The Effects of Tulsa's Pre-K Program on Middle School Student Performance

William T. Gormley, Jr.
Deborah Phillips
Sara Anderson

Abstract

As states have upgraded their commitment to pre-K education over the past two decades, questions have arisen. Critics argue that program effects are likely to fade out or disappear over time, while supporters contend that program effects are likely to persist under certain conditions. Using data from Tulsa Public Schools, three neighboring school districts, and the state of Oklahoma, and propensity score weighting, we estimate the effects of Tulsa's universal, school-based pre-K program on multiple measures of academic progress for middle school students. We find enduring effects on math achievement test scores, enrollment in honors courses, and grade retention for students as a whole, and similar effects for certain subgroups. We conclude that some positive effects of a high-quality pre-K program are discernible as late as middle school.

INTRODUCTION

In recent years, the debate over the efficacy of state-funded pre-kindergarten (pre-K) programs has shifted from school readiness to longer-term effects (Bailey et al., 2017; Duncan & Magnuson, 2013). A key question is whether short-term positive effects on cognitive test scores, which have now been amply documented, persist or fade out over time. In general, taxpayers and public officials would like some assurance that investments in pre-K have enduring impacts that can justify these expenditures.

In this paper, we examine the effects of Tulsa's school-based pre-K program on middle-school children. This study is the first to evaluate the effects of a universal, public school pre-K program on middle school outcomes. Building on previous work (Gormley et al., 2005; Gormley, Phillips, & Gayer, 2008), we ask whether short-term impacts on early literacy and early math skills endure over time. We focus on several measures of academic progress: performance on standardized tests, GPA, enrollment in either a gifted program or honors courses; grade retention, special education placement, absenteeism, and suspensions.

The Tulsa pre-K program is of special interest because it is relatively mature, reaches a relatively large percentage of four-year-olds, and is of relatively high quality. The Tulsa pre-K program has also figured prominently in national debates over the merits of universal pre-K, because the program has been studied in depth over more than a decade, because Oklahoma was the second state in the nation to adopt a universal pre-K program, and because President Obama mentioned Oklahoma explicitly in his 2013 State of the Union remarks endorsing universal pre-K as a national policy.

Journal of Policy Analysis and Management, Vol. 00, No. 0, 1–25 (2017)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22023

From a methodological perspective, our study, though nonexperimental, has several strengths. First, in seeking contemporary data for Tulsa Public School students who entered kindergarten in the fall of 2006, we have incorporated data from three additional school districts in the Tulsa metropolitan area and, for some outcomes, all public schools in the state of Oklahoma, enabling us to reduce sample attrition. Second, we are able to draw comparison group students from the same local population. Third, we take advantage of a parent survey, conducted in the fall of 2006, to assemble an unusually rich array of covariates. Fourth, we conduct a number of sensitivity tests, subjecting our conclusions to an extensive battery of validity checks. These techniques, when used in tandem, have considerable potential to reduce selection bias and enhance validity (Wong, Valentine, & Miller-Baine, 2017).

As predicted by pre-K program critics, we find evidence of diminishing program impacts on standardized test scores over time. However, as predicted by pre-K program supporters, we also find evidence of program impacts on math test scores, enrollment in honors courses, and grade retention. These findings remain intact after several robustness checks. We begin by reviewing a growing body of literature on longer-term pre-K effects. We articulate some expectations for our Tulsa pre-K inquiry and explain the nuances of our methodology, which relies primarily on propensity score weighting. We then present and discuss our findings.

LITERATURE REVIEW

A mounting body of evidence suggests that participation in state-funded public pre-K programs can boost the cognitive skills of children at school entry, sometimes dramatically (Phillips et al., 2017). Studies of universal pre-K programs in Georgia (Henry et al., 2004), Oklahoma (Gormley, Phillips, & Gayer, 2008), and Boston (Weiland & Yoshikawa, 2013), and targeted pre-K programs in New Jersey (Frede et al., 2007), New Mexico (Hustedt et al., 2009), and North Carolina (Peisner-Feinberg & Schaaf, 2011) have documented impressive short-term gains in cognitive growth for students who attended pre-K programs in these jurisdictions. A multi-state study (Wong et al., 2008) reached similar conclusions. Several of these studies use a regression discontinuity design, which offers better safeguards against selection bias than conventional regression-based research designs (Lee & Lemieux, 2010).

But the question of whether these immediate impacts of state pre-K programs are short-lived or sustained remains unsettled. Earlier studies of some highly celebrated pre-K programs in Ypsilanti, MI, Chapel Hill, NC, and Chicago, IL provide evidence that high-quality early childhood programs can have lasting effects on educational attainment, socio-emotional indicators, health, and crime decades after the intervention (Campbell et al., 2012; Reynolds et al., 2011; Schweinhart et al., 2005). Whether these results can be generalized to state pre-K programs is uncertain either because the intervention was unusual in intensity or because it occurred years ago, when counterfactual circumstances were quite different. Also, these programs exclusively served low-income students.

Evidence on the long-term effects of Head Start is mixed. Using techniques such as sibling comparisons and regression discontinuity, several studies of Head Start have reached positive conclusions about longer-term impacts (Currie & Rossin-Slater, 2015; Currie & Thomas, 1995; Deming, 2009; Ludwig & Miller, 2007). However, a national, randomized control trial of multiple Head Start programs found that initial positive effects faded quickly over time (Administration for Children and Families, 2010, 2012). Programmatic differences also raise questions about the implications of findings on Head Start for universal pre-K programs, which, unlike Head Start, serve children from diverse socioeconomic backgrounds.

Recent evidence on the longer-term effects of pre-K programs is also mixed, though tending towards positive conclusions. Several studies, using quasi-experimental or non-experimental methods, have assessed universal programs in Tulsa, Georgia, and Florida and targeted programs in Texas, New Jersey, and North Carolina. These studies, focusing on elementary school academic outcomes, have found persistent advantages for pre-K alumni in states with universal programs (Bassok & Miller, 2014; Fitzpatrick, 2008; Hill, Gormley, & Adelstein, 2015) and targeted programs (Andrews, Jargowsky, & Kuhne, 2012; Barnett et al., 2013; Dodge et al., 2017; Peisner-Feinberg & Schaaf, 2010). In contrast, evidence from a randomized control trial examining Tennessee's targeted pre-K program found that positive effects at the end of pre-K disappeared after one year, and estimated effects as of third grade were in some cases negative (Lipsey, Farran, & Hofer, 2015). A literature review by a diverse group of early childhood scholars recently concluded: "The available evidence about the long-term effects of state pre-K programs offers some promising potential but is not yet sufficient to support confident overall and general conclusions about long-term effects" (Phillips et al., 2017, p. 27).

THE TULSA PRE-K PROGRAM

In 1998, Oklahoma established the nation's second universal pre-K program, available to all four-year-old children, irrespective of income. That program is administered by the state's school districts, which provide pre-K services directly or form partnerships with other providers, such as local Head Start programs. This arrangement is different from the more common "mixed services delivery" model that one finds in states such as Georgia, New York, North Carolina, Florida, and Tennessee, where day care centers, public schools, private schools, and Head Start programs receive funds directly from the state and provide services to four-year-olds. Under state law, all state-funded pre-K programs in Oklahoma must maintain high quality standards, as measured by specific inputs: All teachers must have a B.A. degree and must be early childhood certified; and child/staff ratios of 10 to one must be maintained.

The Tulsa Public Schools (TPS) pre-K program, like other Oklahoma pre-K programs, adheres to these standards. We know from extensive, systematic classroom observations of virtually every TPS pre-K classroom, using the Classroom Assessment Scoring System or CLASS (Pianta, LaParo, & Hamre, 2008), that the quality of early childhood education in Tulsa pre-K classrooms was relatively high, at least during the 2005/2006 school year (Phillips, Gormley, & Lowenstein, 2009). For example, Tulsa's pre-K classrooms received higher scores for instructional learning formats (4.65 vs. 3.81), concept development (2.84 vs. 1.90), and quality of feedback (3.35 vs. 1.89), compared to school-based pre-K classrooms in 11 other states (Phillips, Gormley, & Lowenstein, 2009). Tulsa's pre-K teachers also allocated considerable time to academic subjects (Phillips, Gormley, & Lowenstein, 2009).

In contrast to other state-funded pre-K programs that have been evaluated in recent years, the Tulsa pre-K program is relatively mature, with a relatively high penetration rate. At the time of our study, the Tulsa pre-K program was eight years old and Oklahoma enrolled 68 percent of all four-year-olds. As argued below, we believe that these contextual factors may be important. An older program is less subject to start-up difficulties than an embryonic program, and a program with a relatively high take-up rate may help to convince elementary school teachers to ratchet up their pedagogy, with positive consequences for students generally and especially those students who, thanks to pre-K, are better prepared for more advanced elementary school content.

Tulsa is a particularly good venue for research on early childhood education because it is a relatively large school district and because its students come from diverse racial and ethnic backgrounds. At the time of our study, the TPS school system enrolled more students (40,729) than any other school district in Oklahoma. It also had a very diverse student body with respect to race-ethnicity, family income, and home language.

EXPECTATIONS

Based on our review of the literature, we have several expectations as we conduct our empirical investigation.

Test Scores and GPA

We expect some persistence of cognitive effects over time. The initial effect sizes for the Tulsa pre-K program were moderate to large, ranging from .36 to .98, when estimated using a regression discontinuity design (Gormley, Phillips, & Gayer, 2008), and from .27 to .45, when estimated using a propensity score matching design (Hill, Gormley, & Adelstein, 2015). Therefore, we anticipate bigger standardized test score differences between treatment and control group students in reading and math at grade seven than one might find for a pre-K program that produced smaller initial differences (Duncan & Magnuson, 2013). Because early reading and especially early math skills are thought to be fundamental and transferable (Bailey et al., 2017; Claessens & Engel, 2013; Sarama et al., 2012), we are also inclined to predict superior grade point averages for pre-K participants.

Grade Retention and Special Education

Because pre-K boosted academic skills, at least in the short run, it should promote academic progress, which should reduce grade retention. Similarly, pre-K should reduce the need for special education, at least for children with relatively minor learning disabilities. The nascent literature on longer term pre-K effects supports both of these propositions (e.g., Dodge et al., 2017; Muschkin, Ladd, & Dodge, 2015), and a study using ECLS-K data finds a negative relationship between early math skills and grade retention through eighth grade (Claessens & Engel, 2013).

School Attendance and Suspensions

We might expect pre-K to have positive long-term effects on school attendance by reducing the likelihood that a student will struggle academically (e.g., Connolly & Olson, 2012), which can discourage attendance. Also, we might expect pre-K's positive impact on academic achievement to lower school suspension rates by reducing the inclination to "act out" in school. This would be consistent with prior evidence of less adult crime among pre-K alumni (Heckman et al., 2010; Reynolds et al., 2011; Schweinhart et al., 2005). On the other hand, school attendance and suspensions depend on a host of factors (Ehrlich et al., 2014).

METHODS

To assess long-term impacts in the absence of random assignment, we selected a propensity score weighting approach. In prior research, focusing on short-term impacts on this cohort of students, we used a regression discontinuity design, in which we compared pre-K alumni entering kindergarten with children entering pre-K at the same time (Gormley, Phillips, & Gayer, 2008), after controlling for age and other demographic variables. At this point, nearly a decade later, all students in the regression discontinuity sample have received pre-K, thus requiring us to find a different control group. In a recent paper (Hill, Gormley, & Adelstein, 2015), we estimated the effects of the TPS pre-K program on standardized test scores as of third grade, using propensity score matching. Since that time, some students have exited the TPS system and migrated to a) another state; b) a private school or charter school; or c) another school district. As a partial solution, we determined the three adjacent school districts (Union, Broken Arrow, and Jenks) that are the most common landing areas for students who depart from TPS and have included these school districts in our study. For state standardized test scores and grade retention, we also were able to access *state* administrative data, which strengthens our analysis for these important outcomes by increasing our sample size and by reducing the likelihood of differential attrition.

Study Sample

We begin with a pool of 4,033 students who were in kindergarten at TPS in the fall of 2006; this included children previously in pre-K (40 percent), Head Start (11 percent), or neither (49 percent).[1] For the middle school follow-up, we accessed data in two ways. We first relied on TPS and the three neighboring districts to collect administrative data on students. Of the original sample, 2,269 students remained in one of the four districts (56 percent). We then used statewide administrative data to track additional students, for standardized test scores and grade retention only. We tracked 3,045 student records from our original sample across the state (76 percent).

We excluded Head Start alumni from our present study, using data from the Community Action Project of Tulsa County. We regard Tulsa Head Start students as their own treatment group and have reported on their middle school outcomes in another paper (Phillips, Gormley, & Anderson, 2016). We excluded 359 Head Start alumni when analyzing state data and 277 Head Start alumni when analyzing local data. (There were a total of 428 Head Start students in the original sample.)

Our matching strategy resulted in a final pool of 1,992 students for our Tulsa area sample (TPS + 3 school districts), which applies to all outcomes other than standardized test scores and grade retention. When analyzing the latter, with state administrative data, our working sample consists of 2,656 students. These figures, presented in Table 1, represent the maximum number of student records matched; however, not all students had records on all variables for a variety of reasons, so tables in outcome analyses sometimes reflect a smaller sample size.[2]

MEASURES

This study uses data from three sources: (1) state/district administrative data from 2006/2007 and 2013/2014 for children enrolled in TPS and the three neighboring districts; (2) parent survey data from children enrolled in TPS as collected in August 2006; and (3) U.S. Census data. We also used data from 2014/2015 for more limited purposes, as described below.

[1] The mix of students enrolled in K in 2006/2007 should not be used to infer a pre-K penetration rate for 2005/2006, because some TPS pre-K students did not attend TPS K the following year, while some K students who did not attend TPS pre-K were ineligible to attend pre-K because they lived elsewhere at the time.
[2] The student retention percentage figures in Table 1 (55 percent and 73 percent) differ somewhat from those cited above (56 percent and 76 percent), because Table 1 excludes Head Start students.

Table 1. TPS, TPS and vicinity, and state samples by treatment at middle school follow up, excluding Head Start alumni.

	TPS		TPS + 3 districts		State	
	# of observations	% in group	# of observations	% in group	# of observations	% in group
Treatment group	838	51.3%	991	49.7%	1,278	47.6%
Control group	797	48.7%	1,001	50.3%	1,378	52.4%
Total sample size	1,635	100%	1,992	100%	2,656	100%
% of original cohort, excluding Head Start participants ($N = 3605$)	45%		55%		73%	

Treatment

We define pre-K participation based on enrollment in pre-K in 2005/2006 and on attendance using TPS administrative records. To be included in our treatment group, students must have attended pre-K for at least 50 percent of the academic year (90 days or more). The comparison group was thus youth who were not in pre-K or in Head Start (though they could have attended less than 50 percent of the days). A portion of students ($N = 293$) enrolled in TPS pre-K for fewer than 50 percent of the days of the school year, and these students remain in our comparison group, as do students who were in Head Start for less than 50 percent of the school days.

Independent Variables

TPS staff provided administrative data for each child enrolled in TPS K during the 2006/2007 academic year. From administrative records, information was available on: the child's TPS pre-K program participation, school attended, date of birth, race/ethnicity, gender, and school lunch status. We also used TPS administrative data to designate the number of siblings each student had and whether the student was the oldest sibling.

The parent survey was administered in August 2006, while children were taking cognitive development tests at school registration. This two-page survey contained questions about the child's previous preschool experience, parental marital status, whether the child currently lived with his or her biological father, the highest level of education attained by each parent, the primary language spoken at home, the child's place of birth, the availability of internet access at home, whether or not the child had health insurance, and the perceived health status of the child. The overall response rate was approximately 64 percent.

Finally, we obtained the block group median income, representing neighborhood economic resources, from when the children were in K as an additional covariate. We also used information about Tulsa public housing units to designate whether a student was living in public housing in 2006/2007.

Table 2. Descriptive statistics for student outcomes by treatment status.

	Treatment					Control				
	N	Mean	SD	Min	Max	*N*	Mean	SD	Min	Max
State Math OCCT	1,256	695.05	89.07	400	990	1,380	690.04	92.96	400	990
State Reading OCCT	1281	704.69	83.57	400	990	1,403	700.81	87.16	400	990
GPA	942	2.75	0.79	0	4	937	2.74	0.83	0	4
Absences	929	9.48	9.38	0	75.5	952	9.89	9.82	0	85.5
In-school suspension	1,014	0.36	1.24	0	16	1,024	0.33	1.22	0	13
Out-of-school suspension	999	0.36	1.16	0	16	1,011	0.31	0.89	0	7
	N	Proportion				*N*	Proportion			
Honors	1,003	0.48				1,014	0.47			
Gifted	1,001	0.19				1,013	0.19			
Special Ed	1,019	0.23				1,036	0.26			
Repeat	1,360	0.23				1,505	0.32			
Chronic absenteeism	967	0.20				965	0.26			

Dependent Variables

TPS, neighboring district, and state data from the 2013/2014 academic year were used for our dependent variables, whose descriptive statistics appear in Table 2. We examined state standardized test score data as collected during the spring of 2014 (for most students, seventh grade; for students retained in grade once, sixth grade). Our data source, the Oklahoma Core Curriculum Test (OCCT), is a criterion-referenced state assessment administered annually. Prior longer-term research of this sample (Hill, Gormley, & Adelstein, 2015) was only able to use test score data from non-grade retained students. To address this limitation, we included in our analysis the 6th-grade records of seventh graders who were retained in grade once.[3] For the group of grade-retained students, we also obtained test scores from their 7th-grade year (2014/2015). We analyzed state test score data in two ways: 1) test scores for all students in the 2013/2014 academic year (our preferred approach); and 2) test scores for all 7th-grade students in either 2013/2014 or 2014/2015 (a robustness check).

We also obtained school administrative data on GPA, honors coursework, and enrollment in the gifted and talented program. The GPA was either provided from the relevant district or calculated from school transcript data for the 2013/2014 school year, including all grades from all courses. An indicator for whether the student enrolled in honors coursework was identified from names of courses on transcripts like "advanced" or "honors." Finally, whether a student had been identified as gifted and talented by the district was used as an outcome. Oklahoma has relatively clear standards for enrollment in gifted and talented education, primarily relying on scoring in the top 3 percent of the distribution on an intelligence test, but school principals have some discretion based on creativity and class performance.

[3] Only 18 identified students were retained more than one time in TPS or a proximal district. Seventeen were retained two times and one was retained three times. To avoid excessive delays, additional data requests, and data reconciliation challenges, we opted not to include them in our analyses.

We also obtained data for several other outcome variables. Special education status was based on having an active Individualized Education Plan (IEP). Most students with IEPs in this sample were designated as having a specific learning disability. Grade retention was defined by current grade, or being in seventh grade instead of eighth grade at the start of the 2014/2015 academic year. Total absences in the 2013/2014 school year were also examined. If the student missed more than 10 percent of the school year days (or 18 days in OK), he or she was characterized as chronically absent (Ehrlich et al., 2014). Finally, out of school and in school suspensions in the 2013/2014 school year were included as outcome variables. The most common reasons for suspensions were fighting and disruptive conduct.

Analytic Plan

To generate impact estimates, we used propensity score weighted multiple regression. We estimated the ATT (average treatment effect on the treated) rather than the ATE (average treatment effect) because pre-K is universally available but not mandatory in Oklahoma; current pre-K enrollments in Oklahoma are about as high as any "universal" pre-K program has gotten. Due to sample attrition, we also used multiple imputation to appropriately estimate missingness on covariates. We discuss propensity score estimation and balance, program impact estimating techniques, and strategies for dealing with attrition and missing data in turn.

Propensity Scores

To generate propensity scores, we calculated the probability that a given child would have attended pre-K, given observable characteristics. The Appendix contains a regression table (Table A1) representing the relation between the propensity score covariates and pre-K participation, but this table should be used only to understand relations between pre-K and the covariates and not as the final model because we used boosted regression to generate weights (and the complexity of that model is not amenable to presentation in a table).[4] As recommended by Stuart (2010), we used a comprehensive set of covariates to predict the probability that a student attended pre-K and used students' observed scores to obtain a predicted probability of attending pre-K. In practice, this meant that we included as many variables as possible in generating propensity scores; in contrast, we were more selective and parsimonious in choosing variables for our final weighted regression models (Table 3).

We used boosted logistic regression modeling techniques, which utilize a machine learning approach, to estimate the propensity scores as our primary analytic technique (McCaffrey, Ridgeway, & Morral, 2004). Estimating the ATT with propensity scores involves assigning the treated participants a weight of 1 and the control participants a weight equal to the predicted odds of being in a treatment case, (ρ_i / (1-ρ_i)) (Hirano, Imbens, & Ridder, 2003). This weighting strategy up-weights the comparison participants whose observed covariate values best match those of treatment participants and down-weights participants whose observed covariate values are *unlike* those of treated participants. Other algorithms for propensity score analysis exist (e.g., matching), but there is no consensus on the single best approach (Guo & Fraser, 2010; Stuart, 2010).

[4] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 3. Variables included as covariates and in propensity score models.

Variable	Covariates	Propensity score models
Race	Yes	Yes
Mother's marital status at K	Yes	Yes
Mother's education at K	Yes	Yes
Lunch status at K	Yes	Yes
Female	Yes	Yes
Internet access at home	Yes	Yes
Lives with father at K	Yes	Yes
Overage at K	Yes	Yes
In TPS at grade 8	Yes	Yes
Foreign born	No	Yes
Attended daycare at someone's home at age 3	No	Yes
Attended non-TPS pre-school at age 3	No	Yes
Attended Head Start at age 3	No	Yes
Attended some type of center based care at age 3	No	Yes
Insurance at K	No	Yes
Number of siblings	No	Yes
Oldest sibling	No	Yes
Public housing in K	No	Yes
Neighborhood median income (in thousands)	Yes	Yes

Propensity Score Balance

Our estimation approach focused on achieving the best covariate balance (Harder, Stuart, & Anthony, 2010), and weighting by the odds produces well-balanced groups. For example, we selected iterations, non-linearities, and interactions to optimize the model and minimize absolute standardized differences (ASD) between the treatment and control cases (the difference in means for each covariate divided by the pooled standard deviation). We did not use hypothesis testing to examine balance because it relies upon sample size and can be misleading, as hypothesis tests conflate changes in balance with changes in sample size (Stuart, 2010).

Table 4 provides the descriptive information on individual and family background characteristics used in propensity score models. Prior to applying weights (found in the first five columns), differences in treated and control participants were evident with regard to race, free lunch status, neighborhood median income, and child care history, as measured by the ASD. After the weights were applied to the control group, the differences between the treated and control participants decreased substantially (final three columns). Indeed, the ASD for 31 of 32 variables across imputed datasets was below the conservative threshold of .10, representing very balanced groups (Harder, Stuart, & Anthony, 2010; see Appendix Table A2).[5] We also examined ASDs for our subgroups. Although these were typically not as small as for the overall group, for no subgroup did they exceed .10 more than two times.

Regression Models

As the final analytic step, we use multiple regression, either OLS or logistic depending on the nature of the outcome, with school fixed effects, with weights and a subset

[5] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 4. Means, standard deviations (SD), and absolute standardized differences (ASD) for pre-K and comparison group characteristics used in propensity score models, with and without weights.

	Unweighted, baseline model					Propensity score weighted, analysis model		
	Tulsa Pre-K		Comparison			Final comparison sample		
Covariate	M	SD	M	SD	ASD	M	SD	ASD
Race (%) +								
White	32%	0.47	41%	0.49	0.19	34%	0.55	0.04
Black	37%	0.48	25%	0.43	0.26	34%	0.62	0.04
Hispanic	20%	0.40	20%	0.40	0.01	20%	0.48	0.01
Asian	1%	0.12	1%	0.12	0.00	1%	0.15	0.00
Native American	9%	0.29	12%	0.33	0.09	9%	0.31	0.00
Mother's marital status at K (%) +								
Never married	28%	0.60	29%	0.68	0.02	29%	0.67	0.01
Married	53%	0.64	46%	0.69	0.11	51%	0.73	0.02
Remarried	3%	0.23	4%	0.31	0.02	3%	0.22	0.01
Separated	5%	0.28	7%	0.38	0.06	6%	0.28	0.02
Divorced	9%	0.35	11%	0.48	0.07	9%	0.37	0.00
Widowed	2%	0.26	3%	0.28	0.01	3%	0.31	0.01
Mother's education at K (%) +								
No high school	20%	0.50	22%	0.57	0.03	21%	0.59	0.02
High school	28%	0.57	28%	0.67	0.00	28%	0.69	0.00
Some college	40%	0.58	37%	0.73	0.04	39%	0.72	0.01
College	12%	0.40	13%	0.42	0.01	12%	0.46	0.01
Lunch Status at K (%) +								
Free lunch	65%	0.48	65%	0.48	0.00	66%	0.58	0.01
Reduced price lunch	12%	0.33	10%	0.29	0.08	11%	0.40	0.02
Full price lunch	23%	0.42	25%	0.43	0.06	23%	0.49	0.00
Female (%) +	48%	0.50	47%	0.50	0.02	46%	0.63	0.03
Internet access at home (%) +	51%	0.59	43%	0.65	0.12	48%	0.70	0.04
Lives with father at K (%) +	58%	0.59	52%	0.63	0.11	56%	0.69	0.03
Overage at K +	1%	0.07	2%	0.16	0.17	2%	0.16	0.13
In TPS at grade 8+	63%	0.48	58%	0.49	0.10	62%	0.59	0.02
Foreign born (%)	21%	0.47	18%	0.50	0.06	19%	0.54	0.03

Table 4. Continued.

	Unweighted, baseline model					Propensity score weighted, analysis model		
	Tulsa Pre-K		Comparison			Final comparison sample		
Covariate	M	SD	M	SD	ASD	M	SD	ASD
Attended daycare at someone's home at age 3 (%)	20%	0.62	17%	0.53	0.06	19%	0.65	0.02
Attended non-TPS pre-school at age 3 (%)	13%	0.47	2%	0.57	0.21	14%	0.53	0.02
Attended Head Start at age 3 (%)	9%	0.29	2%	0.13	0.32	6%	0.46	0.08
Attended some type of center-based care at age 3 (%)	47%	0.73	50%	0.69	0.05	47%	0.76	0.00
Insurance at K (%)	89%	0.40	84%	0.50	0.11	88%	0.42	0.03
Number of siblings	1.19	1.60	1.24	1.71	0.03	1.22	1.83	0.02
Oldest sibling (%)	25%	0.47	31%	0.55	0.11	26%	0.53	0.03
Public housing in K (%)	2%	0.15	3%	0.15	0.03	3%	0.19	0.03
Neighborhood median income (in thousands) +	3.70	1.65	3.96	2.01	0.14	3.74	1.74	0.02

Note: Results combined across 40 imputed datasets. + = variables also used as covariates in regression outcome models.

of covariates as additional controls for doubly robust estimation (Bang & Robins, 2005; Duncan et al., 2007).[6] We employed the K school as a fixed effect because of possible differential sorting of students into schools based on pre-K attendance; children also typically attended the same school in K as pre-K. We controlled for K school, rather than middle school, because a child's elementary school experience lasts longer than a child's middle school experience. We were also concerned that controlling for middle school might inadvertently mean controlling for a choice that depends in part on academic success (e.g., enrollment in a competitive magnet middle school), which depends in part on pre-K enrollment. Finally, we ran models stratified by gender, free/reduced price lunch status, race, and ELL status.

Attrition and Missing Data

Over time, control group students exited from TPS at a higher rate than treatment group students. As of the fall of 2006, 45.5 percent of TPS kindergarten entrants were pre-K alumni, after excluding Head Start alumni. Had we continued to look only at active TPS students in the fall of 2014, 51.3 percent of our sample (after excluding Head Start alumni) would have been pre-K alumni. Instead, we looked at the Tulsa metropolitan area, where, after excluding Head Start alumni, 49.7 percent of our students were pre-K alumni, and the state of Oklahoma, where 47.6 percent of our students were pre-K alumni.

We examined differences between retained and attrited students for the Tulsa metro area (TPS + 3 school districts) and state samples to determine generalizability (see Table A3).[7] The students who were not enrolled in TPS or a neighboring school district in 2013/2014 possessed roughly similar demographic characteristics as those who were. A notable exception is racial/ethnic differences. Participants who could not be located in the 2013/2014 academic year were more likely to be white and Native American and less likely to be black and Hispanic than identified students. Also, their parents were significantly less likely to be married, born in another country, and to speak a language other than English at home. For students in the state sample, there were differences in race/ethnicity and the mother's educational achievement. Importantly, however, both groups demonstrated similar proportions of students receiving free, reduced, and paid lunch, and had similar scores on kindergarten assessments of Woodcock-Johnson tests for math and language skills. Because the middle school and kindergarten samples were not identical, we generated weights to minimize discrepancies between the middle school sample and our kindergarten sample (Reynolds et al., 2011). We discuss this below as an auxiliary technique.

We also examined several patterns of baseline, or covariate, missing data. For our treatment variable and several demographic variables, missing data were not an issue as the data were derived from complete administrative records. However, given the extent of missing data for the parent survey, we examined missingness on observable characteristics. Minority students in low-income neighborhoods and those who received a free lunch were less likely to have complete parent data. Parents of pre-K participants were somewhat more likely to fill out the survey than parents of non-participants. We also found small correlations between missingness on the parent survey and 7th-grade test scores (lower test scores for those with missing

[6] In some instances, perfect prediction of the outcome in fixed effects analyses precluded a logit estimate so linear probability models were run instead.
[7] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

data). Because we have these covariates, multiple imputation can help to rectify such problems (Rubin, 1987).

To cope with missing covariate data (we did not impute outcomes), we used multiple imputation to replace missing values of covariates used in our propensity score estimation and regression models generating program effects (Little & Rubin, 2014; Rubin, 1987). Multiple imputation predicts missing values using other variables in the data set. Missing data on covariates from administrative data were rare and so values were not imputed for special education status, gender, race, English language learner status, and free lunch status. We imputed values for missing covariates with the Stata *mi estimate* program using imputation by chained equations to create 40 complete data sets based on observed data (StataCorp, 2011; White, Royson, & Wood, 2011). Each of these 40 data sets was analyzed individually, and the results were combined to produce our final parameter estimates and standard errors by adjusting for variability across imputed datasets (Rubin, 1987). There is relatively little consensus on how to cope with missing data in a propensity score framework, so we attempted other propensity score estimation strategies as well; we used the propensity score weight generated from the non-imputed dataset, which matched on covariate missingness. We generated propensity score and outcome calculations with data from middle school students after excluding attrited students.

RESULTS

We report findings below for students overall and for certain subgroups, sorted by gender, race/ethnicity, school lunch eligibility, and ELL status.

Weighted Multiple Regressions

The results from propensity score weighted multiple regression estimates are found in Tables 5 and 6. All models include the covariates as shown and estimates were averaged across 40 multiply imputed datasets using Stata.[8] We report effect sizes within the text (regression coefficient divided by the comparison group standard deviation of the outcome for continuous outcomes; marginal differences for dichotomous outcomes) and unstandardized regression coefficients within the tables.

For students as a whole (see Table 5), we saw a statistically significant relationship between pre-K enrollment and standardized math test scores eight years later ($ES =$ .10). We also found a statistically significant relationship between pre-K enrollment and enrollment in an honors course eight years later (6 percentage points more pre-K children enrolled in honors courses than comparison youth). Pre-K enrollment also reduced the likelihood of grade retention. Pre-K was associated with a 7 percentage point reduction in grade retention, to 16 percent (the treatment group; covariate adjusted) from an adjusted baseline of 23 percent (the comparison group). Estimated effects on reading test scores, letter grades, special education, designation as a gifted student, absenteeism, and suspensions are substantively small and not statistically distinguishable from zero.

We now turn to results across subgroups and focus on those that were significantly different from zero at the $p < .05$ level (see Table 6). The positive effects on enrollment in honors are more than twice as large for males (by 10 percentage points) as they are for females (by 4 percentage points). Female students who

[8] Occasionally, models could not be effectively run because in some sub-groups for some imputations, there were empty cells predicting the outcome, which resulted in some coefficients having no value, in which case results could not be combined across imputations.

Table 5. Unstandardized regression coefficients (with standard errors) for pre-K predicting outcomes for full sample.

Outcome	B (SE)
State Math OCCT	8.93*
	(3.72)
State Reading OCCT	5.50
	(3.41)
GPA	0.05
	(0.04)
Honors	0.34*
	(0.14)
Gifted	–0.00
	(0.02)
Special education	–0.17
	(0.13)
Repeat	–0.50***
	(0.12)
Number of absences	0.00
	(0.48)
Chronic absenteeism	–0.01
	(0.02)
In-school suspension	0.01
	(0.02)
Out-of-school suspension	–0.00
	(0.02)

Note: All models are weighted by propensity scores and with covariates as described. ***$p < 0.001$; *$p < 0.05$.

were enrolled in pre-K had marginally higher math test scores on the OCCT (*ES* = .11) and were marginally less likely to be designated as requiring special education services; coefficients for boys were of similar magnitude but smaller in size. Both groups demonstrated a significantly lower likelihood of being retained in grade.

Results by free-lunch status demonstrate a somewhat varied pattern. For free-lunch-eligible students, there was a statistically significant relationship between pre-K enrollment and 7th-grade math test scores (*ES* = .11) and between pre-K enrollment and enrollment in an honors course (9 percentage points more free lunch students enrolled). Pre-K enrollment also reduced the likelihood of grade retention for both free lunch (10 percentage points) and reduced price lunch (22 percentage points) students. For paid-lunch students, there was also a marginally significant relationship between pre-K enrollment and 7th-grade math test scores (*ES* = .16) and enrollment in an honors course. For paid-lunch students, there were marginally significant negative relations between pre-K enrollment and special education services, chronic absenteeism, and suspensions in school. It is important to note that although coefficients vary across groups stratified by free-lunch status, the size and magnitude of coefficients are similar; perhaps differences across models are due to smaller sample sizes and imprecision of estimates.

We focus next on English language learners (at any point in their schooling). ELL students demonstrated marginally higher 7th-grade math test scores (*ES* = .21; this relation was not evident among non-ELL students) and were more likely to have enrolled in an honors course (13 percentage points more ELL students enrolled in honors courses; a similar association was found for non-ELL students). English language learners who attended pre-K were approximately half as likely to

Table 6. Unstandardized regression coefficients (with standard errors) for pre-K predicting outcomes across subgroups.

Outcome	Male	Female	Free lunch	Reduced lunch	Paid lunch	ELL	Non-ELL	White	Black	Hispanic
State Math OCCT	8.46	9.56^{+}	9.29*	5.47	13.76^{+}	18.36^{+}	7.06	15.71*	–1.57	13.72
	(5.23)	(5.03)	(4.60)	(12.06)	(7.12)	(10.27)	(4.40)	(6.15)	(6.23)	(8.62)
State Reading OCCT	6.05	4.74	6.05	–3.06	8.54	13.32	3.89	5.34	–3.27	12.55^{+}
	(4.96)	(4.46)	(4.17)	(11.58)	(6.28)	(9.07)	(3.91)	(5.87)	(5.88)	(7.56)
GPA	0.06	0.04	0.06	–0.13	0.01	0.13	0.04	0.04	–0.04	0.11
	(0.06)	(0.05)	(0.05)	(0.11)	(0.08)	(0.09)	(0.04)	(0.07)	(0.06)	(0.08)
Honors	0.56**	0.22	0.37*	–0.01	0.52^{d}	0.14^{***,a}	0.06**	0.08	0.04^{a}	0.12^{***,a}
	(0.21)	(0.21)	(0.17)	(0.61)	(0.36)	(0.05)	(0.02)	(0.29)	(0.04)	(0.04)
Gifted	0.02	0.00	0.00	–0.01	0.01	0.05	0.00	0.00	0.02	0.01
	(0.22)	(0.03)	(0.02)	(0.07)	(0.05)	(0.04)	(0.02)	(0.04)	(0.03)	(0.03)
Special education	–0.06	–0.35^{+}	–0.08	–0.08^{a}	–0.07$^{+,a}$	–0.06^{a}	–0.03	–0.04^{a}	0.03	–0.04^{a}
	(0.17)	(0.22)	(0.16)	(0.06)	(0.04)	(0.05)	(0.02)	(0.04)	(0.04)	(0.05)
Repeat	–0.54***	–0.42*	–0.48***	–1.11^{*,b}	–0.42	–0.86^{**,c}	–0.60***	–0.43*	–0.52**	–0.83***
	(0.14)	(0.18)	(0.12)	(0.49)	(0.32)	(0.29)	(0.13)	(0.19)	(0.19)	(0.25)
Number of absences	–0.20	–0.16	0.25	1.13	–1.00	–1.66	0.00	–0.43	0.65	–1.61
	(0.67)	(0.73)	(0.66)	(1.16)	(0.68)	(1.05)	(0.48)	(0.81)	(0.88)	(0.98)
Chronic absenteeism	–0.03	–0.00	–0.00	0.06	–0.05^{+}	–0.07	–0.01	–0.02	0.01	–0.07
	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)	(0.05)	(0.02)	(0.04)	(0.03)	(0.04)
In-school suspension	0.11	0.00^{a}	0.17	0.05^{a}	–0.05$^{+,a}$	–0.07^{a}	0.01	–0.03^{a}	0.09^{a}	–0.06^{a}
	(0.20)	(0.02)	(0.17)	(0.05)	(0.03)	(0.04)	(0.02)	(0.03)	(0.04)	(0.04)
Out-of-school suspension	0.05	–0.02^{a}	0.01	0.04^{a}	–0.04^{a}	0.01^{a}	–0.00	–0.00^{a}	0.01^{a}	0.00^{a}
	(0.21)	(0.02)	(0.17)	(0.04)	(0.03)	(0.04)	(0.02)	(0.02)	(0.04)	(0.04)

Note: All results are weighted by subgroup propensity scores. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < .10$. Standard errors in parentheses. There is only one *N* for ELL students because ELL was derived from the district (TPS) sample.
[a]LPM coefficient instead of Logit.
[b]27 imputations.
[c]35 imputations.
[d]33 imputations.

be retained in grade as English language learners who did not, similar to non-ELL students.

Results also varied by race/ethnicity and were less positive for black students than for other students. For white students, there was a statistically significant relation between pre-K enrollment and 7th-grade math test scores ($ES = .17$). For Hispanic students, there was a marginally significant relationship between pre-K enrollment and 7th-grade reading test scores ($p < .10$) and a statistically significant relationship between pre-K enrollment and enrollment in an honors course (12 percentage points more). For whites, blacks, and Hispanics, there were statistically significant relations between pre-K enrollment and grade retention as well. No other results that met conventional thresholds of statistical significance ($p < .05$) were found.

ROBUSTNESS CHECKS

We attempted several other estimation strategies as robustness checks (see Table 7). Results, as discussed, are remarkably robust to alternative estimation strategies.[9]

As a first robustness check, we used attrition weights to produce a 2014 sample that more closely resembled our 2006 sample. This approach gave more weight to observations with demographic characteristics that were more common in 2006, less weight to observations with demographic characteristics that were less common in 2006 (Reynolds et al., 2011; Ridgeway et al., 2015). We calculated attrition weights based on the same characteristics that we used for our propensity score analysis. Instead of predicting pre-K attendance, we predicted the likelihood of attrition from 2006 to 2014. We achieved good balance across covariates using this weighting scheme. We then used the product of the propensity score weight and the attrition weight in our outcomes analysis, as described above. The results of this exercise yielded statistical significance levels that were strikingly similar to those reported above.

We next used a variety of alternative strategies to generate propensity score weights, none of which produced discernibly different regression results. First, we used propensity score weights from non-imputed data only, which matched on missingness. Second, we used a product of the attrition weight with matching on non-imputed data and the pre-K propensity score weight, also matched on missingness. Third, we excluded parent survey items from our analysis as covariates in both the generation of propensity score weights and regression analyses. We did so because the parent survey was administered at the beginning of kindergarten rather than at the beginning of pre-K. Thus, it is conceivable that some of the parent survey responses were influenced by pre-K participation, making them questionable candidates for control variables. When we used this trimmed version of our original model, we found statistically significant impacts of pre-K on reading and marginally significant impacts of pre-K on special education placement, in addition to the other significant results reported. However, we continue to prefer our more conservative estimates of program impact, because we believe that the parent survey provides valuable information about several important variables (maternal education, the presence of the biological father at home, the number of siblings, Internet access at home) and because we doubt that these particular variables were influenced by pre-K participation or that they changed significantly over one year.

As a fourth set of robustness checks, we employed alternative propensity score matching methods, including nearest neighbor matching and regression-adjusted

[9] LPM models were used for robustness checks for ease of interpretation for dichotomous outcomes.

Table 7. Robustness checks for full sample models.

Dependent variable	Attrition weight	M = 0 weight	M = 0 and attrit M = 0	Pre-K weight demos only	Pre-K wt. attrit wt. demos only	NN matching	IPTW matching	CEM	No fixed effects
OCCT Math	9.66^{*}	7.15^{*}	7.92^{*}	11.67^{**}	12.18^{**}	9.45^{*}	7.09^{*}	8.24^{*}	7.36^{*}
	(3.74)	(3.64)	(3.68)	(3.42)	(3.46)	(3.66)	(3.50)	(3.42)	(3.57)
OCCT Reading	6.55^{+}	4.68	5.87^{+}	7.82^{*}	7.85^{*}	6.25^{+}	4.28	4.63	5.35^{+}
	(3.50)	(3.45)	(3.55)	(3.25)	(3.27)	(3.47)	(3.28)	(3.20)	(3.25)
GPA	0.05	0.03	0.03	0.06	0.05	0.07^{+}	0.05	0.03	0.04
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Honors courses	0.06^{*}	0.05^{*}	0.05^{*}	0.06^{*}	0.05^{*}	0.04^{+}	0.04^{+}	0.25^{+}	0.06^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.13)	(0.02)
Gifted	0.00	0.00	0.00	0.01	0.00	0.00	0.00	–0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Special education	–0.03	–0.02	–0.03	–0.04	–0.02	–0.02	–0.03	–0.14	-0.04^{+}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.12)	(0.02)
Repeated grade	-0.09^{**}	-0.09^{**}	-0.09^{**}	-0.10^{**}	-0.08^{**}	-0.07^{**}	-0.08^{**}	-0.46^{***}	-0.09^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.10)	(0.02)
Number of absences	0.08	0.15	0.22	–0.30	–0.28	–0.13	0.02	–0.01	–0.07
	(0.51)	(0.47)	(0.50)	(0.45)	(0.48)	(0.48)	(0.45)	(0.47)	(0.48)
Chronic absenteeism	0.00	0.00	0.00	–0.01	–0.01	–0.02	–0.04	–0.00	–0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
In school suspension	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.04	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.14)	(0.02)
Out of school Suspension	0.00	0.00	0.00	–0.01	0.00	–0.01	–0.01	0.17	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.14)	(0.02)

Note: Attrition weight: original model with an attrition weight. M = 0 weight: applying weight from the non-imputed data to outcome analyses. M = 0 and attrit M = 0: attrition and pre-K weights from non-imputed data. Demos only: pre-K weight from demographic characteristics exclusively. NN matching: nearest neighbor matching. IPTW matching: inverse probability of treatment weighting with covariates. CEM: coarsened exact matching. No FE: typical model run without K school fixed effects. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.1$.

inverse-probability weighting (Guo & Fraser, 2010) with the STATA commands *teffects pscore* and *teffects ipwra*. These commands cannot be used with the multiply imputed data so we used a missing-indicator regression approach, as in a prior study with this sample (Jenkins et al., 2016). For both techniques, we rejected participants who were not within the range of common support, which resulted in dropping four participants. Once again, results were remarkably consistent with prior matching strategies. An exception was that the association between pre-K and honors course taking was marginal with both matching strategies. The demonstrated associations between pre-K and math OCCT scores and a lower odds of grade repetition were replicated.

Next, we used coarsened exact matching (CEM), which involves specifying several key variables up front, matching treatment and control group observations based on all of them, and then calculating program impact estimates while controlling for other covariates (Blackwell et al., 2009). An advantage of this strategy is that it matches exactly all the nominal-level variables deemed to be most important and matches approximately the interval-level variables deemed to be most important (through the use of strata). Because they are widely used as predictors of academic outcomes, we determined the most important variables to be gender, school lunch eligibility, race/ethnicity, and neighborhood income. Of these, only the last was stratified (by quintiles) before matching. CEM produced results that closely mirrored the results reported.

Finally we ran analyses without kindergarten school fixed effects. Once again, results were consistent with prior analyses.

In addition to the results reported in Table 7, we conducted a Lee bounds analysis, to determine which of our findings would hold up if we were to make certain extreme assumptions about missing data due to sample attrition (e.g., that students lost through attrition were either exceptionally good or exceptionally bad performers). As proposed by Lee (2009), this approach identifies a block of students who are less likely to attrite (either the treatment group or the control group) and then sequentially trims the upper and lower tails of the distribution (e.g., OCCT test scores) to equalize the number of students in both groups. Because we have more missing students in our control group, we trimmed our treatment group so that it equaled the control group in size. Specifically, we trimmed our treatment group observations by 11 or 12 percent at the state level (depending on the variable) and by 19 percent at the metropolitan level. We used the user-created command "leebounds" in STATA to calculate the bounds (Tauchmann, 2013), consistent with recommendations by Lee (2009). Even these extreme assumptions about missing data yield consistent statistically significant positive estimates for the impact of pre-K on math test scores and honors course taking, and statistically significant negative estimates for grade retention (see Table A4).[10]

DISCUSSION

Differences between pre-K alumni and a comparable group of non-alumni in middle school are statistically significant, for several outcomes. The positive effects of Tulsa's early childhood education program on standardized test scores diminish over time, but pre-K alumni continue to do better than other students in math. Pre-K alumni also excel, relatively speaking, in honors course enrollment and grade retention—outcomes that were not measurable at kindergarten entry. But are these

[10] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

differences substantively significant? Does it matter that pre-K alumni have somewhat higher math test scores in middle school, are somewhat more likely to enroll in honors courses in middle school, and are less likely to be retained in grade?

Although these questions should be more fully investigated, we do know that middle school math test scores help to predict later academic success (Duncan & Magnuson, 2011, online appendix, Table 3.a9; Silver, Saunders, & Zarate, 2008). We also know that students who take more advanced coursework in middle school, such as honors courses, are more likely to be college and career ready (ACT, 2008). More broadly, middle school academic performance is a proven predictor of later success (Hein, Smerdon, & Sambolt, 2013).

We also have hard evidence on grade retention, from Tulsa itself. By merging our Tulsa data on grade retention with NLSY data on grade retention, crime, and adult earnings, and ACS data on adult earnings in the Tulsa metropolitan area, we were able to estimate the benefits of pre-K participation. By combining this information with information on pre-K costs, we calculated a benefit/cost ratio: 2.1/1 (Bartik et al., 2017). This number captures the crime reduction and adult earnings increase benefits due to pre-K's effects on grade retention only. Positive effects mediated through other variables, such as improved math test scores, or effects on other outcomes, such as substance abuse, would presumably yield an even higher benefit/cost ratio.

Our subgroup analysis, across a range of outcomes, yields no simple conclusions about which groups of children benefit more from pre-K in the longer run than others. There may be a tendency for Hispanics to benefit somewhat more and for blacks to benefit somewhat less than others. What might explain this pattern of results? One possibility is that black students have more limited access to better schools and better teachers, because of where they live or for some other reason (Sass, et al., 2012; Stipek, 2004). Another possibility is that the counter-factual condition is more favorable for blacks than Hispanics, because the former are more likely to be enrolled in some type of early childhood education program, while the latter are more likely to stay at home. Whatever the reasons, we are troubled by the relative paucity of benefits for black middle school students who attended pre-K years earlier. Still, it is worth emphasizing that, as late as seventh grade, males and females, disadvantaged and middle-class students, and students from diverse racial and ethnic backgrounds all benefit from participation in Tulsa's high-quality pre-K program to some extent.

Can the success of the Tulsa pre-K program be replicated in other cities and states? Because it administers a universal program, with a very high pre-K penetration rate, Tulsa Public Schools have been obliged to hire a substantial number of highly-qualified school teachers. At the time of our study, TPS was doing this successfully, as measured by strong levels of instructional support in pre-K classrooms. Jurisdictions with stronger labor markets might find it more difficult to recruit and retain talented pre-K teachers.

The presence of a strong Head Start program in Tulsa probably helped TPS, by reducing the number of students to be served and by helping some three-year-olds to strengthen their skills before enrolling in TPS as four-year-olds (Gormley, Phillips, & Gayer, 2008; Jenkins et al., 2016). On the other hand, the Tulsa Head Start program also competed with TPS in the market for highly qualified pre-K teachers. Jurisdictions with weaker Head Start programs would pose less of a recruitment threat to school-based pre-K programs but would contribute less to the goal of school-ready kindergarten entrants. The other big advantage of a strong Head Start program, which improves school readiness for disadvantaged students, is that it makes it easier for elementary school teachers to cover more advanced material if they choose to do so.

Inevitably, one wonders why we find persistent pre-K effects in Tulsa, as late as seventh grade, when researchers found no positive pre-K effects in Tennessee, as early as first grade. One possibility is that the Tulsa pre-K program, at the time of our study, was superior in quality to the Tennessee pre-K program, at the time of that study. We can neither confirm nor dismiss that possibility, given reliance on different measures of quality in the two studies. We know that the Tulsa pre-K program offered higher quality early education experiences than other state pre-K programs, based on both CLASS and Early Academics Snapshot scores (Phillips, Gormley, & Lowenstein, 2009). Early Childhood Environment Rating Scale (ECERS) comparisons of the Tennessee pre-K program with several other pre-K programs suggest that the Tennessee program is not noticeably better or worse than average (Lipsey, Farran, & Durkin, 2010).

At the time of our study, the Tulsa pre-K program was relatively mature (eight years old), with a high penetration rate (66 percent). In contrast, the Tennessee pre-K program, when evaluated, was four to five years old, with a penetration rate of 21 to 22 percent. In general, a more mature pre-K program has had more time to learn from experience and improve, by refining curriculum and professional development, recruiting talented teachers, and increasing participation. Such changes may enhance program impacts on children in both the short run and the longer run. Additionally, a pre-K program with a higher penetration rate may trigger upgrades in elementary school pedagogy, with teachers focusing on more advanced math and reading content because pre-K alumni are better able to handle more difficult material. These changes could help to sustain learning gains over time by avoiding redundant instruction, which has been shown to limit educational gains for young children (Engel, Claessens, & Finch, 2013). A key question that we cannot resolve is whether Tulsa pre-K alumni and non-alumni benefit equally or differently from exposure to advanced material in elementary school (Claessens, Engel, & Curran, 2014).

In the future, it would be useful for researchers to examine several successive cohorts for a nascent state or local government pre-K program to see whether pre-K program quality improves, whether elementary school curriculum upgrades occur, and whether curriculum upgrades are linked to test score gains. Researchers should also investigate whether teachers differentiate instruction to more closely match the skill sets of students who did and did not attend pre-K and, if so, who benefits. As pre-K investigations become both broader and deeper, we will be better able to explain different findings in different states.

Limitations

Before closing, we should acknowledge some limitations of our analysis. First, like most longer-term studies, our research suffers from sample attrition, though less so for our test outcomes and grade retention outcomes than for other variables. Second, we cannot specify exactly when the test score gap between treatment group and control group children declined. Third, propensity score analysis is an imperfect mechanism for overcoming the absence of random assignment. Fortunately, an unusually rich parent survey enabled us to reduce selection bias by controlling for a number of demographic variables not always available to researchers.

One specific worry about studies like this is the possibility that parents of more eager learners might be more likely to enroll their child in pre-K, thus inflating program impact estimates. In a national study using ECLS data, Crosnoe et al. (2016) found little empirical support for this idea (the "enrichment elicitation" hypothesis). If anything, there was modest support for the opposite idea—the "compensatory elicitation" hypothesis. If parents sometimes take their child's characteristics into account when choosing between pre-K and other options for a four-year-old, they would seem to put a slower learner into pre-K more often, which could lead to a small underestimate of program impact when using nonexperimental research designs.

CONCLUSION

Some fade out in the wake of a program intervention is perhaps inevitable, whether that intervention is a job training program (Heckman, 1994), a juvenile justice program (Lipsey et al., 2010), or a pre-K program (Duncan & Magnuson, 2013). But there is a critical difference between effects that diminish and effects that disappear. In this paper, we have shown that the effects of the Tulsa pre-K program on students' academic success do not disappear by middle school. Or, to put it positively, short-term gains in math skills persisted over time and pre-K alumni were more likely to be enrolled in honors courses. Grade retention reductions were evident for students as a whole and for many subgroups as well, including blacks. However, black students experienced fewer long-lasting benefits than students from other racial and ethnic groups.

Overall, our findings parallel similar findings for the CAP of Tulsa County Head Start program, which also receives funding under Oklahoma's UPK program (Phillips, Gormley, & Anderson, 2016). Like the TPS pre-K program, CAP's Head Start program yields persistent gains in math and a reduction in grade retention. It is not associated with higher enrollment in honors courses, but it is associated with less chronic absenteeism.

The persistence of pre-K's positive impacts over an eight-year period is promising, even if one acknowledges that differences between treatment and control group children, as of middle school, are rather modest. Oklahoma is a very poor state, with poorly funded schools. Yet a high-quality pre-K program funded by the state has left an indelible imprint on students who participated in it. This warrants celebration, and it also warrants further exploration.

WILLIAM T. GORMLEY, JR. is a Professor at the McCourt School of Public Policy at Georgetown University, 305 Old North, Washington, DC 20057 (e-mail: gormleyw@georgetown.edu).

DEBORAH PHILLIPS is a Professor in the Department of Psychology at Georgetown University, 301F White-Gravenor, Washington, DC 20057 (e-mail: deborah.dap4@gmail.com).

SARA ANDERSON is an Assistant Professor in the Department of Learning Sciences and Human Development at West Virginia University, P.O. Box 6122, Morgantown, WV 26506-6122 (e-mail: sara.anderson@mail.wvu.edu).

ACKNOWLEDGMENTS

We are indebted to the Tulsa Public Schools, the Union School District, the Jenks School District, the Broken Arrow School District, and the Oklahoma Department of Education for granting us access to student data. We would like to thank Liz Stuart and Mike Bailey for valuable statistical advice and we gratefully acknowledge the help of some great research assistants: Karin Kitchens, Howard Huo, Yuqi Liao, Victoria Rosenboom, Jon Belford, Trellace Lawrimore, Billy Hoo, Esther Owolabi, Paula Ganga, Kristina Rodriguez, Jake Ford, and Brent Howell. The editor, co-editor, and three reviewers offered many constructive suggestions. We would also like to thank the Heising-Simons Foundation, the Foundation for Child Development, and the Stranahan Foundation for their generous financial support.

REFERENCES

ACT. (2008). The forgotten middle: Ensuring that all students are on track for college and career readiness before high school. Iowa City, IA: ACT.

Administration for Children and Families. (2010). Head Start impact study: Final report. Washington, DC.

Administration for Children and Families. (2012). Third grade follow-up to Head Start impact study. Washington, DC.

Andrews, R., Jargowsky, P., & Kuhne, K. (2012). The effects of Texas's targeted pre-kindergarten program on academic performance. Cambridge, MA: National Bureau of Economic Research.

Bailey, D. H., Duncan, G., Odgers, C., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10, 7–39.

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962–973.

Barnett, W. S., Jung, K., Youn, M., & Frede, E. (2013). Abbott preschool program long term effects study: Fifth grade follow-up. New Brunswick, NJ: National Institute for Early Education Research.

Bartik, T. J., Belford, J. A., Gormley, W. T., & Anderson, S. (2017). A benefit-cost analysis of the Tulsa universal pre-k program. Washington, DC: Center for Research on Children in the U.S., Georgetown University.

Bassok, D. & Miller, L. C. (2014). Do children benefit from widely-available public preschool? Evidence from Florida's Voluntary Prekindergarten Program. Albuquerque, NM: Proceedings from the Annual Meeting of APPAM.

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: Coarsened exact matching in Stata. Stata Journal, 9, 524–546.

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B., Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. Developmental Psychology, 48, 1033–1043.

Claessens, A. & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. Teachers College Record, 115, 1–29.

Claessens, A., Engel, M., & Curran, C. (2014). Academic content, student learning, and the persistence of preschool effects. American Educational Research Journal, 51, 403–434.

Connolly, F. & Olson, L. S. (2012). Early elementary performance and attendance in Baltimore city schools' pre-kindergarten and kindergarten. Baltimore, MD: Baltimore Education Research Consortium.

Crosnoe, R., Purtell, K. M., Davis-Kean, P., Ansari, A., & Benner, A. D. (2016). The selection of children from low-income families into preschool. Developmental Psychology, 52, 599–612.

Currie, J. & Rossin-Slater, M. (2015). Early-life origins of life-cycles well-being: Research and policy implications. Journal of Policy Analysis and Management, 34, 1–35.

Currie, J. & Thomas, D. (1995). Does Head Start make a difference? American Economic Review, 85, 341–364.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. American Economic Journal: Applied Economics, 1, 111–134.

Dodge, K. A., Bai, Y., Ladd, H., & Muschkin, C. (2017). Impact of statewide early childhood programs and policies on educational outcomes in elementary school. Child Development, 88, 996–1014.

Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A., Klebanov, P., . . . & Japel, C. (2007). School readiness and later achievement. Developmental Psychology, 43, 1428–1446.

Duncan, G. & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. Duncan & R. Murnane, (Eds.), Whither opportunity? Rising inequality, schools, and children's life chances. (pp. 47–69) New York, NY: Russell Sage.

Duncan, G. & Magnuson, K. (2013). Investing in preschool programs. Journal of Economic Perspectives, 27, 109–132.

Ehrlich, S., Gwynne, J., Pareja, A., & Allensworth, E. M. (2014). Preschool attendance in Chicago public schools. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (Mis)Alignment between mathematics instructional content and student knowledge in kindergarten. Educational Evaluation and Policy Analysis, 35, 157–178.

Fitzpatrick, M. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. B. E. Journal of Economic Analysis & Policy, 8, Article 46.

Frede, E., Jung, K., Barnett, W. S., Lamy, C., & Figueras, A. (2007). The Abbott preschool program longitudinal effects study (APPLES). New Brunswick, NJ: National Institute for Early Education Research.

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

Gormley, W., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. Science, 320, 1723–1724.

Guo, S. & Fraser, M. W. (2010). Propensity score analysis. Thousand Oaks, CA: SAGE Publications, Inc.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods, 15, 234–249.

Heckman, J. (1994). Is job training oversold? The Public Interest, 115, 91–115.

Heckman, J., Moon, S., Pinto, R., Savelyez, R., & Yavitz, A. (2010). The rate of return to the HighScope Perry preschool program. Journal of Public Economics, 94, 114–128.

Hein, V., Smerdon, B., & Sambolt, M. (2013). Predictors of postsecondary success. Washington, DC: American Institutes for Research.

Henry, G., Ponder, B. D., Rickman, D. K., Masburn, A. J., Henderson, L. W., & Gordon, C. S. (2004). An evaluation of the implementation of Georgia's pre-K program. Atlanta, GA: Georgia State University.

Hill, C., Gormley, W., & Adelstein, S. (2015). Do the short-term effects of a strong preschool program persist? Early Childhood Research Quarterly, 32, 60–79.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71, 1161–1189.

Hustedt, J. T., Barnett, W. S., Jung, K., & Goetze, L. D. (2009). The New Mexico pre-K evaluation: Results from the initial four years of a new state preschool initiative—Final report. New Brunswick, NJ: National Institute for Early Education Research.

Jenkins, J., Farkas, G., Duncan, G., Burchinal, M., & Vandell, D. (2016). Head Start at ages 3 and 4 versus Head Start followed by state pre-K: Which is more effective? Educational Evaluation and Policy Analysis, 38, 88–112.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Review of Economic Studies, 76, 1071–1102.

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of Economic Literature, 48, 281–355.

Lipsey, M., Farran, D., & Hofer, K. (2015). A randomized control trial of a statewide voluntary prekindergarten program on children's skills and behaviors through third grade. Nashville, TN: Peabody Research Institute, Vanderbilt University.

Lipsey, M., Howell, J., Kelley, M., Chapman, G., & Carver, D. (2010). Improving the effectiveness of juvenile justice programs. Washington, DC: Center on Juvenile Justice Reform, Georgetown University.

Little, R. J. & Rubin, D. B. (2014). Statistical analysis with missing data. Hoboken, NJ: Wiley & Sons.

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

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.

Muschkin, C. G., Ladd, H. F., & Dodge, K. A. (2015). Impact of North Carolina's early childhood initiatives on special education placements in third grade. Educational Evaluation and Policy Analysis, 37, 478–500.

Peisner-Feinberg, E. S., & Schaaf, J. M. (2010). Summary of key findings: Long-term effects of the North Carolina More at Four pre-kindergarten program. Chapel Hill, NC: Frank Porter Graham Child Development Institute.

Peisner-Feinberg, E. S., & Schaaf, J. M. (2011). Effects of the North Carolina More at Four pre-kindergarten program on children's school readiness skills: Key findings. Chapel Hill, NC: FPG Child Development Institute.

Phillips, D. A., Gormley, W., & Anderson, S. (2016). The effects of Tulsa's CAP Head Start program on middle school academic outcomes and progress. Developmental Psychology, 52, 1247–1261.

Phillips, D. A., Gormley, W., & Lowenstein, A. (2009). Inside the pre-kindergarten door: Classroom climate and instructional time allocation in Tulsa's pre-K programs. Early Childhood Research Quarterly, 24, 213–228.

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., ... Weiland, C. (2017). Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. A consensus statement. In K. Dodge (Ed.), Issues in pre-kindergarten programs and policy (pp. 19–30). Washington, DC: Brookings.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). Classroom assessment scoring system (CLASS) manual: K-3. Baltimore, MD: Paul H. Brookes Publishing Company.

Reynolds, A. J., Temple, J. A., Ou, S. R., Arteaga, I. A., & White, B. (2011). School-based early childhood education and age 28 well-being: Effects by timing, dosage, and subgroups. Science, 333, 360–364.

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Burgette, L., & Griffin, B. A. (2015). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. Washington, DC: RAND Corporation.

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

Sarama, J., Lange, A., Clements, D., & Wolfe, C. (2012). The impacts of an early mathematics curriculum on oral language and literacy. Early Childhood Research Quarterly, 27, 489–502.

Sass, T., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2012) Value added of teachers in high-poverty schools and lower poverty schools. Journal of Urban Economics, 72, 104–122.

Schweinhart, L. J., Montie, J., Ziang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The HighScope Perry Preschool study through age 40. Monographs of the High/Scope Educational Research Foundation, 14. Ypsilanti, MI: HighScope Press.

Silver, D., Saunders, M., & Zarate, E. (2008). What factors predict high school graduation in the Los Angeles Unified School District? Santa Barbara, CA: UC-Santa Barbara.

StataCorp. (2011). Stata statistical software: Release 12. College Station, TX: StataCorp LP.

Stipek, D. (2004) Teaching practices in kindergarten and first grade: Different strokes for different folks. Early Childhood Research Quarterly, 19, 548–568.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25, 1–21.

Tauchmann, H. (2013). Lee's treatment effect bounds for non-random sample selection—An implementation in Stata. Berlin, Germany: Sonderforschungsbereich (SFB) Discussion Paper 823.

Weiland, C. & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. Child Development, 84, 2112–2130.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30, 377–399.

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

Wong, V., Valentine, J., & Miller-Baine, J. (2017). Empirical performance of covariates in educational observational studies. Journal of Research on Educational Effectiveness, 10, 207–236.

APPENDIX

Table A1. Unstandardized LPM regression coefficients (with standard errors) predicting pre-K attendance.

Covariates	Coefficient (Standard error)
Neighborhood income (in $10,000)	–0.01*
	(0.01)
Race	
Asian	0.16**
	(0.03)
Black	0.01
	(0.05)
Hispanic	–0.07
	(0.09)
Native American	0.01
	(0.03)
Mom's marital status	
Married	0.08*
	(0.04)
Remarried	0.03
	(0.08)
Separated	–0.02
	(0.06)
Divorced	0.00
	(0.05)
Widowed	–0.03
	(0.09)
Mom's education	
High school or GED	0.01
	(0.04)
Some college	0.02
	(0.04)
College degree	0.00
	(0.05)
Lives with father	0.07*
	(0.03)
Lunch status (2006-07)	
Reduced price lunch	0.04
	(0.03)
Full price lunch	–0.01
	(0.03)
Care at age 3	
Day care	0.04
	(0.03)
Preschool	–0.08+
	(0.04)
Center-based care	–0.05
	(0.04)
Head Start	0.30**
	(0.05)
Parent foreign born	0.11+
	(0.06)
Female	0.01
	(0.02)

Table A1. Continued.

Covariates	Coefficient (Standard error)
English spoken at home	0.02
	(0.06)
Internet access at home	0.09**
	(0.03)
Insurance	0.13**
	(0.03)
Oldest child	–0.06*
	(0.02)
Number of siblings	–0.01
	(0.01)
Public housing at K	–0.04
	(0.07)
Constant	0.22**
	(0.08)

Note: Standard errors in parentheses. $^{**}p < 0.01$; $^{*}p < 0.05$; $^{+}p < 0.1$.

Table A2. Average, minimum, and maximum standardized bias (SB) by variable, across 40 imputations.

	Unweighted SB			Weighted SB		
	Avg	Min	Max	Avg	Min	Max
Female	0.02	0.02	0.02	0.03	0.00	0.05
Lives with father	0.13	0.10	0.20	0.04	0.03	0.06
Race/ethnicity						
White	0.20	0.20	0.20	0.04	0.02	0.06
Black	0.24	0.24	0.24	0.05	0.04	0.07
Hispanic	0.01	0.01	0.01	0.01	0.00	0.03
Asian	0.00	0.00	0.00	0.01	0.00	0.03
Native American	0.10	0.10	0.10	0.01	0.00	0.02
Marital status						
Never married	0.04	0.00	0.11	0.02	0.00	0.05
Married	0.15	0.05	0.22	0.03	0.01	0.05
Remarried	0.05	0.00	0.16	0.02	0.00	0.05
Separated	0.10	0.01	0.23	0.02	0.00	0.06
Divorced	0.10	0.03	0.26	0.01	0.00	0.04
Widower	0.04	0.00	0.12	0.02	0.00	0.07
Maternal education						
No HS	0.05	0.00	0.11	0.03	0.00	0.07
HS	0.02	0.00	0.09	0.02	0.00	0.04
Some college	0.05	0.00	0.13	0.02	0.00	0.06
College	0.02	0.00	0.06	0.02	0.00	0.07
Free lunch status						
Free	0.00	0.00	0.00	0.01	0.00	0.03
Reduced	0.08	0.08	0.08	0.03	0.00	0.04
Paid	0.06	0.06	0.06	0.01	0.00	0.02
Child care at 3						
Daycare	0.08	0.01	0.15	0.03	0.00	0.06
Preschool	0.16	0.04	0.24	0.04	0.01	0.06
Head start	0.24	0.24	0.24	0.11	0.09	0.14
Center based	0.08	0.02	0.17	0.01	0.00	0.05
Foreign born	0.07	0.01	0.11	0.03	0.01	0.05
English language	0.03	0.00	0.06	0.02	0.00	0.03
Internet at home	0.15	0.06	0.22	0.06	0.03	0.09
Health insurance	0.16	0.10	0.22	0.04	0.02	0.05
Public housing at K	0.03	0.03	0.03	0.03	0.00	0.06
Number of siblings	0.03	0.01	0.07	0.02	0.00	0.04
Oldest sibling	0.13	0.10	0.20	0.03	0.02	0.04
Nb. median income	0.16	0.16	0.16	0.02	0.01	0.04

Table A3. Proportions or means by sample and longitudinal sample retention status.

	Tulsa metro area sample		State sample	
	Not retained	Retained	Not retained	Retained
TPS pre-K	0.35	0.44	0.29	0.42
Gender				
Male	0.52	0.53	0.52	0.53
Female	0.48	0.47	0.48	0.47
Lives with father	0.55*	0.62	0.61	0.59
Race				
White	0.40*	0.31	0.39*	0.34
Black	0.27*	0.34	0.27*	0.32
Hispanic	0.20*	0.26	0.27*	0.22
Asian/Hawaiian	0.01*	0.02	0.01*	0.01
Native American	0.12*	0.08	0.07*	0.11
Mom marital status				
Never married	0.24*	0.26	0.19	0.26
Married	0.52*	0.56	0.59	0.54
Remarried	0.03*	0.02	0.02	0.03
Separated	0.07*	0.05	0.07	0.06
Divorced	0.12*	0.09	0.12	0.10
Widowed	0.01*	0.01	0.01	0.01
Mom education				
No high school/GED	0.18*	0.20	0.18*	0.19
High school/GED	0.26*	0.26	0.22*	0.27
Some college	0.43*	0.36	0.37*	0.39
College degree	0.14*	0.17	0.22*	0.15
Lunch status				
Free lunch	0.69	0.68	0.69	0.68
Reduced price lunch	0.10	0.10	0.08	0.10
Full price lunch	0.21	0.22	0.23	0.21
Redshirted in K	0.00	0.02	0.00	0.01
Born abroad	0.16*	0.24	0.22	0.20
English	0.86*	0.81	0.81	0.83
Internet	0.51	0.50	0.54	0.50
Day care at 3	0.16	0.16	0.15	0.17
Preschool at 3	0.11*	0.16	0.15	0.14
Head Start at 3	0.15	0.16	0.14	0.16
Center-based care at 3	0.48*	0.54	0.51	0.51
Neighborhood median income	3.80 (1.81)	3.82 (1.87)	3.89 (2.06)	3.79 (1.80)

Note: Standard errors are included in parentheses for neighborhood median income because it is a continuous variable. $^{*}p < 0.05$.

Table A4. Results from Lee bounds analyses.

Dependent variable	No. of selected obs/ No. of obs.	Overall trimmed proportion	Lower and upper bound	Effect 95% confidence interval
Math	2636/3605	0.12	10.96, 39.02	2.26, 48.31
Reading	2684/3605	0.12	9.77, 35.07	1.52, 43.56
Retained	2865/3605	0.11	–0.21, –0.12	–0.25, –0.08
Honors	2017/3605	0.19	0.05, 0.19	0.01, 0.23

Note: The table reports the bounds for the treatment effect of the four outcomes as well as 95 percent confidence intervals for the treatment effect.