Estimates of the pre-K program were obtained by comparing children who had finished pre-K (and were just starting kindergarten), with children who had not yet been exposed to 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. The RDD cannot be used to estimate longer-term impacts because both groups of children were eventually exposed to pre-K.

[8] Our typical boosted model specified 5 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 those recommended by Schonlau (2005).

[9] 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 in Footnote # 6 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.

[10] The coefficient for girls is statistically significant at the $p<0.05$ level when using propensity score weights.

[11] The effect size for black girls is -0.08, and not statistically significant. Results on other race*gender subgroups are available from the authors.

[12] Supplementary calculations based on analysis of Gormley et al. 2010.