Developmental Psychology
2016, Vol. 52, No. 8, 1247–1261

0012-1649/16/$12.00 http://dx.doi.org/10.1037/dev0000151

The Effects of Tulsa's CAP Head Start Program on Middle-School Academic Outcomes and Progress

Deborah Phillips, William Gormley, and Sara Anderson
Georgetown University

This study presents evidence pertinent to current debates about the lasting impacts of early childhood educational interventions and, specifically, Head Start. A group of students who were first studied to examine the immediate impacts of the Tulsa, Oklahoma, Community Action Project (CAP) Head Start program were followed-up in middle school, primarily as 8th graders. Using ordinary least squares and logistic regressions with a rich set of controls and propensity score weighting models to account for differential selection into Head Start, we compared students who had attended the CAP Head Start program and enrolled in the Tulsa Public Schools (TPS) as kindergarteners with children who also attended TPS kindergarten but had attended neither CAP Head Start nor the TPS pre-K program as 4-year-olds. CAP Head Start produced significant positive effects on achievement test scores in math and on both grade retention and chronic absenteeism for middle-school students as a whole; positive effects for girls on grade retention and chronic absenteeism; for white students on math test scores; for Hispanic students on math test scores and chronic absenteeism, and for students eligible for free lunches on math test scores, grade retention, and chronic absenteeism. We conclude that the Tulsa CAP Head Start program produced significant and consequential effects into the middle school years.

Keywords: Head Start, academic success, academic progress, middle school

Supplemental materials: http://dx.doi.org/10.1037/dev0000151.supp

The Head Start program has long been at the center of debates over lasting impacts of early childhood interventions. This has been fueled by the availability of medium- and long-term impact data on multiple cohorts of Head Start graduates combined with the prominence of the program as the most longstanding federal investment in promoting school readiness among low-income children. Taxpayers and public officials seek assurance that investments in Head Start have enduring impacts that can justify these expenditures, now over $8.5 billion annually (Office of Head Start, 2015). Researchers are interested in questions of the malleability of development and the persistence of positive changes in developmental trajectories.

Deborah Phillips, Department of Psychology, Georgetown University; William Gormley and Sara Anderson, McCourt School of Public Policy, Georgetown University.

We thank the Tulsa, Jenks, Union, and Broken Arrow Public Schools, and the Oklahoma Department of Education for access to school test data and administrative data. The authors would also like to thank the Foundation for Child Development (Grant 06-2014), the Heising–Simons Foundation (Grant 2014-25), and the Stranahan Foundation (Grant 2725) for their generous financial assistance. This study could not have been conducted without the generous assistance of many individuals in the district and Oklahoma state public school offices, notably Anna Holt, Sam Davidson, Sherri Fair, Lisa Muller, Ellen Dollarhide, and Janet Dunlop. Finally, we thank Karin Kitchens, Howard Huo, Yuqi Liao, Daniel Ready, and Billy Hoo for excellent research assistance.

Correspondence concerning this article should be addressed to Deborah Phillips, Department of Psychology, Georgetown University, 3700 O Street NW, Washington, DC 20057. E-mail: deborah.dap4@gmail.com

In this paper, we examine the effects of Tulsa's Community Action Project (CAP) Head Start program on middle-school students, most of whom had completed 7th grade and were in 8th grade at the time of the study. We also ask if effects differ for subgroups of children defined by gender, race/ethnicity, and both English language learner (ELL) and free-lunch status. Effects are assessed for the following indicators of a student's academic outcomes and successful progress through school: performance on state standardized tests, grade point average (GPA), designation as a gifted student and enrollment in honors courses, grade retention, special education placement, absenteeism, and suspensions.

Head Start and Children's Schooling Outcomes

Evidence on the school readiness impacts of Head Start gathered from the program's inception to the current day has demonstrated that participation in Head Start can boost the cognitive skills of low-income children at school entry (Bitler, Hoynes, & Domina, 2014; Currie & Thomas, 1995; Currie & Thomas, 2000; Deming, 2009; Gormley, Phillips, Adelstein, & Shaw, 2010; V. E. Lee, Brooks-Gunn, & Schnur, 1988; Ludwig & Phillips, 2008; U.S. Department of Health and Human Services [USDHHS], 2005; Zhai, Brooks-Gunn, & Waldfogel, 2011). The most recent and best-available evidence comes from the Head Start Impact Study (HSIS), the only large-scale randomized experiment in Head Start history, conducted in 2002 with 84 Head Start grantees and nearly 5,000 3- and 4-year-old children who were randomly assigned to either Head Start or a control group (USDHHS, 2005). Children who enrolled in Head Start demonstrated significant, but modest, impacts (effect sizes for 4-year-olds ranged from .09 to .31) at the

end of a year in Head Start on a series of early literacy and prewriting assessments and, for those in the 3-year-old cohort, on early math skills as well, relative to the control group children (USDHHS, 2005).

Today, however, the more pressing question is whether these short-term positive effects of Head Start on school readiness persist or dissipate over time. In general, the evidence on lasting impacts of Head Start suggests that the early benefits of the program dissipate quickly once children enter elementary school. Most of the immediate school readiness effects of Head Start diminished considerably or disappeared altogether by the end of first grade (USDHHS, 2010). By the end of third grade, the HSIS reported only two significant cognitive impacts: a positive impact on the Early Childhood Longitudinal Study–K Reading assessment for the 4-year-old cohort and a negative impact on parent reports of grade promotion for the 3-year-old cohort (Puma et al., 2012).

Nevertheless, a series of well-crafted, quasi-experimental studies of earlier cohorts of Head Start children by Currie and colleagues (Currie & Thomas, 1995, 2000; Garces, Thomas, & Currie, 2002) and others (Deming, 2009; Ludwig & Miller, 2007; Ludwig & Phillips, 2008) have found longer-term impacts of Head Start on grade repetition, high school graduation, college attendance, and earnings. Deming (2009), for example, using data from the National Longitudinal Survey of Youth, reported significant kindergarten test score gains that faded out to less than half of the immediate impacts by 11–14 years of age, but significant impacts on a summary score of young adult outcomes that included rates of high school graduation and college attendance. Similarly, Ludwig and Miller (2007), using data from the National Education Longitudinal Study (NELS), reported significant impacts of Head Start on these measures of young adult schooling attainment despite a lack of significant impacts on 8th grade reading and math test scores or grades.

Taken together, this research suggests that long-term schooling-related impacts of Head Start can occur in the context of short-term convergence of test scores for the treatment and comparison groups (Yoshikawa et al., 2013). This pattern of results has also been found for the Perry Preschool program (Schweinhart, Barnes, & Weikart, 1993; Schweinhart et al., 2005), the Abecedarian program (Campbell et al., 2012), and the Chicago Child-Parent Centers (Reynolds, Temple, Ou, Arteaga, & White, 2011), all of which produced large initial impacts on achievement test scores, the effect sizes for which diminished over time. Yet, these programs also produced large program effects on adult schooling attainment and earnings, among other highly consequential outcomes. It is important to note that similar to Head Start, these programs exclusively served low-income children, many of them children of color. But this evidence comes from children who attended these programs in the 1960s and 1970s and cannot be assumed to reflect the Head Start of today. They were also small-scale model programs with questionable applicability to the at-scale Head Start program both then and now. The data reported here follow a 2005–2006 cohort of Tulsa CAP Head Start participants into 8th grade, thus providing a more contemporary portrait of Head Start's longer-term impacts on children's schooling outcomes.

For Whom Do Benefits Persist?

Efforts to gain a deeper understanding of the circumstances under which long-term impacts emerge despite early test score convergence have included examinations of patterns of outcomes for subgroups of children who experienced Head Start. These inquiries are framed in terms of "for whom" the strongest impacts emerge and endure over time, and in terms of the heterogeneity of program impacts. Subgroups defined by race, gender, and extent of economic disadvantage or household risk have been examined.

It is difficult to discern a consistent story from the evidence. For example, in various studies using different datasets and analytic approaches, relatively stronger impacts for high school graduation and college attendance have been reported for Blacks (Deming, 2009), Whites (Currie & Thomas, 1995; Garces et al., 2002; Ludwig & Miller, 2007), and Hispanics (Currie & Thomas, 1999). The negative parent reports regarding grade promotion in the HSIS third-grade follow-up were significant only for White children (Puma et al., 2012). The third-grade HSIS subgroup analyses also compared treatment and control children from high-risk households, defined with a set of risk indices that included receipt of public assistance; parents who did not have a high school degree and were not employed or in school; and a biological mother who was a single parent and who was 18 or younger when the child was born (Puma et al., 2012). At third grade, the highest risk Head Start subgroup (4–5 risk factors) performed at significantly higher levels on two teacher-reported measures of reading and language arts ability than did the highest-risk control children.

Gender differences are rarely examined, but when they are found, some results favor boys and some favor girls (Deming, 2009; Oden, Schweinhart, Weikart, Marcus, & Zie, 2000). Ludwig and Miller (2007) reported somewhat larger long-term outcomes on high school graduation and college attendance for girls. Anderson (2008) reported much larger benefits for girls than boys in his evaluation of the Perry Preschool, Abecedarian, and Early Training Project programs, and results from studies of community-based child care suggest adverse impacts on boys more so than on girls (Desai, Chase-Lansdale, & Michael, 1989; Howes & Olenick, 1986; Winer & Phillips, 2012).

More recent efforts to examine subgroup impacts in the HSIS have found that children who are English language learners with married, poorly educated mothers were more likely than children with other constellations of risk factors to experience lasting early literacy skill benefits into first grade (Cooper & Lanza, 2014). This latter finding is consistent with a growing body of evidence suggesting that children of immigrants and Spanish-speaking dual language learners may benefit more than other subgroups from early education (Gormley, 2008; Magnuson, Lahaie, & Waldfogel, 2006; Weiland & Yoshikawa, 2013). A reanalysis of the HSIS data through first-grade outcomes reached similar conclusions regarding more lasting impacts for children who spoke Spanish at home (Bitler et al., 2014).

The Tulsa CAP Head Start Program

In 1998, Oklahoma established the nation's second universal state pre-K program, available to all 4-year-old children irrespective of income. The program is administered by the state's school districts that provide pre-K services directly or through partnerships with other providers. The Tulsa Public Schools (TPS) col-

laborates with the CAP of Tulsa County Head Start program. Under their agreement, CAP Head Start in Tulsa receives 93% of the funding per student that would otherwise flow to TPS school-based pre-K programs absent a collaboration (Gormley et al., 2010).

All state-funded pre-K programs, including Tulsa's CAP Head Start program, must maintain high quality standards as measured by specific "input" requirements: all teachers must have a bachelor's degree and be early childhood certified; group sizes of 20 and child/staff ratios of 10:1 must be maintained. Though not required by state law, the CAP Head Start program pays its teachers on the K–12 public school wage scale, yet, because the Head Start teachers are employed for 12- rather than 9-months, they do not earn an equivalent annual hourly wage nor do they receive equivalent benefits to TPS pre-K teachers (C. Decker, personal communication, Nov. 12, 2015). Both the CAP Head Start and TPS pre-K classrooms were observed to offer high-quality early education (Phillips, Gormley, & Lowenstein, 2009), with minor differences across programs. Relative to the TPS teachers, the teachers in Head Start spent significantly less classroom time on math instruction and more time on social studies. No program differences were found on the subscales of the Classroom Assessment Scoring System (CLASS: Pianta, La Paro, & Hamre, 2008). All of the Head Start classrooms were full-day; approximately two thirds of the TPS pre-K classrooms were full-day. However, among low-income students enrolled in TPS pre-K programs, 76% attended full-day programs. As with most Head Start programs nationally, Tulsa's CAP Head Start serves 3-year-olds in addition to 4-year-olds; TPS pre-K programs serve only four year-olds.

The CAP Head Start program serves students from diverse subgroups defined by race/ethnicity and ELL status, as well as children with special needs. With the explicit exception of the latter group, virtually all other enrollees must live at or below 100% of the federal poverty line—a requirement that is not the case with the universal school-based TPS pre-K program. Parents of low-income children may choose to enroll them in either a CAP Head Start or a school-based TPS pre-K program. Among the low-income (free-lunch eligible) children who were enrolled in state-funded preschool in 2005–2006, 73% attended TPS pre-K programs and the remaining 27% attended CAP Head Start. Beyond some evidence in Tulsa that parents, in general, select into preschool programs based on geographic proximity, we lack qualitative data that could inform questions about how low-income parents and those whose children have special needs decide between CAP Head Start and TPS pre-K. We do know, however, that the two subsamples differ. Specifically, comparisons at kindergarten entry between the students who had attended CAP Head Start or TPS pre-K the prior year (2005–2006) have revealed that 88% of the Head Start sample was free-lunch eligible as compared to 66% of the TPS pre-K sample. Moreover, the Head Start participants included larger shares of black and Hispanic children. When restricted to free-lunch eligible students in both samples, the Head Start participants were still less likely to be white, more likely to be Hispanic, and less likely to have Internet access at home than the TPS pre-K participants. There were, however, no differences in mother's education, percent black, or the presence of a biological father in the home in these trimmed subsamples (Gormley et al., 2010).

Prior results regarding effects of the CAP Head Start program have been restricted to school readiness outcomes at kindergarten entry using a regression discontinuity design in which Head Start participants entering kindergarten were compared with children entering Head Start at the same point in time (Gormley, Phillips & Gayer, 2008; Gormley et al., 2010). The Head Start participants demonstrated significant school readiness impacts on the Woodcock–Johnson (WJ) III Letter–Word ID (prereading), Spelling (prewriting), and Applied Problems (prematematics) subtests. Effect sizes ranged from .33 for Spelling to .51 for Letter–Word ID, which translate into gains of three to five months of additional learning. The results remained significant when the sample was trimmed to free-lunch eligible children, who composed the large majority of the Head Start participants. These impacts were comparable to the impacts of the school-based TPS pre-K programs in Tulsa for free-lunch eligible students on prematematics skills, but were smaller for the assessments of prereading and prewriting skills (Gormley et al., 2008). These strong initial impacts of the CAP Head Start program are directly pertinent to one of the salient hypotheses regarding test score fade-out for early childhood programs, namely that initial effects on school readiness outcomes are not always potent enough to persist in the long-term (Brooks-Gunn, 2011; Yoshikawa et al., 2013). The CAP Head Start program arguably got its participants off to a strong start in school.

The Present Study

The question addressed in this study is whether Tulsa's CAP Head Start program produced significant impacts in middle school for outcomes associated with academic success and school progress. It extends the important contributions of prior research on longer-term Head Start impacts by providing evidence from a more contemporary sample, using a rigorous nonexperimental design to compare the achievement-related outcomes of middle-school students who had attended the CAP Head Start program to their peers who had not (see sample details, below). Because the CAP Head Start program offers relatively high-quality early education (Phillips et al., 2009), this study is most appropriately viewed as offering an examination of what is possible, more so than what is typical, for children who attended Head Start a decade ago.

We examined middle school impacts for children as a whole and for subgroups defined by gender, race/ethnicity, and ELL and free-lunch status (as of kindergarten), using a propensity score weighting analytic approach, with multiple imputation for missing independent variables. Essentially, we identified comparison group students who most closely resembled CAP Head Start participants and weighed them more heavily than students who did not closely resemble Head Start participants. Because we had access to demographic variables from multiple sources, including administrative data and a parent survey (developed by the Tulsa research team) in the fall of 2006, we were able to match students based on a wide variety of important demographic variables.

We hypothesized that the CAP Head Start participants would demonstrate significant middle-school test score effects given the program's relatively substantial initial effects on kindergarten test scores. We also anticipated positive impacts on grade retention given the prior quasi-experimental evidence on lasting impacts of Head Start. Because higher levels of academic performance and

remaining at grade level are likely to make school more rewarding to students, we also hypothesized that the CAP Head Start program in Tulsa would reduce absenteeism. We had no prior hypotheses with regard to GPA given that grades depend on factors other than cognitive abilities, such as motivation, or with regard to school suspensions for which there is little prior evidence to offer guidance.

With regard to special education status, a strong Head Start program that promotes early learning could either reduce the need for students to maintain special education services as they progress through school or, as a result of careful screening, ensure more accurate identification of students who will benefit from special education in elementary school. In the latter case, reductions in special education students would not necessarily occur. As such, we had no hypotheses regarding special education. We also had no expectations for whether students would demonstrate academic strengths through participation in the TPS gifted and talented program or through enrollment in honors courses.

Given the mixed evidence in the existing literature regarding subgroup impacts, we had no prior expectations with regard to relative impacts on boys and girls or for children from different racial-ethnic backgrounds. Nor did we expect differing results for the free-lunch eligible Head Start participants given that they constitute the vast majority of children in Head Start in Tulsa. We did, however, expect ELL students to benefit relatively more given replicated results in the existing literature examining heterogeneity of program impacts, including reports on the Tulsa school-based pre-K program (Gormley, 2008).

Method

Participants

All students who were enrolled in a TPS kindergarten in 2006–2007 were eligible for inclusion in this study ($N = 4{,}033$). Figure 1 illustrates the sampling approach and provides detailed sample size information for the state and district samples. Table 1 provides corresponding descriptive data for the samples. To be included in the middle school follow-up, students had to be enrolled in an Oklahoma public middle school in 2014–2015 when the major share of the kindergarteners from 2006–2007 were in 8th grade, and to have state achievement test data for that year ($N = 3{,}045$ or 76% of the 4,033 eligible students). We included 7th graders who had been retained in grade once in this eligible pool, in addition to the 8th graders, to minimize the possibility of a skewed sample that could result if, relative to CAP Head Start participants, comparison students were more likely to be retained. The remaining 988 students were no longer in public school records, due presumably to moves to private school or to schools outside of Oklahoma.

Of these 3,045 identified students, 1,271 were excluded from the analyzed middle-school sample because (a) they had attended a TPS school-based pre-K program in 2005–2006 ($N = 1{,}195$), (b) had attended the CAP Head Start program as a 3-year-old only ($N = 21$), or (c) did not have valid state achievement test data ($N = 55$). The TPS participants were excluded from the sample, as in our prior reports (Gormley et al., 2010), because they are seen as their own treatment group given the close collaborative relation-

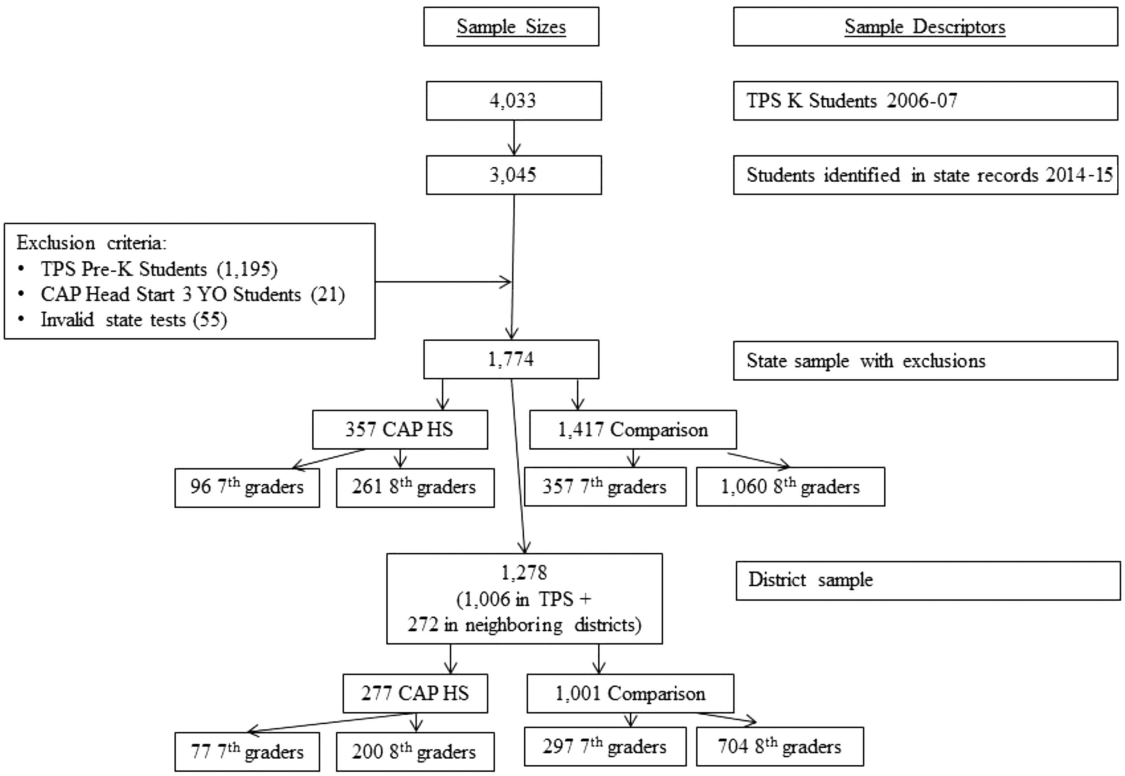

Figure 1. State and district samples with exclusions.

Table 1
Proportions and Means for Student Characteristics in District and State Samples, by Treatment Status, With Absolute Standardized Differences (Stnd. Diff.) for the Unweighted and Weighted District Sample

	Unweighted—State		Unweighted—District			Propensity score weighted—District	
	CAP HS (N = 357)	Comparison (N = 1,417)	CAP HS (N = 277)	Comparison (N = 1,001)		Comparison (N = 1,001)	
Covariate	*M*/%	*M*/%	*M*/%	*M*/%	Stnd. diff.[a]	*M*/%	Stnd. diff.[a]
Female (%)	49%	47%	47%	48%	.02	48%	.01
Live with father at K (%)	55%	52%	56%	56%	.01	55%	.01
Race (%)							
White	10%	42%	10%	38%	.65	12%	.05
Black	42%	24%	39%	26%	.35	37%	.03
Hispanic	41%	20%	46%	24%	.65	45%	.01
Asian	1%	1%	1%	2%	.16	1%	.02
Native American	6%	13%	5%	10%	.23	6%	.02
Mother's marital status at K (%)							
Never Married	35%	29%	32%	28%	.09	34%	.02
Married	44%	46%	46%	49%	.05	47%	.01
Remarried	3%	4%	3%	3%	.02	3%	.01
Separated	11%	7%	12%	7%	.23	10%	.02
Divorced	5%	11%	4%	10%	.29	4%	.03
Widowed	3%	2%	3%	2%	.06	2%	.03
Mother's education at K (%)							
No high school	26%	22%	27%	23%	.12	29%	.03
High school	34%	28%	36%	27%	.2	36%	.01
Some college	31%	38%	27%	35%	.16	28%	.01
College	9%	13%	9%	15%	.18	8%	.04
Lunch status at K (%)							
Free lunch	89%	65%	90%	64%	.47	88%	.06
Reduced price lunch	6%	10%	5%	9%	.25	5%	.03
Full price lunch	5%	25%	6%	27%	.6	7%	.06
Neighborhood median income (in thousands)	34.24 (13.47)	39.84 (20.14)	33.81 (12.12)	40.70 (21.42)	.22	34.28 (20.43)	.03
Overage (6 years old) at K (%)	1%	3%	1%	4%	.27	2%	.09
Foreign born (%)	39%	18%	41%	20%	.63	38%	.03
English language spoken at home at K (%)	67%	85%	65%	81%	.26	66%	.01
Attended daycare at someone's home at age 3 (%)	19%	17%	11%	30%	.4	28%	.18
Attended non-TPS preschool at age 3 (%)	10%	18%	6%	21%	.4	7%	.02
Attended some type of center based care at age 3 (%)	53%	50%	49%	53%	.07	49%	.01
Internet access at home (%)	32%	43%	32%	45%	.24	29%	.03

Note. Results combined across 20 imputed data sets. Propensity score weighted results omitted from state sample (available upon request), as figures are very similar to the district sample upon weighting. Balance statistics are similarly comparable. TPS = Tulsa Public Schools.
[a] Standardized differences presented for district sample.

ship between TPS and CAP Head Start (see above). The students who had attended CAP Head Start as 3-year-olds were excluded because they received exposure to the treatment, albeit not as 4-year-olds, and thus did not belong in either the treatment or comparison groups.

The remaining 1,774 students constituted the final analytic sample for this study, 357 of whom had attended the CAP Head Start program as 4-year-olds in 2005–2006 for at least 50% of the academic year (90 days or more, which was the case for 94% of Head Start participants)—a criterion designed to capture a reasonable exposure to the treatment. These students composed the treatment sample, referred to as the CAP Head Start participants. The remaining 1,417 students constituted the comparison sample. We refer to these treatment and comparison students as the state sample.

Of these 1,774 students, 1,278 were enrolled in a middle school in TPS (N = 1,006) or in one of three neighboring school districts that account for a major share of student migration from TPS to other local middle schools (N = 272). For this subsample of students, we were able to obtain much richer data regarding school achievement and progress than were available from state administrative data (see the Measures section, below). Of these 1,278 students, 277 had attended CAP Head Start as 4-year-olds in 2005–2006 and 1,001 were comparison students. We refer to these students as the district sample.

It is important to note that the comparison samples for the current study differ from that used in the Head Start Impact Study, namely children from Head Start waiting lists who were randomly selected into the control group. The current comparison samples also differ from our prior studies of the immediate impacts of the CAP Head Start program in Tulsa, namely children who participated in the CAP Head Start program a year after those in the treatment sample as determined by a strict birthday cut-off date for enrollment. We return to these distinctions in the discussion.

Measures

This study uses data from four sources: (a) Oklahoma State Department of Education data, (b) District administrative data from TPS and three adjacent school districts, (c) Parent survey data

collected in 2006 when the students were entering TPS kindergarten, and (d) U.S. Census data. The school administrative data provided the outcome measures, as well as demographic covariates, and the parent survey and Census data provided the remaining covariates.

Outcome variables. Five indicators of school achievement and six measures of school progress constituted the outcomes for this study (see Table 2). The school achievement outcomes are: State standardized test scores (Oklahoma Core Curriculum Test [OCCT] in math and reading/language arts administered in the spring of 2014), GPA for all academic grades as of the end of the prior school year, whether or not the students had been enrolled in honors-level courses during the prior school year (including pre-Advanced Placement, pre-International Baccalaureate, advanced/accelerated, or otherwise advanced for the current grade of the student), and whether or not the student was currently designated as gifted and talented (as determined either by an intelligence test or by a teacher/school administrator). The OCCT is a criterion-referenced state assessment administered annually to assess student achievement. The math test included 47 multiple choice questions on algebraic reasoning, number sense and operations, geometry, measurement, and data analysis. The reading test had 50 multiple-choice questions covering vocabulary, comprehension, literature, and research and information (Oklahoma State Department of Education, 2014).

School progress was measured by whether or not the student was currently receiving special education services defined as having an active Individualized Education Plan (IEP), whether or not he or she had been retained in grade (being in 7th grade instead of 8th grade in 2014–2015), number of days absent during the prior school year and whether or not the student had been chronically absent (absent more than 18 days or 10% of the academic year), and whether or not the student had been suspended in or out of school during the prior academic year.

Family background and demographic covariates. At the start of the 2006–2007 school year, in conjunction with required achievement testing of all entering kindergarteners, TPS staff distributed a brief family demographics survey designed by our research team to all parents who accompanied their child to the testing session. The surveys were collected by TPS staff for pick-up by the research team. Close to two thirds of students from both the state (60%) and district (61%) samples had parent survey data. The following survey variables were included as covariates in the propensity score models: maternal education and marital status, presence of Internet in the home, primary language spoken at home, whether the parent was foreign born, whether the child currently lives with the father, and the child's previous child care experiences. It is important to note that 53% of the comparison students with parent survey data had experienced center-based child care prior to kindergarten entry and 22% were in a home-based care arrangement, excluding parental care. TPS school administrative data, collected in the fall of 2006, provided additional covariates and were the source of subgroup assignments: whether the student qualified for free/reduced price lunches, gender, and race/ethnicity. An additional covariate—neighborhood median income during the kindergarten year—was obtained by geocoding each student's 2006 home address using ArcGIS (Environmental Systems Research Institute, 2013) and linking it to federal identifiers of the block group. These data were then linked to U.S. Census estimates of the neighborhood median income of the block group in 2006.

Analytic Strategy

Given the absence of an experimental design and the current examination of long-term outcomes, which precludes a regression discontinuity design, we took advantage of naturally occurring differences in children's participation in CAP Head Start and

Table 2
Unweighted Proportions or Means and Standard Deviationss of Middle School Outcome Variables by Treatment Status

Outcome variables	CAP HS		Comparison	
	M/%	*SD*	*M*/%	*SD*
State sample				
Achievement outcomes				
OCCT math score	680.17	78.79	690.22	93.03
OCCT reading score	686.48	73.48	701.00	87.60
District sample				
Achievement outcomes				
GPA	2.54	0.81	2.70	0.84
Honors student	20%		26%	
Gifted status	10%		19%	
School progress outcomes				
Total absences	9.21	9.37	10.68	10.47
Student chronically absent (>10% days)	18%		25%	
Received in school suspension at least 1 time	18%		15%	
Received out of school suspension at least 1 time	24%		17%	
Repeated a grade by 8th grade	33%		36%	
Special education services	24%		26%	

Note. Results combined across 20 imputed data sets. OCCT = Oklahoma Core Curriculum Test.

employed propensity score estimation to balance treatment and comparison groups on observable covariates. Accordingly, we estimated the difference in Head Start effects between children who did and did not participate in Head Start, taking into account observable characteristics obtained from the parent survey, 2006 school administrative data, and Census data.

We used boosted regression modeling techniques to estimate the propensity scores (McCaffrey, Ridgeway, & Morral, 2004; Schonlau, 2005). These techniques incorporate nonparametric regression or classification trees to find the best model fit that minimizes prediction error. The algorithm iteratively splits the data according to covariate values and applies increasing weights to observations that were not successfully classified in previous iterations. The final boosted model combines results across the iterations to produce a prediction of Head Start participation. Our boosted model specified three interactions with 2,000 iterations. We employed the Twang package in R to estimate the propensity score, which is then converted into a weight to estimate the average treatment effect on the treated (ATT; see McCaffrey et al., 2004; Ridgeway, McCaffrey, Morral, Burgette, & Griffin, 2015).

To estimate the ATT with propensity scores, we weighted our sample by the odds of Head Start participation, whereby the Head Start participants received a weight of 1 and the comparison students received a weight equal to their propensity score (ρ_i) converted to the odds scale ($\rho_i/[1-\rho_i]$; Hirano, Imbens, & Ridder, 2003). This strategy up-weights the comparison students whose observed covariate values best matched those of Head Start participants and down-weights comparison students whose observed covariate values were *unlike* those of Head Start participants. In other words, the comparison group was weighted to most closely resemble the Head Start group, given observable characteristics. Many other algorithms for propensity score analysis exist (i.e., matching), but there is no consensus on the single best approach in all settings (Caliendo & Kopeinig, 2008; Guo & Fraser, 2010; Stuart, 2010). Our approach focuses on achieving the best covariate balance (Harder, Stuart, & Anthony, 2010) and, as shown in the next section, ATT weighting produced well-balanced groups overall and by imputed dataset and variable.

We then used multiple regression (either ordinary least squares [OLS] or logistic depending on whether the outcome variable was continuous or dichotomous) with kindergarten school fixed effects, with weights and covariates as additional controls. We also controlled for school district effects by adding school district dummies to the state-level (TPS vs. all other districts) and district-level (dummies for each of the four districts) models. With 67 districts represented in the state sample, the state-level models that included individual district fixed effects did not converge. The decision to rely on kindergarten, rather than middle school, fixed effects models was made for three reasons. First, the follow-up outcome data on the students in this study reflect, at most, their first two years of middle school (e.g., grade retention and special education status as of the start of 8th grade). In contrast, the students had spent the prior six years in elementary school and, in Tulsa, approximately 70% of TPS students remain in a single elementary school from K–5th grade. As a result, their elementary school environments will arguably have a larger influence than their middle schools on their developing attitudes about learning, views of their own abilities, and motivational dispositions, all of which play a powerful role in shaping the outcomes addressed in this study (Stipek, 2002). Second, reliance on elementary school fixed effects is common in the pertinent empirical literature examining longer-term impacts of preschool programs (Claessens, Duncan, & Engel, 2009; Magnuson, Ruhm, & Waldfogel, 2007; Reynolds et al., 2011). Third, any effects of Head Start participation on selection into formal schooling are likely to affect the most proximal (i.e., elementary) school transition at least as much as the transition to middle school, as suggested by findings from the HSIS (Puma et al., 2012).

Regression covariates included dummy variables for gender, free-lunch status, race, presence of father in the home, mother's education and marital status, the presence of Internet in the home, whether the child was overage in kindergarten, an indicator of current school district (with TPS as the omitted referent), and neighborhood median income (in thousands). All analyses were conducted in Stata 13.0 and included the survey command to incorporate propensity score weights. We also ran these models stratified by gender, race/ethnicity, and the child's ELL and free-lunch status.

To minimize the threat of bias from excluding subjects with missing data, we employed multiple imputation, which estimated missing variables on the covariates using a series of chained equations, filling in missing data as estimated from complete data in an iterative process (Little & Rubin, 2014). We created 20 imputed datasets using Stata 13.0 employing the *mi impute chained* command (StataCorp, 2011). Each of these imputed datasets was then used to estimate propensity scores and outcome models, and results were combined incorporating variability in estimates across imputed datasets (Rubin, 1987).

We report the results of two robustness checks (coarsened exact matching [CEM] and attrition weights) after presenting our outcome results for students as a whole and for the subgroups. Briefly, CEM (Blackwell, Iacus, King, & Porro, 2009) matches subjects exactly on a set of predetermined characteristics, as chosen by the analyst. A benefit of CEM over propensity scores is that the analyst predetermines the desired balance; hence, balance does not need to be assessed following propensity score estimation. In this case, we required exact matching on gender, race/ethnicity, free lunch status, and neighborhood median income (within quintiles). Matched subjects were then compared using traditional tests of mean differences, with weighted multiple regression.

In addition to CEM, we ran models with weights to take into account the potential for differential attrition from kindergarten to middle school. Although the characteristics of subjects were overall very similar between the kindergarten and middle school state and district samples, some differences by race in particular were noted for the district sample (see the appendix in the online supplementary materials). Therefore, we created weights using a similar procedure to that used in the creation of our propensity scores, though instead of predicting Head Start participation, we predicted the likelihood of attrition, using the same suite of background characteristics and Head Start participation. Analyses were conducted with 20 imputed datasets.

Results

Descriptive Statistics and Results of Propensity Score Matching

Descriptive statistics on the pretreatment covariates and unstandardized outcome measures can be found in Table 1, which in-

cludes the child and family characteristics employed as covariates in the propensity score and regression analyses for the state sample and for the district sample before and after propensity weighting, with the accompanying average standardized bias statistic (absolute standardized difference [ASD] or the absolute difference between the CAP Head Start and comparison groups for the district sample).

Prior to the weighting, differences between CAP Head Start participants and comparison students in the state and district samples were evident with most of the covariates. Specifically, Head Start participants in the state sample were less likely to be white and to have parents who attended some college. They were more likely to be foreign born, and less likely to speak English at home than the comparison sample. In the district sample, the Head Start participants were more likely to be black or Hispanic and less likely to be white; more likely to have a never-married mother and one with relatively less education; more likely to receive free lunches (thus, to be low-income) and to live in a neighborhood with a relatively low median income; to be foreign born and to be an ELL student, and less likely to have Internet at home. These differences between the groups likely arise from the fact that Head Start participants in Tulsa must be from households at or below 100% of poverty, among other criteria. Nonetheless, after the weights were applied to the comparison groups, the differences between the Head Start and comparison students decreased substantially (district sample ASDs shown in Table 1; state ASDs were very similar so omitted for parsimony). The standardized biases for all variables across imputed datasets were below the conservative threshold of 0.10, representing very balanced groups with regard to observed covariates (Harder et al., 2010). We also examined balance by subgroups and individual imputation, and the standardized biases remained under the 0.10 threshold with few exceptions. As seen in Table 2, prior to propensity-score weighted analyses, CAP Head Start and comparison group students differed on all student outcomes. Head Start students generally demonstrated lower achievement outcomes and poorer evidence of school progress, with the notable exception of chronic absenteeism.

Effects of CAP Head Start

Tables 3 and 4 present the results for academic achievement overall and then by subgroup, respectively. Tables 5 and 6 provide the same information for the measures of school progress. All percentages reported are adjusted holding covariates at means and balanced.

Academic achievement. Results for the academic achievement outcomes are presented in Table 3. CAP Head Start participants received higher math, but not reading, state test (OCCT) scores than did the comparison students ($ES = 0.13$). No significant differences between the Head Start and comparison students in the district sample were found for GPA, honors course taking, or gifted status.

Subgroup effects, presented in Table 4, were found for race/ethnicity and free-lunch eligibility and, at a marginal level, for gender, although some of the models would not converge for specific subgroups. Specifically, the findings for math achieve-

Table 3
Coefficients (and Standard Errors) for CAP Head Start Participation Predicting Academic Achievement

Variable	(1) Math OCCT	(2) Reading OCCT	(3) GPA	(4) Honors	(5) Gifted
CAP Head Start	12.54 (5.64)*	7.97 (5.61)	.01 (.06)	.18 (.23)	−.01 (.02)
Female	8.45 (5.48)	27.07 (5.01)***	.33 (.06)	.22 (.23)	−.03 (.07)
Lived with father at K	8.30 (8.41)	3.29 (8.81)	.09 (.09)	.27 (.33)	−.001 (.04)
Black	−12.33 (9.35)	−23.94 (8.43)*	.05 (.09)	−.01 (.38)	−.10 (.05)†
Hispanic	−0.53 (9.90)	−13.08 (8.92)	−.42 (.10)	−.16 (.39)	−.07 (.05)
Asian	15.90 (24.02)	0.12 (28.46)	.32 (.19)†	.62 (1.25)	.11 (.20)
Native American	−1.58 (10.86)	−9.17 (10.04)	.003 (.14)	−.51 (.47)	−.14 (.05)**
Mother's marital status at K (comp w/never married)					
Married	1.62 (8.63)	3.24 (7.86)	.08 (.10)	.23 (.37)	.01 (.03)
Remarried	19.00 (22.47)	3.86 (21.76)	−.10 (.26)	1.20 (.77)	.28 (.14)*
Separated	1.52 (12.50)	8.39 (13.67)	−.22 (.13)	.54 (.51)	.03 (.06)
Divorced	5.75 (16.64)	5.33 (14.85)	.27 (.15)†	−.11 (.73)	.02 (.07)
Widowed	7.71 (20.73)	−4.90 (21.38)	−.42 (.27)	−.01 (1.15)	−.06 (.07)
Mother's education at K (comp w/No high school)					
HS/GED	−11.85 (9.89)	−3.57 (8.71)	.05 (.10)	−.19 (.37)	.01 (.04)
Some college	1.58 (10.51)	10.78 (8.22)	.06 (.12)	.27 (.43)	.06 (.05)
College degree	15.69 (16.15)	23.02 (16.47)	.14 (.16)	.13 (.56)	.15 (.08)*
Free lunch status (comp. w/free)					
Reduced	16.69 (11.08)	28.06 (9.14)**	.16 (.11)	1.07 (.43)*	.03 (.07)
Paid	9.59 (12.01)	4.99 (11.08)	.099 (.10)	.56 (.43)	.03 (.06)
Internet in home	20.97 (7.31)**	17.79 (7.35)*	.07 (.09)	.02 (.32)	−.01 (.03)
Overage at K	14.81 (15.43)	37.18 (15.28)*	.44 (.26)†	.94 (.90)	−.04 (.08)
Neighborhood median income (in thousands)	3.24 (1.99)	0.71 (1.86)	−.001 (.002)	.004 (.01)	.0005 (.001)
Constant	657.62	672.99	2.55	−4.30	.26
Observations	1,774	1,774	1,176	1,209	1,278

Note. Results combined across 20 imputed data sets. Models also include current school district and fixed effects for K school as covariates. Observations vary because of intermittent missingness on the student outcomes and differences in observations between state and local district data. CAP = Community Action Project; OCCT = Oklahoma Core Curriculum Test.
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Coefficients (and Standard Errors) for CAP Head Start Participation Predicting Academic achievement, by Gender, Race/Ethnicity, ELL Status, and Free Lunch Status

Subgroup	(1) Math OCCT	(2) Reading OCCT	(3) GPA	(4) Honors	(5) Gifted
Gender					
Male	9.43 (8.29)	5.68 (8.33)	−.07 (.09)	.14 (.38)	—
Female	13.99 (7.64)†	9.86 (7.14)	.12 (.09)	.32 (.33)	.30 (.66)
Race					
White	24.45 (12.46)*	19.43 (12.37)	.05 (.11)	1.35 (1.15)	−.11 (.79)[a]
Black	1.77 (8.24)[b]	−3.70 (8.48)[b]	−.07 (.10)	—	—
Hispanic	22.69 (10.54)*	13.35 (9.21)	.01 (.11)	—	—
Language					
ELL	15.80 (10.23)	15.19 (9.02)†	.03 (.11)	.28 (.38)	—
Free Lunch status					
Free lunch only	12.99 (6.19)*	8.18 (6.20)	.01 (.07)	.16 (.24)	−.01 (.03)

Note. Results combined across 20 imputed data sets. Models include propensity score weights and covariates as per Table 3 plus current school district and fixed effects for K school. *N*s (available upon request) vary by subgroup, outcome (state vs. district), and missingness, but range from 336 (e.g., GPA for ELL students) to 926 (e.g., OCCT Math state data for males). CAP = Community Action Project; OCCT = Oklahoma Core Curriculum Test; ELL = English language learner. Dashes indicates insufficient imputations.
[a] 13 imputations. [b] 18 imputations.
† $p < .10$. * $p < .05$.

ment were restricted to white and Hispanic students (*ES*s are 0.13 and 0.24, respectively) and to students who qualified for free-lunch status (*ES* = 0.14). The math test score outcome for girls was marginally significant (*ES* = 0.15). There was also a marginally significant positive impact on reading test scores for ELL students, 76% of whom were Hispanic (*ES* = 0.17). No statistically significant associations were found between Head Start participation and academic achievement for boys or for black students, although the coefficients were generally in the anticipated direction.

School progress. Attending Head Start was associated with a significantly lower likelihood of being retained in grade prior to 8th grade (see Table 5). The share of Head Start participants retained in grade was 6 percentage points lower than the share of comparison students (*OR* = 0.60). Head Start participants were also less likely to have been chronically absent (*OR* = 0.54); the share chronically absent was 3 percentage points lower than the share of the comparison group. No significant associations were found for special education status, days absent, or in- or out-of school suspensions.

Turning to subgroup analyses (see Table 6), as with the measures of academic achievement, free-lunch eligible students demonstrated significantly better academic progress than did similarly very low-income students who did not attend Head Start. Specifically, they were significantly less likely to have repeated a grade (7 percentage points lower than the comparison group, *OR* = 0.64.) and were significantly less likely to be chronically absent (by 5 percentage points; *OR* = 0.58). Girls showed a similar pattern of results with significant Head Start impacts for grade retention (13 percentage points lower for female Head Start participants than for comparison group females, *OR* = 0.30) and chronic absences (by 11 percentage points; *OR* = 0.25). These results did not characterize boys.

With regard to subgroups defined by race/ethnicity and ELL status, neither white nor black Head Start participants differed significantly on the indicators of academic progress from their comparison group peers. Hispanic Head Start participants, however, were significantly less likely to have been chronically absent (1 percentage point lower among Hispanic Head Start participants, *OR* = 0.36). Finally, both Hispanic and ELL students demonstrated a marginally lower likelihood of repeating a grade (*OR* = 0.56).

Robustness checks. Running CEM-weighted regressions with 20 multiply imputed datasets resulted in coefficients and significance levels that were extremely similar to the propensity-score weighted analyses for all outcomes, and identical conclusions were drawn. (Full results are available from Deborah Phillips upon request.) Second, we created attrition weights from the kindergarten sample (*N* = 4,033), predicting the likelihood of remaining active in our sample by the middle school assessment period using the same set of comprehensive covariates across 20 imputed datasets. We then applied the product of the attrition and propensity score weights to our outcome analyses with the imputed datasets for our main set of analyses. Once again, the results are very similar with the addition of attrition weights.

Discussion

The aim of this study was to examine impacts of Tulsa's CAP Head Start program on children's school-related outcomes in middle school, approximately eight years after having attended Head Start. Using OLS and logistic regressions with a rich set of controls and propensity score weighted models to account for differential selection into CAP Head Start, the outcomes of the Head Start participants were compared to those of their counterparts who did not attend Head Start or a TPS school-based pre-K program as 4-year-olds. The results partially supported our hypotheses regarding test scores, grade retention and absenteeism. The Head Start program had positive impacts on middle-school students' math achievement test scores, but not on their reading scores. In addition, the Head Start participants were 31% less likely to be retained in grade by 8th grade and 34% less likely to be chronically absent, based on adjusted estimates. The Head Start participants did not

Table 5
Coefficients (With Standard Errors) for CAP Head Start Participation Predicting School Progress

Variable	(1) Special ed	(2) Repeat a grade	(3) Days absent	(4) Chronic absenteeism	(5) Suspensions—In school[a]	(6) Suspensions—Out of school
CAP Head Start	−.13 (.21)	−.51 (.21)*	−.85 (.80)	−.62 (.22)**	.11 (.25)	.28 (.23)
Female	−.45 (.21)*	−.86 (.20)***	−.11 (.66)	−.06 (.22)	−.85 (.25)***	−.71 (.24)**
Lived with father at K	−.11 (.32)	−.35 (.33)	−.70 (1.26)	−.47 (.36)	−.08 (.38)	−.67 (.42)
Race (comp w/White)						
Black	−.37 (.33)	−.19 (.37)	−4.52 (1.18)***	−.43 (.38)	−.05 (.45)	.45 (.45)
Hispanic	−.66 (.39)†	−.25 (.37)	−4.84 (1.24)***	−.55 (.37)	−.22 (.47)	−.05 (.48)
Asian	−1.93 (1.11)†	−1.11 (1.14)	−6.01 (1.98)**	.40 (.47)	Omitted	.60 (1.12)
Native American	−.03 (.43)	−.41 (.52)	−1.86 (1.86)	−.40 (.59)	−.24 (.63)	.07 (.60)
Mother's marital status at K (comp w. never married)						
Married	.03 (.30)	.45 (.33)	−1.11 (1.33)	−.12 (.37)	.06 (.49)	−.24 (.36)
Remarried	−.64 (.82)	.49 (.78)	−1.66 (2.73)	−.70 (1.27)	.71 (.87)	.45 (.96)
Separated	−.12 (.41)	.28 (.55)	1.27 (2.08)	.40 (.48)	.50 (.58)	.81 (.46)†
Divorced	−.40 (.65)	−.76 (.66)	−2.24 (1.85)	−.40 (.50)	−.23 (.79)	−1.02 (.80)
Widowed	−.58 (1.30)	−.08 (.92)	2.06 (3.67)	.25 (.92)	−.09 (1.08)	−.49 (1.11)
Mother's education at K (comp w/no high school)						
HS/GED	.09 (.35)	−.28 (.31)	.08 (1.25)	−.02 (.35)	.10 (.48)	.09 (.37)
Some college	.11 (.40)	−.62 (.36)†	.65 (1.62)	−.04 (.45)	.66 (.48)	−.09 (.48)
College degree	−.24 (.60)	−.46 (.59)	−1.02 (1.60)	−.60 (.77)	−.03 (.74)	.77 (.77)
Free/reduced lunch at K (comp. w/Free)						
Reduced	−.07 (.43)	−.78 (.44)†	−2.87 (1.35)*	−.65 (.48)	.003 (.47)	−.55 (.52)
Paid	.17 (.47)	−.22 (.52)	1.94 (1.18)	−.21 (.44)	−.74 (.76)	−.18 (.66)
Internet in home	−.16 (.33)	−.25 (.35)	−.36 (1.00)	−.16 (.34)	.03 (.40)	−.13 (.36)
Overage at K	−1.05 (.79)	−1.17 (.85)	1.15 (2.24)	.16 (.72)	.51 (.84)	−.87 (.77)
Neighborhood median income (in thousands)	.004 (.01)	−.02 (.01)†	−.04 (.03)	−.02 (.01)	.01 (.01)	.003 (.01)
Constant	.06	.10	16.22	.98	−.43	−1.06
Observations	1,276	1,278	1,207	1,207	1,208	1,222

Note. Results combined across 20 imputed data sets. Models also include current school district and fixed effects for K school as covariates. Observations vary because of intermittent missingness on the student outcomes.
[a] 16 imputations.
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6
Coefficients (and Standard Errors) for CAP Head Start Participation Predicting School progress, by Gender, Race/Ethnicity, ELL Status, and Free Lunch Status

Variable	(1) Special ed	(2) Repeat a grade	(3) Days absent	(4) Chronic absenteeism	(5) Suspensions—In school	(6) Suspensions—Out of school
Gender						
Male	.06 (.30)	−.09 (.31)	−1.01 (1.19)	—	—	—
Female	−.53 (.35)	−1.12 (.36)**	−1.52 (1.11)	−1.04 (.38)**	.60 (.57)	.01 (.43)
Race						
White	.17 (.68)	−1.20 (.78)	−.92 (1.98)	−1.40 (.90)[a]	—	—
Black	—	—	−.77 (1.39)	—	—	—
Hispanic	—	−.57 (.35)†	−1.61 (1.18)	−.98 (.44)*	.46 (.49)	—
Language						
ELL	—	−.56 (.31)†[b]	−2.02 (1.27)	—	—	—
Free Lunch Status						
Free lunch only	−.17 (.23)	−.45 (.23)*	−.47 (.87)	−.55 (.24)*	.02 (.26)	.22 (.25)

Note. Results combined across 20 imputed data sets. Models include propensity score weights and covariates as per Table 3 plus current school district and fixed effects for K school. *N*s (available upon request) vary by subgroup, outcome (state vs. district), and missingness, but range from 336 (e.g., absences for ELL students) to 652 (e.g., grade repetition for males). ELL = English language learners. Dashes indicates insufficient imputations.
[a] 18 imputations. [b] 14 imputations.
† $p < .10$. * $p < .05$. ** $p < .01$.

differ significantly from their classmates who had not attended Head Start as 4-year-olds in their GPAs, gifted status, enrollment in honors classes, special education status, or suspensions. This pattern of findings is roughly comparable to that reported in a companion paper focused on the middle-school impacts of the TPS school-based pre-K program (Gormley, Phillips, & Anderson, 2016).

The magnitude of these treatment effects is larger than the estimated impacts on test scores reported by Deming (2009) for children enrolled in Head Start between 1984 and 1990, with effect sizes of 0.06 for children aged 11–14. It is important to note that Deming reports outcomes for these same children in young adulthood, showing long-term impacts on such consequential outcomes as high school graduation and college attendance (combined effect size of 0.23). Similarly, Ludwig and Miller (2007), using the NELS, found no Head Start impacts on achievement test scores in eighth grade, yet reported evidence for long-term effects on years of schooling completed. When combined with the enduring impacts on grade retention in the current sample, this pattern of results is cause for cautious optimism that our promising middle school results may yield longer-term schooling impacts, though the jury is still out.

There are several reasons why the CAP Head Start program may have generated significant impacts on achievement test scores into middle school. Using a regression discontinuity design, the initial impacts of the Head Start program on the children's WJ preliteracy and prematheamtics scores were relatively strong (as noted above) compared to prior evidence on the immediate impacts of Head Start using both propensity score and experimental analytic approaches (V. E. Lee et al., 1988; USHDDS, 2005; Zhai et al., 2011). For example, the effect sizes for WJ preliteracy subscales ranged from 0.09 to 0.22 immediately after the Head Start year for the 4-year-olds in the HSIS (USDHHS, 2005). The effect sizes from treatment-on-the-treated (TOT) estimates based on the HSIS results, which may be more comparable to the Tulsa kindergarten results, were between 0.26 and 0.32 among 4-year-old children on the WJ scales (Ludwig & Phillips, 2008). Comparing program impacts across studies, however, requires caution given differing designs and samples.

The middle-school impacts seen for the Head Start program in Tulsa may also derive from the relatively high quality of instruction observed in the CAP Head Start classrooms during the year (2005–2006) that the children in this study were in attendance. The Head Start classrooms were compared to Head Start classrooms in the 11 states represented in the National Center for Early Development and Learning (NCEDL) Multi-State Study of Pre-Kindergarten and State-Wide Early Education Programs (SWEEP) Study (Early et al., 2005) on the CLASS (Pianta et al., 2008). Scores on the CLASS Instructional Learning Formats subscale were 4.93 and 3.90, and were 3.41 and 2.18 on the Quality of Feedback subscale for the Tulsa CAP and the NCEDL/SWEEP Head Start programs, respectively (Phillips et al., 2009). The average percent of classroom time spent on literacy instruction was 26% in the Tulsa CAP Head Start classrooms and 18% in the NCEDL/SWEEP Head Start classrooms. The percent of time spent on math was 11% in Tulsa and 7% in the NCEDL/SWEEP classrooms. Finally, all of the CAP Head Start lead teachers in Tulsa had BA degrees, were certified in early education, taught in classrooms with 1:10 teacher:student ratios, and were provided with closely comparable wages to the TPS school-based pre-K teachers. This was not typical of Head Start in 2005–2006. For example, 2007 Program Information Report data indicate that only 79% of Head Start lead teachers had a degree in early care and education or a related field and only 44% of these teachers had a bachelor's degree or higher (Office of Head Start, 2007). Even now, estimates of Head Start lead teachers with bachelor's degrees or higher range from 51% (Whitebook, Phillips, & Howes, 2014) to 67% (Office of Head Start, 2013).

Recent investigations of more typical Head Start programs that generate stronger impacts also offer insights into the middle school results that have emerged in Tulsa. Walters (2014) has recently reported that full-day Head Start centers boost cognitive skills more than do part-day centers. As noted above, all of the CAP Head Start centers in Tulsa offered full-day programs. Currie and

Neidel (2007), in analyses using the National Longitudinal Survey of Youth, found that Head Start programs spending more per capita had larger effects on children's reading and vocabulary scores aggregated across testing ages. In part due to the high-quality requirements of the Tulsa prekindergarten program, the per child expenditures on Tulsa's CAP Head Start program are higher than is typically the case with Head Start (S. Dow, personal communication, June 6, 2015). Moreover, although not specific to Head Start, Magnuson et al. (2007) have reported that prekindergarten programs that are co-located with schools are more likely to generate positive cognitive impacts. Some, but by no means all, of Tulsa's CAP Head Start programs operating in 2005–2006 were co-located with public elementary schools.

The subgroup results are also of interest. The findings with regard to math achievement test scores extended to the Hispanic and white Head Start participants, to those who were free-lunch eligible, and, at a marginal level, to girls and, for reading test scores, to the English language learners. The free-lunch eligible Head Start participants also displayed significantly lower odds of repeating a grade and lower rates of chronic absenteeism than their counterparts in the comparison group. Among Hispanic students, those who attended Head Start were also less likely to be chronically absent than were the comparison students and marginally less likely to repeat a grade, as was also the case for the English language learners. The significant impact of Head Start on grade retention was also present for girls, but not boys. And, girls demonstrated significant impacts on chronic absenteeism.

As we've noted, the available evidence on subgroup impacts within the Head Start literature fails to offer a consistent story, although the broader literature suggests somewhat stronger cognitive impacts for girls (Anderson, 2008; Oden et al., 2000). Although the reasons for this gendered pattern are not well understood, there is some evidence that boys experience poorer quality care than girls in child care and that early childhood teachers portray their relationships with boys as both more problematic and less close than their relationships with girls (Howes & Olenick, 1986; NICHD Early Child Care Research Network, 1997, 2001; Peisner-Feinberg et al., 2001; Wachs, Gurkas, & Kontos, 2004; Winer & Phillips, 2012). This is of concern given evidence that, at the other end of the schooling years, boys are now less likely to graduate from high school, and less likely to attend and complete college and enroll in graduate training than are girls (DiPrete & Buchmann, 2013). There is an urgent need to ensure that the benefits of preschool education extend to and endure for boys, perhaps by including gender bias as an explicit component of early childhood teacher training and mentoring (see Winer & Phillips, 2012).

With regard to the findings for the Hispanic children and ELLs, the existing evidence is somewhat more consistent. Earlier work focused on the TPS school-based programs in Tulsa (Gormley, 2008; Gormley, Gayer, Phillips, & Dawson, 2005), as well as the broader literature on subgroup impacts of Head Start (Bitler et al., 2014; Bumgarner & Brooks-Gunn, 2015; Cooper & Lanza, 2014; Currie & Thomas, 1999)—with the important exception of the third-grade impacts in the HSIS (Puma et al., 2012)—provides evidence that preschool participation may be particularly effective at preparing young Hispanic and ELL children for kindergarten. Speculation about why this may be the case has centered on the possibility that these children get a double benefit from exposure to both English language and academic content in high-quality early education programs, particularly in comparison to their peers who are cared for in home-based settings (Gormley et al., 2005). Indeed, Bumgarner and Brooks-Gunn (2015) have recently reported that the quality of care in other-home arrangements appears lower than in center-based arrangements for Hispanic children, in particular. Fortunately, quality assessments that are specific to the experiences of dual language learners are now available (Castro, Espinosa, & Paez, 2011) and offer the promise of further elucidating the circumstances under which Head Start and other early education settings can foster school readiness among this rapidly growing group of young children.

The findings for the free-lunch eligible students correspond to a long-standing literature that finds larger impacts for a wide range of early education programs on relatively more disadvantaged children (Phillips & Lowenstein, 2011; Johnson, in press). That this pattern of results is found even in Tulsa's CAP Head Start program, which primarily serves very disadvantaged children, is somewhat surprising. It does, however, correspond to findings from the third-grade follow-up of the children in the HSIS which reported more sustained impacts, particularly with regard to early literacy outcomes, for children growing up in higher-risk households (Puma et al., 2012).

Finally, it is of interest that we did not find that participation in CAP Head Start decreased special education placement during the elementary and early middle-school grades. This has been a prominent outcome in the literature examining the enduring impacts of early care and education, including recent studies of state pre-K programs (Andrews, Jargowsky, & Kuhne, 2012; Barnett, Jung, Youn, & Frede, 2013; Bassok & Miller, 2014; Dodge, Bai, Ladd, & Muschkin, 2014; Peisner-Feinberg & Schaaf, 2010). In the 2005–2006 Tulsa school year, children with special needs were encouraged to enroll in CAP Head Start, rather than in the TPS school-based pre-K classrooms (A. McKenzie, personal communication, 2012). The Head Start program, in turn, has extensive experience with children with special needs given the Congressional mandate that at least 10% of enrolled children have an identified disability. In Tulsa's CAP Head Start program, specifically, children are carefully screened for special needs and aggressively enrolled in early intervention and special education services (S. Dow, personal communication, June 12, 2015). It may be the case, as a result, that Head Start's impact on children with special needs is not to reduce their numbers but rather to increase the odds that children who can benefit from special education are identified early and their families are equipped with the skills needed to link them to appropriate services throughout their school years.

Limitations

It is important to acknowledge that there are several limitations of this study. First, like most longer-term studies, our research suffers from sample attrition. Despite our best efforts, we lost 24% of our kindergarten sample when identifying students still in Oklahoma state public middle schools. Fortunately, our middle school state and district analytic samples did not differ from the Head Start and comparison students in the kindergarten sample in one key respect: school lunch eligibility percentages were strikingly similar. For other variables on which the kindergarten and district sample did differ, such as race/ethnicity and mother's

education, we have made appropriate adjustments through propensity score weighting and OLS regression to balance our treatment and comparison groups. Yet, propensity score weighting is an imperfect mechanism for overcoming the absence of random assignment. In contrast to experimental research designs and certain quasi-experimental designs, treatment group parents affirmatively chose Head Start for their children, whereas comparison group parents did not. This means that selection bias is possible. On the other hand, propensity score weighting did exactly what it was supposed to do in paving the way for our regression analysis—it did an excellent job of ensuring minimal differences in the demographic characteristics of our CAP Head Start participants and comparison students.

Second, we were unable to explore the contribution to our results of the counterfactual to CAP Head Start participation in Tulsa due to missing data from the kindergarten parent survey regarding the early care and education experiences of the comparison children. In this context, it is important to note that, unlike the children in the comparison sample in the HSIS who had signed up to attend Head Start and in our earlier studies of the kindergarten impacts of CAP Head Start in Tulsa who were just entering Head Start, the children in the current comparison sample for whom we have pertinent data represented a wide range of 4-year-old care experiences. Many had enrolled in center-based programs not affiliated with the state preschool program and a sizable share was in home-based or informal arrangements, or in parental care. The implications of these differing comparison samples are not immediately evident. However, in light of growing evidence that Head Start has stronger impacts when the counterfactual consists of children who were in home-based or parental care (Feller, Grindal, Miratrix, & Page, 2014; R. Lee, Zhai, Brooks-Gunn, Han, & Waldfogel, 2014; Zhai et al., 2011), it is plausible that our results represent stronger evidence of long-term Head Start impacts than would have been the case if we had restricted the comparison sample to children who had not experienced other center-based arrangements.

Third, we made a reasoned decision to define Head Start as having experienced at least 50% of the CAP Head Start year in the program, but there is little in the existing literature to affirm this decision. In fact, prior studies have either failed to define "participation" or have used varying criteria ranging from any days, to presence on the last day, to reliance on a state definition for prekindergarten participation. This is clearly an issue that warrants careful conceptual and empirical attention. Fourth, Tulsa's CAP Head Start program is not representative of Head Start programs across the country. It is documented, for example, to have had higher scores on observational assessments of instructional quality and on time spent on academic instruction than was true of an 11-state sample of Head Start programs assessed at the same time, as discussed above (Phillips et al., 2009). The results presented here may thus not generalize to more typical Head Start programs.

Finally, we do not have data on the children's elementary or middle school experiences. This is significant in light of emerging evidence that instructional practices and the peer environments in children's elementary schools play an important role in either sustaining or overwhelming the effects of preschool education. What is less clear is whether the influence of the subsequent school context takes the form of the failure of low-quality schools to maintain preschool participants' early learning gains, as suggested by Currie and Thomas (2000), or the tendency for children who did not attend preschool to catch up in high-quality classrooms (see Bierman et al., 2014; Magnuson et al., 2007). A lack of alignment between early learning experiences in preschool and subsequent instructional practices could also contribute to test score convergence, as suggested by Claessens, Engel, and Curran (2014). We do know that Oklahoma is 49th in per student state funding for public education (U.S. Department of Education, 2013) and had a high school graduation rate (79%) in 2011–2012 that placed it behind 83% of the states (U.S. Department of Education, 2014). This suggests that schools in Tulsa may be struggling to provide high-quality education.

Conclusion

The key question, addressed in this study is whether children who attended the Tulsa CAP Head Start program, compared to their nonattending counterparts, showed evidence of stronger academic outcomes and progress in middle-school. We found that they did, with relatively larger impacts for girls, Whites and Hispanics, and free-lunch eligible children. That the effects did not extend to boys or to Blacks is of great concern and warrants both empirical and policy attention. The positive impacts on Hispanic, ELL, and free-lunch eligible children offer signs of hope and affirm the importance of outreach and other efforts to enroll these children in Head Start. Each of the effects we have documented is rather modest in isolation, but together they suggest that CAP Head Start, under the conditions that it operates in Tulsa, can provide young children with a strong boost into their subsequent stages of schooling. The challenge for public officials is how to replicate this success. Nationally, Head Start is addressing the quality of its teaching workforce. Other promising approaches would focus on the content and quality of instructional practices in Head Start classrooms; the adult work environment, compensation, and well-being of Head Start teachers; the dosage of Head Start received by the enrolled children; and explicit efforts to link children's experiences and learning trajectories over the course of the Head Start year to their experiences in kindergarten and then into their subsequent elementary schooling. With the nation's attention now turned to preschool education, it is an opportune time to make the investments necessary to ensure that the initial impacts of Head Start are strong enough to start children along an enduring path toward school success.

References

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

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. http://dx.doi.org/10.3386/w18598

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 of Early Education Research. Retrieved from http://nieer.org/sites/nieer/files/APPLES%205th%20Grade.pdf

Bassok, D., & Miller, L. C. (2014, November). *Do children benefit from widely-available public preschool? Evidence from Florida's voluntary prekindergarten program*. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Albuquerque, NM.

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development, 85,* 140–159. http://dx.doi.org/10.1111/cdev.12117

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014, August). *Experimental evidence on distributional effects of Head Start* (NBER Working Paper 20434). Washington, DC: National Bureau of Education Research. http://dx.doi.org/10.3386/w20434

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: Coarsened Exact Matching in Stata. *The Stata Journal, 9,* 524–546.

Brooks-Gunn, J. (2011). Early childhood education: The likelihood of sustained effects. In E. Zigler & W. Gilliam (Eds.), *The pre-k debates: Current controversies and issues* (pp. 200–207). Baltimore, MD: Brookes.

Bumgarner, E., & Brooks-Gunn, J. (2015). The association between early care arrangements, quality, and emergent bilingual Latino American children's math and literacy skills in English. *Early Childhood Research Quarterly, 30,* 32–44. http://dx.doi.org/10.1016/j.ecresq.2014.08.002

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

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., . . . 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. http://dx.doi.org/10.1037/a0026644

Castro, D. C., Espinosa, L., & Paez, M. (2011). Defining and measuring quality early childhood practices that promote dual language learners' development and learning. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 257–288). Baltimore, MD: Brookes.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28,* 415–427. http://dx.doi.org/10.1016/j.econedurev.2008.09.003

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

Cooper, B. R., & Lanza, S. T. (2014). Who benefits most from Head Start? Using latent class moderation to examine differential treatment effects. *Child Development, 85,* 2317–2338.

Currie, J., & Neidel, M. (2007). Getting inside the "black box" of Head Start quality: What matters and what doesn't. *Economics of Education Review, 26,* 83–99.

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

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

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *The Journal of Human Resources, 35,* 755–774. http://dx.doi.org/10.2307/146372

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

Desai, S., Chase-Lansdale, P. L., & Michael, R. T. (1989). Mother or market? Effects of maternal employment on the intellectual ability of 4-year-old children. *Demography, 26,* 545–561. http://dx.doi.org/10.2307/2061257

DiPrete, T. A., & Buchmann, C. (2013). *The rise of women: The growing gender gap in education and what it means for American Schools*. New York, NY: Sage.

Dodge, K. A., Bai, Y., Ladd, H., & Muschkin, C. (2014, November). *Impact of statewide early childhood programs and policies on educational outcomes through the end of elementary school*. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Albuquerque, NM.

Early, D. M., Barbarin, O., Bryant, D., Burchinal, M., Change, F., Clifford, R., . . . Barnett, W. S. (2005, May). *Pre-kindergarten in eleven states: NCEDL's multi-state study of pre-kindergarten and study of state-wide early education programs (SWEEP)*. Chapel Hill, NC: National Center for Early Development and Learning.

Environmental Systems Research Institute. (2013). *ArcGIS Desktop* (Version 10). Redlands, CA: Author.

Feller, A., Grindal, T., Miratrix, L., & Page, L. (2014). Compared to what? Variation in the impacts of early childhood education by alternative care-type settings. *Preliminary draft*. http://dx.doi.org/10.2139/ssrn.2534811

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

Gormley, W., Phillips, D., & Anderson, S. (2016). *The effects of Tulsa's pre-K program on middle-school academic outcomes and progress*.

Gormley, W. T., Jr. (2008). The effects of Oklahoma's pre-k program on Hispanic Children. *Social Science Quarterly, 89,* 916–936. http://dx.doi.org/10.1111/j.1540-6237.2008.00591.x

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

Gormley, W., Phillips, D., Adelstein, S., & Shaw, C. (2010). Head Start's comparative advantage: Myth or reality? *Policy Studies Journal: The Journal of the Policy Studies Organization, 38,* 397–419.

Gormley, W., Phillips, D., & Gayer, T. (2008, June 27). Preschool programs can boost school readiness. *Science, 320,* 1723–1724.

Guo, S., & Fraser, M. W. (2010). *Propensity Score Analysis*. Thousand Oaks, CA: Sage.

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. http://dx.doi.org/10.1037/a0019623

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71,* 1161–1189. http://dx.doi.org/10.1111/1468-0262.00442

Howes, C., & Olenick, M. (1986). Family and child care influences on toddlers' compliance. *Child Development, 57,* 202–216. http://dx.doi.org/10.2307/1130652

Johnson, A. (in press). Child care and child development in the United States. In E. Dearing & E. Votruba-Drzal (Eds.), *Handbook of early childhood programs, practices, and policies*. New York, NY: Blackwell.

Lee, R., Zhai, F., Brooks-Gunn, J., Han, W.-J., & Waldfogel, J. (2014). Head start participation and school readiness: Evidence from the Early Childhood Longitudinal Study-Birth Cohort. *Developmental Psychology, 50,* 202–215. http://dx.doi.org/10.1037/a0032280

Lee, V. E., Brooks-Gunn, J., & Schnur, E. (1988). Does Head Start work? A 1-year follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Developmental Psychology, 24,* 210–222. http://dx.doi.org/10.1037/0012-1649.24.2.210

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. New York, NY: Wiley.

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

Ludwig, J., & Phillips, D. A. (2008). Long-term effects of head start on low-income children. *Annals of the New York Academy of Sciences, 1136,* 257–268. http://dx.doi.org/10.1196/annals.1425.005

Magnuson, K., Lahaie, C., & Waldfogel, J. (2006). Preschool and School Readiness of Children of Immigrants. *Social Science Quarterly, 87,* 1241–1262. http://dx.doi.org/10.1111/j.1540-6237.2006.00426.x

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly, 22,* 18–38. http://dx.doi.org/10.1016/j.ecresq.2006.10.002

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

NICHD Early Child Care Research Network. (1997). Familial factors associated with the characteristics of nonmaternal care for infants. *Journal of Marriage and the Family, 59,* 389–408. http://dx.doi.org/10.2307/353478

NICHD Early Child Care Research Network. (2001). Child care and children's peer interaction at 24 and 36 months: The NICHD study of early child care. *Child Development, 72,* 1478–1500. http://dx.doi.org/10.1111/1467-8624.00361

Oden, S., Schweinhart, L., Weikart, D., Marcus, S., & Zie, Y. (2000). *Into adulthood: A study of the effects of Head Start*. Ypsilanti, MI: High/Scope Press.

Office of Head Start. (2007). *Head Start information report (PIR)*. Washington, DC: U. S. Department of Health and Human Services, Administration for Children and Families.

Office of Head Start. (2013). *Head Start information report (PIR)*. Washington, DC: U. S. Department of Health and Human Services, Administration for Children and Families.

Office of Head Start. (2015). *Head Start fact sheets*. Washington, DC: U. S. Department of Health and Human Services, Administration for Children and Families.

Oklahoma State Department of Education. (2014). *Oklahoma School Testing Program Oklahoma Core Curriculum Tests Grades 3–8 Assessments 2013–14*. Oklahoma City, OK: Office of Assessments.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejin, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72,* 1534–1553.

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: University of North Carolina Frank Porter Graham Child Development Institute.

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

Phillips, D., & Lowenstein, A. (2011). Structure and goals of preschool educational settings. In S. T. Fiske, D. L. Schacter, & S. E. Taylor (Eds.), *Annual Review of Psychology, 62,* 483–495. Palo Alto, CA: Annual Reviews.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) Manual*. Baltimore, MD: Brookes.

Puma, M., Bell, S., Book, R., Heid, C., Broene, P., Jenkins, F., . . . Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study Final Report* (OPRE Report #2012–45). Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Reynolds, A. J., Temple, J. A., Ou, S. R., Arteaga, I. A., & White, B. A. B. (2011, July 15). School-based early childhood education and age 28 well-being: Effects by timing, dosage, and subgroups. *Science, 333,* 360–364. http://dx.doi.org/10.1126/science.1203618

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*. Retrieved from https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a Stata plugin. *The Stata Journal, 5,* 330–354.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The Hish/Scope Perry Preschool study through age 27. Monographs of the High/Scope Educational Research Foundation, 10*. Ypsilanti, MI: High/Scope Press.

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*.

StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, TX: Author.

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th ed.). Needham Heights, MA: Allyn & Bacon.

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

U.S. Department of Education. (2013, July). *Revenues and expenditures for public elementary and secondary education: School year 2010–2011*. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2014, April). *Public high school four-year on-time graduate rates and event dropout rates: School years 2010–11 and 2011–12*. Washington, DC: National Center for Education Statistics.

U.S. Department of Health and Human Services, Administration for Children and Families. (2005). *Head Start Impact Study: First Year Findings*. Washington, DC: Author.

U.S. Department of Health and Human Services, Administration for Children and Families. (2010). *Head Start Impact Study: Final Report*. Washington, DC: Author.

Wachs, T. D., Gurkas, P., & Kontos, S. (2004). Predictors of preschool children's compliance behavior in early childhood classroom settings. *Applied Developmental Psychology, 25,* 439–457. http://dx.doi.org/10.1016/j.appdev.2004.06.003

Walters, C. (2014, October). *Inputs in the production of early childhood human capital: Evidence from Head Start* (NBER Working Paper 20639). Washington, DC: National Bureau of Education Research. http://dx.doi.org/10.3386/w20639

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. http://dx.doi.org/10.1111/cdev.12099

Whitebook, M., Phillips, D. & Howes, C. (2014). *Worthy work, STILL unlivable wages: The early childhood workforce 25 years after the National Child Care Staffing*. Berkeley, CA: Center for the Study of Child Care Employment.

Winer, A., & Phillips, D. (2012). Boys, girls, and the "two worlds" of child care. *Merrill–Palmer Quarterly, 58,* 22–49. http://dx.doi.org/10.1353/mpq.2012.0004

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinoza, L. M., Gormley, W. T., . . . Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Washington, DC: Society for Research in Child Development and the Foundation for Child Development.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Developmental Psychology, 47,* 134–152. http://dx.doi.org/10.1037/a0020784

Received June 29, 2015
Revision received April 14, 2016
Accepted April 24, 2016 ■