CHILD DEVELOPMENT

Child Development, September/October 2020, Volume 91, Number 5, Pages 1401–1422

The Effects of Enrolling in Oversubscribed Prekindergarten Programs Through Third Grade

Christina Weiland
University of Michigan

Rebecca Unterman
MDRC

Anna Shapiro
University of Michigan

Sara Staszak
MDRC

Shana Rochester, and Eleanor Martin
University of Michigan

This study leverages naturally occurring lotteries for oversubscribed Boston Public Schools prekindergarten program sites between 2007 and 2011, for 3,182 children (M = 4.5 years old) to estimate the impacts of winning a first choice lottery and enrolling in Boston prekindergarten versus losing a first choice lottery and not enrolling on children's enrollment and persistence in district schools, grade retention, special education placement, and third-grade test scores. There are large effects on enrollment and persistence, but no effects on other examined outcomes for this subsample. Importantly, children who competed for oversubscribed seats were not representative of all appliers and almost all control-group children attended center-based preschool. Findings contribute to the larger evidence base and raise important considerations for future prekindergarten lottery-based studies.

Decades of research have shown that attending preschool improves children's cognitive and socio-emotional skills at kindergarten entry (Duncan & Magnuson, 2013). This evidence, along with dramatic increases in maternal employment, has helped to fuel currently high levels of public support and parental demand for public preschool. Via a combination of public dollars and parental spending, attending preschool is now the typical experience for U.S. 4-year-olds (Chaudry, Morrissey, Weiland, & Yoshikawa, 2017). In all, 69% of 4-year-olds attend some form of center-based preschool in the year before they enter kindergarten, though children from the top income quintile are much more likely to attend preschool than those in the bottom quintile (83% vs. 50%, respectively; Whitehurst & Klein, 2015). Approximately 43% of 4-year-olds access preschool through public funding, via state or local prekindergarten programs or Head Start (Barnett et al., 2017).

While the evidence is nearly incontrovertible that children who attend preschool enter kindergarten better ready to learn (Phillips et al., 2017; Yoshikawa et al., 2013), questions about how long the benefits of preschool persist are long-standing, dating back to the first major public investment in preschool in the United States—Head Start in the 1960s (Cicirelli, 1969). The overall pattern in the older literature is that the language, literacy, and mathematics test scores of preschool participants and nonparticipants tend to converge in the early elementary grades (i.e., by around third grade), sometimes partially and sometimes fully (Phillips et al., 2017; Yoshikawa et al., 2013). But in the

This study is funded by the Institute of Education Sciences RFA R305A140059. Some of Anna Shapiro's time was also supported by Institute of Education Sciences RFA R305B150012.

Thanks to the Boston Public Schools (BPS), Jason Sachs, Brian Gold, the BPS Department of Early Childhood coaches and staff, the BPS Office of Data and Accountability (particularly Nicole Wagner, Erin Cooley, Barry Kaufman, and Peter Sloan), Kamal Chavda, Carrie Conaway and the Massachusetts Department of Elementary and Secondary Education, Abt Associates, and the Wellesley Centers for Women. Special thanks also to Hirokazu Yoshikawa, Howard Bloom, Richard Murnane, David Deming, Catherine Snow, Caroline Ebanks, Parag Pathak, Gina Biancarosa, and the three anonymous *Child Development* reviewers.

Correspondence concerning this article should be addressed to Christina Weiland, School of Education, 610 E. University Ave, Ann Arbor, MI 48104. Electronic mail may be sent to weilandc@umich.edu.

 0009-3920/2020/9105-0001
DOI: 10.1111/cdev.13308

studies examining long-run effects, preschool participants tend to outperform nonparticipants on a wide range of behavioral, health, and educational outcomes into adulthood. Evidence from modern-day, scaled-up programs so far largely mirrors this medium-term pattern, though a group of experts recently concluded that such evidence "is sparse, precluding broad conclusions" (Phillips et al., 2017). Furthermore, long-run evidence is not yet available for modern-day, large-scale programs.

In this study, we address current needs in the literature using lotteries for oversubscribed program sites as a window into the medium-term effects of the Boston Public Schools (BPS) prekindergarten program. Specifically, using data from four cohorts of students, we examine whether children who won their first choice lottery and enrolled in Boston prekindergarten benefit more than children who lost their first choice lottery and ultimately did not enroll in BPS prekindergarten. Our outcomes are drawn from administrative records and include third-grade state-standardized reading and mathematics test scores, K-2 grade retention, and K-3 special education placement. We also examine whether prekindergarten leads children to enroll and persist in the BPS at higher rates, as one of the program's goals was to attract and retain families that might otherwise not have enrolled in BPS schools. Finally, to contextualize the findings, we descriptively examine children's post-prekindergarten schooling environments.

Importantly, given both calls in the field for more rigorous longitudinal studies of prekindergarten (Phillips et al., 2017) and increasing attention to external validity (Stuart, Bradshaw, & Leaf, 2015; Tipton, 2014), the effects of prekindergarten enrollment that we estimate apply to *the subgroup of lottery compliers*—that is children who won or lost their first choice lottery and either enrolled in the program (first choice lottery winners) or did not enroll in the program at all (first choice lottery losers). As we describe, our lotteries were highly concentrated in a small subset of BPS schools (e.g., 75% of lottery applicants competed for about a quarter of eligible district schools) and the children who competed for oversubscribed seats were more advantaged than the average applicant. Virtually all the control group members attended other center-based preschool programs, an unusual counterfactual in the public prekindergarten evaluation literature. To assess external validity, we followed the example of Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak's (2011) seminal lottery-based study and used data on the full set of program applicants and enrollees to examine the generalizability of our results through descriptive and quasi-experimental analyses. These analyses are important given recent attention to how effects for compliers may not represent a generalizable test of the effects of a program on all members of a target population (Chyn, 2018).

When Might Prekindergarten Benefits Persist?

Multiple theoretical frames are relevant to examining whether and when attending prekindergarten might boost children's medium-term academic and school progress outcomes. First, the human capital accumulation theory from economics posits that a strong early foundation sets the stage for acquiring more advanced skills. Heckman (2000) referred to this perspective as "learning begets learning, skill begets skill." Second, developmental cascades theory, which has its origins in the field of developmental psychology, describes the processes by which antecedent conditions have different probabilities of leading to particular outcomes; functioning at a particular level or in a particular developmental domain is hypothesized to affect later competencies in multiple domains (Masten & Cicchetti, 2010). A third theoretical perspective for expecting persistence is based on transactional developmental theory (Sameroff, 2009)—that is, the reciprocal effects of child skills and environmental inputs on subsequent teacher behaviors, and effects of such behaviors in turn on students. Following this theory, the prekindergarten boost may persist because participants' later-grade teachers may respond to their students increased skill level by increasing the instructional opportunities that they offer students. There also may be observer-expectancy effects in which teachers may either consciously or subconsciously behave in ways that facilitate students' progress in accordance with their own expectations of the students (Weinstein, 2004).

Most recently, Bailey et al. (2017) built on these theories and offered three hypotheses for the persistence (or not) of a preschool boost. First, their "sustaining environments" hypothesis posits that the quality (broadly defined) of children's educational settings after preschool is critical in sustaining the preschool boost. As an example, repeating the same content in kindergarten as in preschool would not be a sustaining environment for the preschool boost. Having a high percentage of peers who are well prepared for kindergarten might spark their teacher to increase rigor and therefore sustain the boost. Second, their "foot-in-the-door" hypothesis

posits that attending preschool may get children over an important hurdle in their K-plus experiences and thereby grant them access to a benefit or allow them to avoid a harm (e.g., unwarranted special education placement). They also hypothesized that another key to convergence of outcomes of attenders and nonattenders could be which skills are emphasized and measured in the prekindergarten through third-grade period. They argue for a boost to last, the focal skills must be malleable, fundamental for success, and unlikely to develop in the counterfactual. The boost from a prekindergarten program that focuses on constrained skills (Snow & Matthews, 2016)—for example, the discrete set of basic literacy and mathematics skills that almost all children master by third grade such as letter knowledge and simple counting—is likely to be less enduring than the boost from a program that focuses on students' deeper unconstrained skills, meaning more broadband skills like world knowledge, vocabulary, conceptual thinking, and problem solving.

At this juncture, it remains unclear which of these theories best describes patterns in the empirical evidence base. Empirically, in the medium term, the older evidence has shown that preschool has small-to-moderate effects in reducing grade retention and special education placement in the K-12 years (McCoy et al., 2017; Yoshikawa, Weiland, & Brooks-Gunn, 2016). In both older and more recent studies, language, literacy, and mathematics test scores between preschool participants and nonparticipants tend to partially or fully converge by the end of third grade, though some studies do show some evidence of medium-term persistence (e.g., Bassok, Gibbs, & Latham, 2018; Hill, Gormley, & Adelstein, 2015; Ladd, Muschkin, & Dodge, 2014; Lipsey, Farran, & Durkin, 2018; Phillips et al., 2017; Puma et al., 2012). Recent work on the trajectory of effects suggests most of the eventual medium-term convergence between preschool attenders and nonattenders occurs within 1–2 years after preschool (Hojman, 2015). Specifically, about half of the eventual convergence on cognitive outcomes occurs during kindergarten and then by about half again by the end of second grade (Li et al., 2016).

The relatively small number studies that have followed preschool participants into adulthood have found long-term benefits such as increases in college enrollment, decreases in incarceration rates, and decreases in teen pregnancy, even when in the medium-term there is convergence in test scores (Deming, 2009; Gibbs, Ludwig, & Miller, 2011; Yoshikawa et al., 2016). However, the jury is still out on whether today's preschool programs will yield long-term benefits to participants and society similar to programs from earlier decades, particularly in settings in which participants show medium-term fadeout. By necessity, all the longer term evidence is from participants who attended preschool decades ago and there are important differences in context between older studies versus those of today's preschools and preschoolers. Parents of all social classes today invest more time and money in their children's learning, on average, than in previous generations (Bassok, Finch, Lee, Reardon, & Waldfogel, 2016; Reardon, 2011). Also, more children attend nonparental care than in the past, changing the counterfactual against which a given preschool program is evaluated (Chaudry et al., 2017). Previous work suggests the counterfactual plays a substantial role in preschool evaluations. In a re-analysis of the Head Start Impact Study, Feller, Grindal, Miratrix, and Page (2016), for example, found persistence of positive effects on language through first grade only for children who in the absence of Head Start would have been at home with their parents and not for children who otherwise would have been enrolled in another preschool program.

These more modern-day findings regarding the counterfactual are particularly relevant to this study because, as we detail further in our findings section, an unusually high percentage of our control group compliers (88%) attended a center-based preschool program other than the Boston program and only 6% stayed home with a parent. As such, our study sits between two types of studies: (a) studies that compare a given preschool program against a more mixed counterfactual; and (b) studies in which all children attend the same preschool program but some attend an enhanced version. Examples of the former include the recent Tulsa quasi-experimental propensity score studies in which Tulsa prekindergarten is compared to a counterfactual in which 48% of children were in other center-based preschool programs (Hill et al., 2015) and the Tennessee VPK study in which 34% of comparison group was in other center-based preschool programs (Lipsey et al., 2018). Examples of the latter type of study include a recent preschool mathematics curricula trial that followed children into first grade, with the treatment group receiving an enhanced preschool experience and the control children business as usual preschool (Jenkins et al., 2018).

Increasingly, the evolving prekindergarten context adds nuance to understanding the effects of today's programs as well as raises new outcomes of interest.

For example, one of our study's key medium-term outcomes—post-prekindergarten enrollment in the BPS—has not been a focus in the literature to date, though it increasingly is a focus of localities that administer such programs. The only relevant evidence we are aware of is a recent study that found that Tulsa prekindergarten alumni were somewhat more likely to persist in the Tulsa Public Schools than were non-Tulsa prekindergarten and non-Tulsa Head Start attenders through eighth grade (Gormley, Phillips, & Anderson, 2018). There are, however, on-the-ground reports that schools in both New Orleans and DC—in which parent choice is a central feature of school assignment—have chosen to offer prekindergarten as a strategy to attract and retain families (D. Ewen, personal communication, June 19, 2017; Weixler, Lincove, & Gerry, 2017). In DC, school re-enrollment is also now a measure of school success/progress under the Every Student Succeeds Act (U.S. Department of Education, 2017). Seattle has framed its prekindergarten program not explicitly as a family retention strategy but as part of its affordability agenda (Slote & Kelly, 2015)—though presumably greater affordability would allow more young families to stay in the city. We expect that both the increasingly competitive educational markets and the rising cost of living in large cities will lead to increasing focus on the effects of public prekindergarten on K+ enrollment decisions. And we expect more broadly that the changing prekindergarten landscape will lead to attention to other outcomes new to the literature.

The Boston Prekindergarten Program and the Boston K-3 Context

The Boston Public Prekindergarten program is a modern-day, relatively large-scale program that has been of interest in the recent literature because of its programmatic elements and its documented strong impacts on children's school readiness. The program began an expansion in 2005–2006, under the decree of then-Mayor Thomas Menino who argued that in addition to preparing children for school, the program could help attract families to the BPS who might otherwise leave or choose other options. The program is based entirely in the public schools, pays teachers on the same scale as K-12 teachers, and subjects teachers to the same educational requirements of K-12 teachers (e.g., a master's degree within 5 years). Furthermore, it is open to any child in the city, regardless of income. In our study years, about one third of all 4-year-olds in Boston enrolled in the program and about half of all children who enrolled in BPS kindergarten had attended BPS prekindergarten the year before (Shapiro, Martin, Weiland, & Unterman, 2019).

Since 2007, the program also has utilized a consistent curricula and coaching system. Specifically, the district implemented *Opening the World of Learning*, which targets children's early language and literacy skills and includes a social-skills component embedded in each unit, in which teachers discuss socioemotional issues with children and integrate emotion-related vocabulary words (Schickedanz & Dickinson, 2005). It also implemented *Building Blocks*, an early mathematics curriculum which covers both numeracy and geometry and has a heavy focus on verbal mathematical reasoning (Clements & Sarama, 2007a). Both curricula have shown positive effects on children's outcomes in other studies (Ashe, Reed, Dickinson, Morse, & Wilson, 2009; Clements & Sarama, 2007b; Clements, Sarama, Spitler, Lange, & Wolfe, 2011), though the evidence base for Building Blocks is stronger than that for OWL (Weiland & Yoshikawa, 2013).

In two of our four focal years (2007–2008 and 2008–2009), curricula implementation was supported via trainings and regular coaching, meaning weekly to biweekly on-site support from an experienced early childhood coach trained in both curricula. Thereafter, due to budget cuts, coaching was targeted to new teachers and to prekindergarten and kindergarten teachers in schools undergoing National Association for the Education of Young Children Accreditation, a quality assurance process used in early childhood settings nationally. On the whole, Boston's structural and programmatic choices make it fairly unique among public programs nationally which tend not to require master's degrees, usually do not pay prekindergarten teachers on the same scale as K-12 teachers, do not require a proven, consistent curriculum, and do not employ coaching (Barnett et al., 2017).

The quality of the Boston program has been investigated in prior work using standard classroom observational tools. Boston classrooms score similarly to other systems nationally on structural quality and on emotional support (Weiland, Ulvestad, Sachs, & Yoshikawa, 2013). However, Boston has the highest average instructional quality of a large-scale program to date (Chaudry, Morrissey, Weiland, & Yoshikawa, 2017), scoring, for example, in the 2009–2010 school year 1.7 to 2.4 *SD*s higher on this dimension than current Head Start quality nationally (Weiland, 2016). It also showed strong effects on the language, literacy, mathematics, and executive function skills at kindergarten entry of

children who attended the program in 2008–2009 in a large-scale regression discontinuity study that used the program's long-standing September 1 cutoff as its source of exogeneity (Weiland & Yoshikawa, 2013). Importantly, the care settings for control group children were relatively stronger than has typically been the case in past such studies because Massachusetts has some of the strongest child care standards nationally and approximately two-thirds of control-group children were enrolled in nonparental care during the treatment year, with about 57% in other center-based preschool programs (Weiland & Yoshikawa, 2013).

Post-prekindergarten in our study's focal years, district K-3 teachers implemented the literacy curriculum *Reading Street* and the mathematics curriculum *TERC Investigations.* These curricula do not have a strong evidence base compared to the pre-k curricula used in the district (Agodini, Harris, Thomas, Murphy, & Gallagher, 2010; Gatti & Petrochenkov, 2010; Ladnier-Hicks, McNeese, & Johnson, 2010; What Works Clearinghouse, 2013), nor were they supported by coaching and training as systematically or as frequently as the pre-k program's supports. Reflective of these differing investment levels, classroom quality data collected by the Wellesley Centers for Women in spring 2012 on 84 K-3 classrooms in BPS and in spring 2010 on 83 prekindergarten classrooms and reanalyzed by our study team show that prekindergarten classroom instructional quality was markedly higher on average than K-3 instructional quality (see Table 1 in Appendix S1). Notably, the district responded to this evidence and other related evidence by subsequently (not in our study focal years) developing its own K-2 curriculum and associated professional development program (Boston Public Schools, 2017).

Current Study

Using data from four cohorts of students who applied to the BPS prekindergarten program between the 2007–2008 and the 2010–2011 school year, we aimed to investigate the effects of enrolling in the Boston prekindergarten program versus students' other options. Ultimately, consistent with other lottery-based studies (explained in detail in the next section), we were able to leverage *oversubscribed first choice lotteries* to address our central research question: What is the effect of enrolling in a Boston prekindergarten program versus not at all on children's enrollment and persistence in BPS grades K-3; children's risk of being retained in grade in K-2 or of being classified as special-needs in K-3; and children's third-grade state-standardized test scores in mathematics and reading?

Method

Data Set

We use data from BPS and the Massachusetts Department of Elementary and Secondary Education. We begin with data on students' choices and baseline demographics during the BPS assignment process from the spring of the 2006–2007 through 2009–2010 school years (for enrollment in 2007–2008 through 2010–2011). We merge these data, using each student's unique identifier, with district and state administrative records covering the years students were age-eligible for prekindergarten (at age 4) through third grade.

Sample

Our sample comes from the population of students who applied to the Boston prekindergarten program for 4-year-olds. As shown in Appendix S1 Figure 1, in all, 12,740 families applied to the program in our focal years. Nearly 10,000 of these families applied to the district's school choice lottery (described in greater detail in the next section) in the spring before their child was age-eligible for the program. This is what we call the "standard process"; it included four rounds and from these rounds, we identified naturally occurring lotteries for students' first choice school involving 3,182 students, or 25% of all appliers and 32% of those who applied through the standard process. The distribution of the lottery sample across the four rounds of the standard process is as follows: 99% of the sample is drawn from round one, < 1% is drawn from round two, < 1% is drawn from round three, and no students are drawn from round four. Another 2,769 (22% of appliers) applied via a later process after the four rounds had concluded. Their applications were considered on a rolling, as-space-is-available basis and they are not part of our lottery sample.

Our lottery sample was diverse in their background characteristics. As shown in Table 1, for example, 35% of first choice lottery winners and 40% of control group members were Hispanic; about a quarter overall were Black; another quarter were White; and the rest (~13%) were Asian or Other. About 58% of first choice lottery winners was eligible for free/reduced lunch (57% for control

group children). A little over half spoke English at home, a quarter spoke Spanish, and around 20% spoke a non-English and non-Spanish home language.

Boston Prekindergarten School Assignment Process Details

Under the BPS's school choice plan, in the winter and spring of each school year, families could apply to up to 10 schools they wanted their child to attend for prekindergarten the following fall (i.e., unlike most other systems, children were not automatically assigned to their neighborhood school). Families were assigned different priorities to different schools based on criteria set by the district, such as sibling and walk zone priority, sibling priority only, walk zone priority only, and no priority (listed in order from most to least priority). Importantly, when there was more demand than supply for a given school, the assignment algorithm used family choice lists, school priorities, and a random number to randomly assign some students (and not others) to the school.

In the present analysis, we used data from students' first application to prekindergarten to identify naturally occurring lotteries among students with the same preference to the same oversubscribed school/program (e.g., two students who listed school A's regular education program who both had walk-zone priority to it) that listed the program as their first choice. As discussed in the following section, we constrained our sample to students' first choice lotteries as only these students clearly participated in the equivalent of an experiment. Appendix S2 includes more details on the district's school application and lottery process and how we identified the lotteries used in our analysis.

Appendix S1 Table 2 displays the number of applicants to the prekindergarten program in the focal years, the number of lottery sample members, and the percentage of lottery participants each year. As mentioned earlier, across the study years, 25% of applicants were in an experimental lottery. The percentage of district schools represented across years in the lotteries ranged from 67% to 83% across years. However, some schools were highly over-represented and others were under-represented in the lottery sample—for example, about half of the students competed for just seven schools (10% of schools with prekindergarteners during this time period) and about 75% competed for just 18 schools (26% of schools with prekindergarteners during this time period).

Ultimately, although all lottery winners were offered the opportunity to enroll in the BPS prekindergarten program, 91% did so, according to BPS administrative records. Approximately 90% of lottery winners enrolled in their first choice school, 2% enrolled in a school not in their initial choice list, and 9% did not enroll in Boston prekindergarten. While all the control group students lost the first lottery they competed in, an estimated 62% of them enrolled in the program either by coming off of a waitlist, winning a subsequent lottery, being assigned to an under-subscribed school farther down their choice list, or participating and being assigned in a subsequent

Table 1
Balance on Observables in the First Choice Lottery Sample

	Lottery winners	Control group	Estimated difference	*p*-Value
Race/ethnicity (%)				
Hispanic	35.22	39.90	−4.68**	.003
Black	25.00	23.35	1.65	.271
White	26.73	24.34	2.39	.144
Asian	10.13	7.28	2.85**	.005
Other	2.92	4.14	−1.22	.146
Male (%)	50.27	46.95	3.32	.126
Eligible for free/ reduced lunch (%)	57.66	56.68	0.98	.604
Age	4.51	4.53	−1.97	.117
Country of origin USA (%)	94.89	94.53	0.36	.701
Home language (%)				
English	52.18	55.06	−2.88	.133
Spanish	25.18	25.19	−0.01	.994
Other	22.64	19.75	2.89	.074
N children	1,101	2,081		

Note. There was a small amount of missing data on all baseline characteristics except age: 12 children (0.4%) were missing race/ethnicity and male information, 34 (1.1%) were missing male and free/reduced lunch information, 113 (4.2%) were missing country of origin information, and 5 (0.2%) were missing home language information. Means in the table were computed using nonmissing data. Values for first choice lottery winners are the simple means for each requisite group. Values for the difference between lottery winners and control group members are obtained from a regression of a given baseline characteristic on a series of indicator variables that identify each lottery plus an indicator variable that equals 1 for lottery winners and 0 for lottery losers. The coefficient on lottery indicator equals the difference in the mean baseline characteristic between lottery winners and control group members, respectively. The value for control group members equals the corresponding value for lottery winners minus the estimated difference between lottery winners and control group members. A two-tailed *t*-test was applied to the estimated differences. An *F*-test was used to assess the statistical significance of the overall difference between lottery winners and control group members reflected by the full set of baseline characteristics in the table. The resulting *F* value is not statistically significant ($p = .2004$). Statistical significance levels are indicated as: ** $p < .01$.

assignment round. Ultimately, roughly 13% of the control group enrolled in their first choice school, 29% enrolled in a school lower on their choice list, 23% enrolled in a school not on their initial choice list, and 35% did not enroll in Boston prekindergarten.[1] Taken together, this suggests an estimated BPS prekindergarten enrollment rate difference of 29 percentage points (91% minus 62%), a difference that is low but not uncommon in research designs utilizing naturally occurring lotteries within choice processes (Abdulkadiroglu, Angrist, Narita, & Pathak, 2015; Angrist, Cohodes, Dynarski, Pathak & Walters, 2016). In the data analysis section, we describe how we use these lotteries to estimate the effect of *enrolling* for the target of our analysis—those that won a seat in their first choice BPS prekindergarten and subsequently enrolled versus those lotteried out of the program who do not enroll (i.e., the compliers).

Outcomes

Enrollment and Persistence in BPS

From district administrative records, we coded whether students enrolled in the BPS in kindergarten through third grade. If a student enrolled in BPS at least 1 day in a given year, we set the enrollment variable for that year to 1 and to zero otherwise. From our yearly variables, we constructed a 0/1 coded "ever enrolled" variable for kindergarten to third grade and 0/1 "persistence" variable for continuous enrollment for kindergarten to third grade. As previously mentioned, we included these outcomes because though unusual in the literature, they are increasingly relevant to public prekindergarten programs and because attracting families to the BPS who might otherwise leave or choose other options was an original goal for the program.

Grade Retention and Special Needs Placement

From administrative records, we constructed year-by-year measures of children's K-2 grade retention and K-3 special needs placement, defined as having an Individualized Education Plan (IEP). We also constructed measures of whether the child was ever retained from K-2 or ever had an IEP in K-3.

Notably, since 2008, the average districtwide retention rate in Grades 1–3 in Boston has ranged from 2.9% to 7.5% at each grade level, meaning that the percentage of students ever retained by the end of third grade is around 10% (e.g., 2.9% and 7.5% averaged and multiplied by three; Massachusetts Department of Elementary & Secondary Education, 2013). Nationally, the yearly annual retention rate was 1.5% in 2010 (Warren, Hoffman, & Andrew, 2014). Regarding special education placement, Massachusetts has the second highest rate of special education placement in the United States (Hehir, Grindal, & Eidelman, 2012). Approximately 19% of BPS elementary-school students in 2012 had been diagnosed with a disability (T. Grindal, personal communication, June 9, 2013).

Third-Grade Standardized Test Scores

For third-grade reading and mathematics analyses, we use students' statewide mathematics and reading standardized tests. Cohorts 1, 2, and 3 took the Massachusetts Comprehensive Assessment System (MCAS) in third grade, the test used for state accountability purposes in Massachusetts (see Appendix S3 for psychometric details). In 2015, the state of Massachusetts gave districts the choice between continuing to administer the MCAS or administering instead a new mathematics and English Language Arts (ELA) exam based on the Common Core standards, called the Partnership for Assessment of Readiness for College and Careers (PARCC) assessment (Massachusetts Department of Elementary & Secondary Education, 2015). In all, 54% of districts in the state switched to the PARCC, whereas the rest continued to administer the MCAS. In the three largest school districts in the state—Boston, Worcester, and Springfield—individual schools chose which test to administer. In Boston, all but two schools with third-grade students chose to administer the PARCC.

Amidst these changes, the state recommended that researchers standardize students' estimated theta (i.e., IRT) scores when conducting analyses that require pooling across the MCAS and PARCC exams (Massachusetts Department of Elementary & Secondary Education, 2016). We followed this advice and standardized each student's theta score on the mean and standard deviation of all third graders within the BPS taking the given exam in that year. Test score data in this article accordingly can be interpreted as a given group's performance compared to the average BPS third grader. For both the MCAS and the PARCC, if students were

[1]These are simple counts on control group enrollment rates, presented for descriptive purposes. They approximate but not do perfectly reproduce the 62% crossover rate estimated using our main analytic model and presented in Table 3.

retained, we used their score from their first third-grade test administration.

Covariates

Using administrative records, we constructed a set of student-level covariates. We captured students' race/ethnicity using a set of dichotomous variables that identified whether a student was Asian, Black, Hispanic, White, or mixed/other. Similarly, we used a set of dichotomous variables to identify whether the students' home language was English only, Spanish, or another language. Using student birthdates, we calculated students' age as of September 1 in the year they were applying to prekindergarten. We also created dichotomous variables that identified whether the student was eligible for free-reduced priced lunch; whether the student was male; and whether the student's country of origin was the United States.

School Context Variables

To capture each student's school experience in every follow-up year, we drew on publicly available data from the Massachusetts Department of Elementary and Secondary Education (n.d.), which we merged on to each student's data row by follow-up year and enrolled school ID. If the student was enrolled in multiple schools in a given year, we used the value for the school in which the student was enrolled the longest. We included indicators of the school-level student sociodemographic characteristics—percentage of students from low-income families (see Appendix S3 regarding a definition change in this measure in our last study year); the school's percentage of English language learners; percentage of students with a non-English home language; percentage of students with disabilities; percentage of students who were African American, Asian, Hispanic, or White; and percentage of female students. For schools' academic context, we included the percentage of third-grade students who were proficient or higher on state ELA and mathematics standardized tests (for Cohorts 1–3, the MCAS and for Cohort 4, either the MCAS or the PARCC, depending on which was used in the students' schools in third grade). Finally, we also included measures of schools' percentage of licensed teachers, student/teacher ratio, percentage of teachers rated as exemplary or proficient in the state's rating system, percentage of teachers retained or remained working in the same position compared to the previous school year, the percentage of students who remain in the school throughout the school year (stability rate), and average class size. Percentage of teachers rated as exemplary or proficient in the state's rating system and average class size were available for Cohorts 3 and 4 only. We averaged the characteristics of students' schools across the K-3 grades to create our key analytic variables.

Prekindergarten Year Care Settings

For our first two cohorts, when students applied to the BPS, their parents answered a set of questions about their child's last child care experience. We used these data to identify the care setting of children not enrolled in BPS prekindergarten (e.g., the counterfactual)—Head Start, private preschool, family day care, or parental/relative care. We also used state administrative records that captured whether a student attended preschool in a traditional public school or a charter school. We used district administrative records from the prekindergarten year to identify which sample children attended BPS prekindergarten.

The district changed its data collection form for this information for Cohort 3 and Cohort 4 such that setting type was not available to our study team. For this reason, we used control group care setting data for the first two cohorts only. More details on these data are available in Appendix S3.

Data Analytic Plan

To estimate the impacts of enrolling in the BPS prekindergarten program on study outcomes, we utilized naturally occurring lotteries in the Boston choice system. As is common when applying this experimental, lottery-based approach, our first step was to estimate the effect of *being offered the opportunity* to enroll in a Boston prekindergarten school (intent-to-treat [ITT]) using students' first choice applications (Abdulkadiroğlu et al., 2011; Bloom & Unterman, 2014; Dobbie & Fryer, 2011). We constrained our sample to students' first choice lotteries because when a student is competing in any lottery other than her first lottery, her probability of being assigned to a lower choice may depend in part on her earlier choices (and not just her random number) and thus using these later lotteries could pose a threat to randomization (Bloom & Unterman, 2014). While students may not compete in a lottery for their first choice and may compete in a lottery for a later choice, we focus only on their first choice lotteries to ensure that we have identified a purely experimental sample.

Within our lottery-based research design, a set of students randomly "won" the opportunity to attend their first choice BPS prekindergarten program (the treatment group). Another set of students randomly "lost" the opportunity to attend their first choice BPS prekindergarten program (the control group). Because the lottery randomly assigns students, students in the treatment and control groups were, in expectation, equivalent in all measurable and unmeasurable characteristics. The basic approach for the analysis is to estimate, for each lottery, differences in mean outcomes for winners and control group members, and to average the results across lotteries.

Specifically, we construct the following linear regression model:

$$Y_{ij} = \beta T_{ij} + \sum_{k=1}^{K} \pi_k I_{kij} + \sum_{p=1}^{P} \theta_p X_{pij} + \varepsilon_{ij}, \qquad (1)$$

where Y_{ij} is a relevant short- or medium-term outcome for student i in lottery j; T_{ij} is a lottery winner indicator equal to 1 if student i wins lottery j and 0 otherwise; I_{ij} is a set of k lottery indicators equal to 1 for lottery j and 0 otherwise; X_{ij} is a set of p student-level covariates (race/ethnicity, gender, eligibility for free or reduced-price lunch, age, country of origin, and home language status); and ε_{ij} is a random error for student i that is clustered by the prekindergarten school that students entered after their lottery. This latter information is available only for students that enroll in the Boston prekindergarten program. Thus for this purpose, we assume that students who do not enroll in the program—the majority of whom are in the control group—are not clustered together in another setting. The $\hat{\beta}$ coefficient identifies the effect of winning a lottery on student outcomes and its associated t-statistic identifies statistical significance.

Our ITT estimates represent the effect of winning one's first lottery and thus do not answer the question likely of most substantive interest to practitioners and policymakers—the effect of *enrolling* in Boston prekindergarten versus not doing so. Thus we use students' first lottery participation as an instrument for estimating the effects of *BPS prekindergarten enrollment*— often referred to in the literature as a complier average causal effect (CACE; Gennetian, Morris, Bos, & Bloom, 2005). In this context, the effect of enrollment represents the effect of enrolling in Boston prekindergarten for the subgroup of students—the compliers—who won their first choice lottery and enrolled in BPS prekindergarten compared with those that lost their first choice lottery and ultimately did not enroll in BPS prekindergarten. As previously mentioned, the overwhelming majority of lottery winners enrolled in their first choice school, making our enrollment effect more specifically represent *the effect of enrolling in children's first choice program*, versus not at all.

Because this approach is new to estimating the effects of prekindergarten (though it has been used in contexts with older children; Abdulkadiroğlu et al., 2011; Bloom & Unterman, 2014; Dobbie & Fryer, 2011), it merits some additional explanation. In particular, what is a *complier* in this context? A complier is a student who randomly won or lost his/her first choice lottery and, for winners, enrolled in a Boston prekindergarten program (first choice or otherwise), and for lottery losers, did not enroll in Boston prekindergarten. Notably, some children who lost their first lottery won a slot to a school lower on their choice list and attended Boston prekindergarten in that school. Our estimates of the effect of enrollment ultimately do not apply to them (i.e., known as "always takers" in the literature), just as they do not apply to children who would not have enrolled in the program regardless of whether they won or lost their first lottery (i.e., "never takers"). Our instrument effectively carves out the exogeneous variation in enrollment that is due children's first choice lottery result and uses it to estimate the causal effect of enrollment *for the subgroup of compliers*. Notably, as previously mentioned, effects for compliers in some contexts have been shown not to generalize to the full population (Chyn, 2018). This aspect of our design is why we emphasize that our analysis is a *window* into the medium-term effects of the program, rather than an evaluation necessarily for all students. This is also why we also conduct multiple analyses of the generalizability of our results.

To calculate CACE, we conducted a two-stage least squares analysis. The lotteries we drew on range in size with many of their samples being quite small. To avoid finite sample bias from "weak instruments" (Bound, Jaeger, & Baker, 1995), we estimated the CACE using a single-instrument model (also known as a Wald estimate). This approach has been used for past analyses of randomized experiments and lottery-based studies (Abdulkadiroğlu et al., 2011; Bloom & Unterman, 2014; Gennetian et al., 2005; Ludwig & Kling, 2007). The first stage was specified as:

$$E_{ij} = \beta T_{ij} + \sum_{k=1}^{K} \pi_k I_{kij} + \sum_{p=1}^{P} \theta_p X_{pij} + w_{ij}, \qquad (2)$$

where E_{ij} is a BPS prekindergarten enrollment indicator equal to 1 if student i ever enrolled in BPS

prekindergarten and 0 otherwise, and all other terms are defined as in Equation 1. Our first-stage *F*-statistic equals 11, which is *just* above the recommended threshold for instrument strength (Bloom, Zhu, & Unlu, 2010; Bound et al., 1995).

The second stage equation was specified as:

$$Y_{ij} = \delta \hat{E}_{ij} + \sum_{k=1}^{K} \pi_k I_{kij} + \sum_{p=1}^{P} \theta_p X_{pij} + e_{ij}, \qquad (3)$$

where $\hat{E}_{ij}$ equals the fitted value of the enrollment outcome from the first-stage equation, e_{ij} is a random error that is clustered by the prekindergarten school that students entered after their lottery, and all other terms are defined as in Equation 1. The estimated value of δ is a consistent estimate of the average effect of enrolling in BPS prekindergarten for target BPS prekindergarten enrollees. We fit our CACE models in SAS, using MDRC code described in detail in Bloom and Unterman (2014). Importantly, while the ITT approach meets the What Works Clearinghouse's highest standard of evidence, the CACE approach is considered quasi-experimental (What Works Clearinghouse, 2014).

Finally, there was a small amount of missing data on all covariates except age in our lottery sample, ranging from 0.4% to 4.2% (and likewise in our full sample, ranging from 0.4% to 4.0%; see the Table 1 note). We imputed missing covariates as our primary approach in our lottery approach, using multiple imputation with 40 data sets. Our lottery estimates are not sensitive to problems of missing covariate data (see the "Robustness Checks" section in the following section). We describe missing data on outcomes in the next section.

Results

Balance on Observables and Attrition Analysis

We compared the background characteristics of first choice lottery winners and control group members in the lottery sample (see Table 1 for this ITT analysis). There are 2 (of 12) statistically significant differences between the two groups—lottery winners are 4.7 percentage points less likely to be Hispanic (p = .003) and 2.9 percentage points more likely to be Asian (p = .005). A joint *F*-test used to assess the statistical significance of the overall difference between the first choice lottery winners and control group members could not reject the null hypothesis that there was no difference between the two groups (p = .200). We controlled for these background characteristics (as is suggested by What Works Clearinghouse, 2014), both to improve precision and, for the characteristics for which there was evidence of imbalance, to reduce the threat of possible bias in our estimates. See Appendix S1 Table 3 for the estimated complier averages for these same background characteristics. On average, compliers were quite similar to the full ITT lottery sample.

In an analysis of the availability of our outcome data for our first choice lottery sample, we found that outcome data were missing at relatively low rates (3%–16%) and that differences in outcome missingness by treatment status were relatively small (1–5 percentage points more likely to be missing for the control group, across outcomes). These levels of missingness meet the What Works Clearinghouse's (2014) standards for rigor. Six out of 11 differences across outcomes were statistically significant at conventional levels. The resulting F-value from a joint *F*-test of differences in the background characteristics of children with nonmissing outcome data by treatment status was not statistically significant. See Appendix S4 for full attrition details and results.

Care Settings in the Prekindergarten Year

Table 2 displays results from fitting our ITT and CACE models with care setting information in the prekindergarten year for Cohorts 1 and 2 as the outcomes, for children who had nonmissing counterfactual data. The results shown are important for identifying what Boston prekindergarten is being compared to in our study. From our ITT results, 97% of lottery treatment group members enrolled in BPS prekindergarten and nearly all the treatment group members (99.6%) enrolled in some kind of center-based preschool. In the lottery control group, 97% enrolled in some kind of center-based preschool—substantially exceeding the national average of 69% (Whitehurst & Klein, 2015). Overall, 72% of control group members enrolled in BPS prekindergarten, 14% in private centers, 4% in Head Start, 4% in charters, 3% in other public programs, 1% in family day cares, and 2% were at home with a family member.

By definition, for our BPS enrollment effect estimates, all our treatment group compliers attended BPS and none of control group compliers did so. Among control group compliers, 88% were in other center-based preschool programs. All told, 48% of control group compliers attended private programs, 17% Head Start, 12% charters, 12% other public

Table 2
Children's Care Settings (Cohorts 1 and 2) in the Prekindergarten Year

	ITT				CACE			
	Lottery winners	Control group	Estimated difference	*p*-Value	Lottery winner compliers	Control group compliers	Estimated difference	*p*-Value
Any center-based preschool	**99.55**	**96.72**	2.83***	< .001	**100.00**	**88.40**	11.60	< .001
Preschool types								
BPS	96.64	72.26	24.39***	< .001	100.00	0.00	100.00***	< .001
Non-BPS center-based preschool	2.91	24.47	−21.56***	< .001	0.00	88.40	−88.40***	< .001
Private	2.01	13.62	−11.60***	< .001	0.00	47.57	−47.57***	< .001
Head Start	0.00	4.18	−4.18***	.008	0.00	17.14	−17.14***	.008
Public	0.00	2.81	−2.81***	< .001	0.00	11.54	−11.54***	< .001
Charter	0.89	3.86	−2.96	.067	0.00	12.15	−12.15	.067
Other settings								
Family day care	**0.00**	**1.39**	–1.39	.372	**0.00**	**5.71**	–5.71	.372
At home	**0.45**	**1.88**	–1.44	.512	**0.00**	**5.88**	–5.88	.512
Total	**100.00**	**100.00**	—	—	**100.00**	**100.00**	—	—

Note. Care setting types were reported by parents at the time of application to Boston kindergarten (e.g., the winter, spring, or summer preceding kindergarten fall), were pulled from Boston prekindergarten enrollment records, or were pulled from age 4 state administrative records on traditional public school or charter school enrollment. Values were obtained from fitting our primary ITT and CACE equations with each care setting as the requisite outcome. Data were missing for 11.5% of students. Bolded numbers sum to 100. Statistical significance levels are indicated as: *$p < .05$; **$p < .01$; ***$p < .001$. ITT = intent-to-treat; CACE = complier average causal effect; BPS = Boston Public Schools.

programs, 6% family day cares, and 6% were at home.

In recent preschool evaluations, about a third to half of the control group has attended other center-based preschools programs (e.g., 34% in Tennessee, 48% in Tulsa, about 50% in Head Start; Bloom & Weiland, 2015; Hill et al., 2015; Lipsey et al., 2018). These lottery-sample counterfactual findings accordingly are quite distinctive within the current evidence base.

Impacts

Examining lottery sample members' K-3 enrollment in the BPS, we found that first choice lottery winners enrolled in the BPS at higher rates at each grade compared to the control group. As shown in Table 3 Column 2, effects of winning a first choice lottery (ITT) ranged from about 7 to 10 percentage points ($p < .0001$). The effects for compliers who enrolled in BPS (CACE, see Column 6) at each grade level K-3 were large, ranging from 24 to 34 percentage points ($p < .0001$). There was also a large difference of 34 percentage points in consistent K-3 enrollment in the district between lottery winner compliers and control group compliers ($p < .0001$). Treatment and control group complier means shown in Column 8 and 9 further illuminate these findings; 74% of lottery winner compliers enrolled in BPS continuously from K-3 versus just 39% of control group compliers. These findings demonstrate that pre-K enrollment markedly increased later enrollment in district public schools.

For other examined outcomes—children's grade retention, special education placement, and standardized test scores—findings in Table 3 demonstrate that there were no effects of winning a first choice lottery (ITT; Column 2) nor of enrolling in Boston prekindergarten (CACE; Column 6). The effects of winning a first choice lottery (ITT) for these outcomes were uniformly small, close to zero in magnitude, and not statistically significant. Compliance rates across these outcomes were around 29 percentage points (see Column 4; $p < .0001$). CACE estimates are larger than the ITT estimates; given the compliance rate of ~29 percentage points across outcomes, the magnitude of the CACE estimates reflects the low compliance rate difference. Also notable, CACE confidence intervals (Column 10) were relatively wide, ranging from substantially negative to substantially positive. For example, for "ever placed in special education," the point estimate was 0.8% with a 95% confidence interval of –10% to 12%.

For these outcomes too, the treatment and control complier means are illuminating (see Columns 8–9). In kindergarten, for example, very few students were retained in grade—1.6% of treatment

Table 3
First Choice Lottery Sample ITT Impacts, Compliance Rates Difference, and CACE

	ITT	ITT (*SE*)	First stage (compliance)	First stage *SE*	CACE	CACE (*SE*)	Lottery winner compliers mean	Control group compliers mean	Confidence interval
Enrolled in BPS (%)									
Prekindergarten	29.47***	1.06	—	—	—	—	100	0	—
Kindergarten	7.02***	1.73	.29***	.01	23.81***	5.97	91.06	67.25	12.51, 35.89
First grade	10.14***	1.90	.29***	.01	34.41***	6.53	85.60	51.19	22.16, 48.77
Second grade	8.49***	2.00	.29***	.01	28.80***	6.88	79.05	50.24	15.78, 42.76
Third grade	7.54***	2.04	.29***	.01	25.57***	7.02	75.77	50.2	12.22, 39.75
Enrolled K-3	10.05***	2.07	.29***	.01	34.10***	7.13	73.58	39.49	20.68, 48.62
Ever enrolled	8.24***	1.48	.29***	.01	27.97***	5.09	100	72.03	18.45, 38.39
Retained in grade (%)									
Retained in kindergarten	0.04	0.59	.29***	.01	0.14	2.05	1.55	1.41	−3.87, 4.16
Retained in first grade	1.24	0.80	.29***	.01	4.28	2.77	4.19	−0.09	−1.14, 9.71
Retained in second grade	−0.03	0.62	.29***	.01	−0.09	2.15	1.93	2.01	−4.30, 4.12
Ever retained	1.22	1.10	.30***	.01	4.20	3.79	7.50	3.31	−3.23, 11.63
Special education classification (%)									
SPED in kindergarten	−0.36	1.26	.29***	.01	−1.23	4.36	7.47	8.71	−9.78, 7.30
SPED in first grade	0.21	1.41	.29***	.01	0.73	4.87	10.55	9.81	−8.82, 10.28
SPED in second grade	1.5	1.53	.29***	.01	5.18	5.27	13.56	8.39	−5.15, 15.51
SPED in third grade	0.01	1.66	.29***	.01	0.02	5.74	15.94	15.91	−11.22, 11.27
Ever SPED	0.25	1.67	.30***	.01	0.84	5.77	17.27	16.42	−10.46, 12.15
MCAS and PARCC									
English language arts	0.01	0.04	.29***	.01	0.02	0.13	0.40	0.38	−0.24, 0.28
Math	−0.05	0.04	.29***	.01	−0.18	0.14	0.35	0.53	−0.45, 0.10

Note. There was no missing data on enrollment variables. Other outcomes were missing data as follows: grade retention 3%–11% across variables; special education 4%–11% across variables; and test scores, 13%–16%. Note that we also calculated ITT and CACE effect sizes for MCAS and PARCC (the continuous outcomes) by dividing the estimated effect by the standard deviation of the control group and found they were nearly identical in magnitude to the ITT and CACE estimates shown in the table. Statistical significance levels are indicated as: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. ITT = intent-to-treat; CACE = complier average causal effect; BPS = Boston Public Schools; MCAS = Massachusetts Comprehensive Assessment System; PARCC = Partnership for Assessment of Readiness for College and Careers.

compliers and 1.4% of control group compliers. These levels are substantially below the aforementioned district average of 2.9% to 7.5% in Grades 1–3 at each grade level in BPS (Massachusetts Department of Elementary & Secondary Education, 2013). For special education, mean levels for lottery compliers more closely approximate the district average of 19% of BPS elementary-school students diagnosed with a disability. Specifically, about 16% of treatment and control compliers were classified as special education students in third grade. For state-standardized tests, both groups scored substantially higher than the average BPS third grader. In ELA, lottery compliers scored 0.40 *SD* higher than the average BPS third grader, whereas control group members scored 0.38 *SD* higher than the average BPS third grader. In math, the means were 0.35 *SD* for lottery compliers and 0.53 for control compliers—a more sizable difference compared to other outcomes but statistically not significant.

Differences in Students' K-3 School Experiences

To provide context for these results, we examined whether differences in students K-3 contexts might have driven our findings, concentrating on the differences in contexts for compliers. Specifically, as described in the measures section earlier, we used publicly available school characteristics data and student enrollment records and calculated the average characteristics of the school students were enrolled in from grades K-3. We then analyzed these student-specific measures as outcomes using our standard CACE model. As shown in Table 4, there were some statistically significant, though relatively small, differences in 9 of 18 characteristics of treatment complier and control complier K-3 environments. For example, treatment compliers had fewer peers who were low-income students (66% vs. 72%, $p < .05$) and African American (26% vs. 37%, $p < .001$) than did control compliers. Treatment compliers also had

Table 4
CACE Estimates of K-3 School Context Differences Between Lottery Winner Compliers and Control Group Compliers

	Lottery winner compliers	Control group compliers	Estimated difference	*p*-Value
Student background characteristics				
% Low-income	65.62	72.21	−6.59*	.019
% ELL	28.35	26.41	1.94	.305
% non-English home language	39.58	35.77	3.81	.065
% Disabilities	17.45	17.81	−0.36	.631
% African American	26.34	36.98	−10.64***	< .001
% Asian	8.07	10.39	−2.33*	.034
% Hispanic	40.43	32.54	7.89**	.001
% White	22.17	16.04	6.12*	.030
% Female	48.23	48.47	−0.24	.511
Student performance—% proficient in third grade				
ELA	44.28	40.61	3.67	.117
Math	51.29	47.69	3.60	.144
Teacher and school characteristics				
% Licensed Ts	97.29	93.79	3.50**	.003
Student-T ratio	13.76	13.23	0.53*	.011
% Exemplary Ts	14.63	11.41	3.22	.109
% Proficient Ts	78.85	79.74	−0.89	.680
% T retention	81.36	78.42	2.94**	.009
Stability	89.12	83.90	5.21***	< .001
Avg class size	19.06	18.44	0.61	.258

Note. Using publicly available data from the Massachusetts Department of Elementary and Secondary Education, we averaged available school-level data across the schools in which a student was enrolled for the longest period of time each year in kindergarten, first, second, and third grade. If data were missing for a student in a given year (e.g., first grade), we used nonmissing data to compute the student's K-3 context averages (e.g., K, second, third). Across variables, data were missing for 8%–11% of students overall and 5% of treatment students were missing data compared with 13% of their control group counterparts. Percentage of teachers scoring proficient or exemplary on state ratings and average class size was available for Cohorts 3 and 4 only. Statistical significance levels are indicated as: *$p < .05$; **$p < .01$; ***$p < .001$. CACE = complier average causal effect; ELA = English Language Arts; ELL, English Language Learners; Ts = teachers.

more peers who were Hispanic (40% vs. 33%, $p < .01$) and White (22% vs. 16%, $p < .05$). They also experienced slightly more licensed teachers (97% vs. 94%, $p < .01$) and were in schools with more stable student bodies (89% stable vs. 84%, $p < .001$). The percentage of children proficient on third-grade tests in children's K-3 schools favored the treatment group compliers (e.g., 44% vs. 41% for ELA), but the difference was not statistically significant.

On the whole, while there were lottery-induced differences in students' K-3 school experiences favoring the treatment group, these were relatively small. Both groups of students attended elementary schools in which their peers were majority low-income and in which the majority of teachers were rated as exemplary or proficient by the state's teacher evaluation system.

Robustness Checks

As a robustness check, we fit third-grade outcome models in the first choice lottery sample without multiple imputation of covariates or outcomes and with multiple imputation for both covariates and outcomes (vs. our primary strategy of imputing covariates but not outcomes; see Appendix S5 Table 1). We also fit school context models that used characteristics of students' K and third-grade schools only, in case averaging over different numbers of years for students with missing data in one or more of their K-3 years was distorting or misrepresenting the schooling context differentials (Appendix S5 Table 2). We re-fit third-grade standardized test models dropping children who were ever retained, in case taking the test at an older age or in a different year from the rest of the cohort somehow biased our estimates even though there were no impacts on retention (for parsimony, these results are available upon request). Also, most of our outcomes were dichotomous; we used linear probability models as our primary modeling strategy because our sample size is well over the threshold for doing so and as these models are more straightforward (Angrist &

Pischke, 2008). We did, however, refit key models with dichotomous outcomes using logistic regression as a sensitivity check (results available upon request). We also refit impact models with prekindergarten enrollment defined as being enrolled in at least 150 days of the school year (rather than 1 day; see Appendix S5 Table 3). We chose 150 days because of the distribution of the enrollment variable in our sample; there is no agreed-upon threshold in the literature and recent preschool studies have used different thresholds (Lipsey, Farran, & Hofer, 2015; Phillips, Gormley, & Anderson, 2016). Across these checks, we found no evidence that our main results were sensitive to our data analytic decisions.

Finally, one assumption underlying our CACE analysis—that always-takers in both the treatment and control groups (i.e., children who would have enrolled in Boston prekindergarten regardless of their first choice treatment assignment status) experienced the same effect of enrollment—is difficult to evaluate. Treatment group always-takers in our study largely enrolled in their first choice school; among control group crossovers, approximately a third did so. If the level at which a student ranked a prekindergarten program is indicative of their match with the program or its quality, it is possible that the two-thirds of the control group crossovers that enrolled in lower choices experienced a lower quality program. Empirically, when we compared the first choice schools and BPS schools actually attended in prekindergarten for the two-thirds of control group crossovers who did not enroll in their first choice but enrolled in a lower choice, we found that their first choice and their school attended differed on 9 of 13 school context characteristics, with first choice school appearing generally somewhat higher quality than the school in which they actually enrolled. However, we also found that school-level context variables were only weakly correlated with observed prekindergarten process quality (Weiland & Unterman, 2019). Furthermore, among all control crossovers, about two-thirds were unassigned to the program after their first round; they were not assigned to a lower choice as part of the first round of the lottery system. These students crossed over later, which might indicate that their parents were particularly highly motivated and that therefore, they might have benefitted *more* from their Boston prekindergarten classroom. Ultimately, the direction of any potential bias from violation of the always-taker CACE assumption is ambiguous.

Gauging External Validity

Following other lottery-based studies (i.e., Abdulkadiroğlu et al., 2011), we explored the external validity of our results using descriptive and quasi-experimental analyses. This work was important in our context, given that (as previously mentioned) students in the lottery sample comprised 25% of all appliers in focal years. Also, some schools were highly over-represented and others were under-represented in the lottery sample—for example, about half of the students competed for just seven schools (10% of schools with prekindergarteners during this time period) and about 75% competed for just 18 schools (26% of schools with prekindergarteners during this time period).

To explore external validity, we first compared the *background characteristics* of first choice lottery sample members to those of children in the full applicant sample. As shown in Appendix S6 Table 1, while the two samples appeared to be similar in age, country of origin, and gender, the lottery sample was more economically advantaged and more likely to be White than all BPS prekindergarten applicants. About 51% of the lottery sample qualified for free-/reduced-price lunch, whereas 65% of all BPS appliers did. Regarding students' race/ethnicity, White students comprised 28% of the lottery sample versus 17% of all BPS prekindergarten appliers; Hispanic students comprised 39% of the lottery sample versus 44% of all BPS prekindergarten appliers. About 21% of the lottery sample was Black versus 28% of the full applicant sample. Fifty-seven percent of the lottery sample spoke English at home versus 50% of the full sample.

Next, we compared the *comparison group care settings* of our lottery sample to the full applicant sample. Among children whose families applied to the BPS prekindergarten but did not enroll (i.e., full sample nonenrollees), 76% attended a non-BPS center-based preschool (vs. 97% of lottery control group members and 88% of lottery control group compliers) and types were markedly different from those in the lottery control group (see Table 2 and Appendix S6 Table 6). For example, 37% of full sample nonenrollees attended private centers, 26% attended Head Start, and 13% attended other public programs. For lottery control compliers, 48% attended private centers, 17% attended Head Start, and 24% attended other public programs. In all, 18% of full sample nonenrollees were at home versus 6% of lottery control group compliers.

Using K-3 school context data, we also examined the representativeness of *schools* in our lottery-based analysis. We defined over-represented schools as schools for which 50% ($N = 6$ schools) or 75% ($N = 17$ schools) of first choice lottery sample members competed. As shown in Appendix S1 Table 4, over-represented schools (75% threshold) in our lottery-based study had a considerably lower average percentage of students from low-income families compared to other district schools (64% vs. 77%, respectively; $p < .001$), proportionately more White students (24% vs. 13%, respectively; $p < .01$), more teachers rated as exemplary by administrators under the state's teacher evaluation system (23% vs. 14%, $p < .05$; Massachusetts Department of Elementary & Secondary Education, 2017), and more third graders scoring advanced/proficient on state standardized tests (47% vs. 33% for ELA, $p < .001$; 54% vs. 41% for math, $p < .01$, respectively). Our lottery estimates therefore are heavily weighted toward applicants to schools with more advantaged, higher performing students than in the district overall, though the over-represented schools also enrolled majority low-income and non-White students.

Finally, we used a propensity-score approach to estimate the relationship between BPS prekindergarten enrollment and our key outcomes and thus to examine the representativeness of our *lottery-based estimates*. Specifically, we predicted the probability that a student would be treated conditional on their background characteristics, their cohort year, and the public school each student lived closest to as a proxy for neighborhood characteristics. We then inverted these propensities to obtain an inverse probability weight (IPW) that we could use in our subsequent regression analysis to counteract selection into the program (Imbens & Wooldridge, 2009; Murnane & Willett, 2010). The covariates available for this work are the key covariates in our impacts work (i.e., race/ethnicity, gender, free/reduced lunch, age, country of origin, and home language; see Table 1). The exception is that in our additional analysis we add a fixed effect for the closest public elementary school to the student, a proxy for neighborhood which we use because lottery blocks by definition are not available for the full sample. These covariates are considerably less rich than those in some other recent prekindergarten evaluations which have been able to include covariates such as parent education, home literacy measures, Internet availability in the home, and number of working parents (Hill et al., 2015; Lipsey et al., 2015; Phillips et al., 2016). Therefore, to gauge whether they captured selection into the program and following Abdulkadiroğlu et al. (2011), we began by replicating our lottery-based findings first.

Specifically, we replicated the lottery-based ITT findings by estimating the association between being in the first choice lottery sample treatment group with grade retention, special education, and test scores in third grade using the sample of *all* students who applied to Boston prekindergarten during the four application rounds ($N \sim 9{,}700$). As shown in Table 5, with this replication sample, we found results that were very similar to our lottery ITT estimates, with the exceptions of a marginally significant and larger result on special education placement in Kindergarten (0.36 percentage points ITT compared with −1.68 percentage points for the replication sample) and a marginally significant result of similar magnitude on third-grade math scores (−0.05 ITT and −0.06 replication). We then used our IPW approach to estimate the association between BPS prekindergarten enrollment and later outcomes on the full sample of prekindergarten applicants and enrollees ($N \sim 11{,}790$), effectively including in the replication sample an additional group of students who enrolled in prekindergarten but did not apply through the standard process (see Appendix S6 for more information on available data in the full sample and other details on our IPW approach). With the full sample, we find larger and statistically significant associations between prekindergarten enrollment and grade retention outcomes (enrollees were 4 percentage points less likely to be retained in grades K-3, $p < .001$), special education placement (enrollees were 7 percentage points less likely to be placed in special education in grade K-3, $p < .001$), and MCAS scores (enrollees scored 0.04 *SD*s higher than the average BPS third grader on both Math and ELA, $p < .05$). We view these findings as best interpreted as *associations* for gauging external validity and not as causal estimates; the internal validity of our IPW findings is bolstered somewhat by the lottery-based validation but ultimately, it is difficult to assess the internal validity of these findings.

Taken together, our analyses gauging external validity point to a first choice lottery sample that was more advantaged than the full applicant sample and raise caution in generalizing our lottery sample findings to all applicants and enrollees.

Discussion

While the evidence is clear that children who attend preschool have stronger school readiness skills at

kindergarten entry than children who do not attend preschool (Duncan & Magnuson, 2013; Phillips et al., 2017), the longer run evidence base on large-scale prekindergarten programs is just emerging. In the current study, we used a rigorous lottery-based approach as a window into the effects of one such program, the BPS prekindergarten program, on key child outcomes through the end of third grade. We also examined counterfactual care settings and K-3 settings for first choice lottery-sample children to contextualize our results and we explored the generalizability of our key results beyond the lottery sample.

For special education placement, retention, and standardized test scores, in the first choice lottery sample, we found no differences in outcomes through third grade between first choice lottery winners who enrolled in BPS prekindergarten and control group members who did not. We did find evidence that Boston prekindergarten succeeded in drawing families into the BPS and in retaining them, which was one of the program's original goals. Effects for compliers on enrollment and persistence in the BPS were large—about 91% of lottery winners who enrolled in Boston prekindergarten also enrolled in BPS kindergarten, versus just 67% of control group members who did not enroll in Boston prekindergarten. Overall, 74% of lottery winner compliers were enrolled in BPS from K-3 versus only 39% of control compliers. In increasingly competitive urban educational markets, offering prekindergarten in the public schools appears to be one avenue for attracting and retaining families that might otherwise enroll elsewhere.

Notably results like these—medium-term convergence of outcomes for prekindergarten attenders and nonattenders in the early elementary grades in our lottery sample—is a common (though not universal) finding overall in the literature (Phillips et al., 2017). The *why* behind this pattern is a puzzle and one that likely has no consistent answer across study contexts, given the wide range in program quality, counterfactuals, child demographics, and elementary school quality nationally. In our context, our results could be seen as surprising, given the high quality of the Boston program and its promising short-term effects on children's school readiness (Weiland & Yoshikawa, 2013). Several factors are highly important in placing our results within the broader context.

First, as we emphasize throughout the article, our analysis is not an evaluation of the effects of all Boston prekindergarten programs, for the full sample of children who attended. Rather, lotteries were highly concentrated in a subset of schools; 75% of lottery applicants, for example, competed for about a quarter of eligible district schools. There were also important differences between children in our first choice lottery-based sample and the full sample generally appearing more advantaged. On the one hand, more popular schools might be higher quality and thus more effective and thus we

Table 5
Lottery ITT Results and IPW Results for the Lottery Replication Sample and the Full Prekindergarten Applicant Sample

	Lottery sample (ITT)	Replication sample	Full sample
Retained in grade (%)			
Retained in kindergarten	0.04 (0.59)	−0.59 (0.37)	−2.46 (0.38)***
Retained in first grade	1.24 (0.80)	0.58 (0.57)	−1.36 (0.45)**
Retained in second grade	−0.03 (0.62)	−0.35 (0.45)	−0.31 (0.37)
Ever retained	1.22 (1.10)	−0.26 (0.77)	−3.87 (0.64)***
Special education classification (%)			
SPED in kindergarten	−0.36 (1.26)	−1.68* (0.83)	−5.28 (0.70)***
SPED in first grade	0.21 (1.41)	−0.61 (0.96)	−5.92 (0.76)***
SPED in second grade	1.50 (1.53)	0.48 (1.06)	−5.31 (0.81)***
SPED in third grade	0.01 (1.66)	0.47 (1.11)	−5.78 (0.85)***
Ever SPED	0.25 (1.67)	−0.01 (1.12)	−6.51 (0.85)***
Third-grade test scores			
English language arts	0.01 (0.04)	−0.01 (0.03)	0.04 (0.02)*
Math	−0.05 (0.04)	−0.06* (0.03)	0.04 (0.02)*

Note. For the lottery sample (ITT), outcomes were missing data as follows: grade retention 3%–11% across variables; special education 4%–11% across variables; and test scores, 13%–16%. For the replication sample, outcomes were missing data as follows: grade retention 6%–10% across variables; special education 3%–10% across variables; and test scores, 15%. For full sample, outcomes were missing data as follows: grade retention 9%–14% across variables; special education 6%–14% across variables; and test scores, 20%. Statistical significance levels are indicated as: *$p < .05$; **$p < .01$; ***$p < .001$. ITT = intent-to-treat; IPW = inverse probability weight.

might have expected *more* persistence of impacts among this sample. However, our generalizability work showed that students enrolled in prekindergarten in these schools appeared more advantaged and thus might have been less likely to benefit from the program than their less advantaged peers in less popular schools. In that case, *less* persistence of impacts might be expected among the lottery compliers. Supporting this hypothesis, in our propensity score work, we found associations suggestive of small benefits for the full population of Boston prekindergarten enrollees on all examined outcomes. Descriptively, lottery sample control group compliers were also quite high performing, scoring 0.38–0.53 *SD* higher than the average BPS third grader on standardized math and literacy tests, versus 0.15–0.17 *SD* for the full sample non-enrollees. Non-enrollees were also less likely to attend other preschool programs than lottery sample control group compliers and more likely to persist in BPS.

Second, previous research has shown that the counterfactual matters greatly in preschool studies (Feller et al., 2016). In our lottery sample, a large majority of the lottery control group attended a center-based preschool program (97% ITT, 88% CACE; see Table 2). More typically, about a third to half of the control group has attended other center-based preschools programs in large-scale causal evaluations of publicly funded preschool programs (e.g., 34% in Tennessee, 48% in Tulsa, about 50% in Head Start; Bloom & Weiland, 2015; Hill et al., 2015; Lipsey et al., 2015). Ours is not a test of preschool versus no preschool; rather, our results indicate that compliers who attended a free public prekindergarten program versus largely a mix of other preschool programs did about equally well at the end of third grade. This may be because ultimately, the treatment-control contrast (Bloom & Weiland, 2015) may not have been large enough to generate lasting impacts for our lottery sample. Unfortunately, we lacked information on the quality of control group care settings that would have allowed us to identify the full treatment-control contrast. Notably, in the regression-discontinuity (RD) evaluation of the Boston program described earlier in this article that found strong impacts on children's school readiness skills, parents of control group children reported that in the year their children were too young to enter the Boston prekindergarten program (e.g., their age 3 year), 57% experienced another type of center-based care and 33% were in parental care (Weiland & Yoshikawa, 2013)—considerably higher than the national average of 42% of 3-year-olds enrolled in preschool programs (Whitehurst & Klein, 2015) but far fewer than in our lottery-based study.

A third reason for nuanced interpretation is that we lack information on children's kindergarten entry skills and thus were unable to identify whether lottery complier children had experienced an initial boost from Boston prekindergarten compared to control compliers. In other words, interpreting our results as either surprising or expected in terms of persistence is complicated by not knowing whether compliers' experienced benefits from the program in the first place. Underscoring this point, few of the children who participated in the previous RD evaluation study of Boston prekindergarten that showed strong impacts on kindergarten readiness (Weiland & Yoshikawa, 2013) were included in our lottery sample. Specifically, the RD sample represented approximately 85% of district schools and 70% of eligible children in those schools in 2008–2009. Only 125 children were in both the previous study's RD treatment group and the current study's lottery winner group—constituting about 47% of the lottery treatment group for the 2008–2009 school year and only 13% of the RD treatment group overall.

The K-3 schooling experiences of children in our sample are also important to highlight to place our results in the context of the "sustaining environments hypothesis"—the idea that sustaining the boost from preschool depends on the quality of K+ schooling environments (Bailey et al., 2017). So far, the evidence on this hypothesis is mixed (Bassok, Gibbs, & Latham, 2018; Bierman et al., 2014; Clements, Sarama, Wolfe, & Spitler, 2013; Jenkins et al., 2018; Johnson, 2013; Swain, Springer, & Hofer, 2015; Zhai, Raver, & Jones, 2012). We found that the quality of K-3 programming in Boston was lower on average than that of the district's prekindergarten program (see Appendix S1 Table 1). Notably, Massachusetts and Boston do show higher performance relative to other states and similar districts nationally, respectively (National Center for Education Statistics, 2013; Reardon, 2017). But relative to other districts in the state, Boston in our focal years had relatively weak third-grade performance, scoring around the bottom 11% of districts on the state third-grade standardized math test and the bottom 5% of districts for third-grade reading (Massachusetts Department of Elementary & Secondary Education, 2014). Also, for our study's cohort years, prekindergarten to third-grade alignment reforms (see Boston Public Schools, 2017) had not yet taken place in the district. Prekindergarten

attenders during our study years may have repeated some of the same content in kindergarten, offering an opportunity for control compliers to catch up; content repetition has been associated with less growth in kindergarteners' math skills in a nationally representative study (Engel, Claessens, & Finch, 2013). However, ultimately, simply knowing that K-3 quality was lower than prekindergarten quality and that Boston scored lower than most other districts does not answer the question of whether there is a threshold of quality needed to sustain effects. More nuanced measurement would have been required to answer questions about threshold effects.

In addition to contributing to the field's understanding of medium-term convergence patterns, our study's lottery-based design also has methodological implications for the field. As recently reviewed by a group of experts (Phillips et al., 2017), the rigor of longitudinal studies of today's large-scale preschool programs thus far has been mixed. This is due in part to the difficulties of randomly assigning children to a given preschool program in localities that already have universal preschool (i.e., Oklahoma, West Virginia), as well as to difficulties not unique to preschool in gaining buy-in/agreement from local stakeholders. However, with the recent expansion of public preschool programs in contexts that, like Boston, use lottery-based assignment algorithms to assign children to preschool (e.g., Washington DC, Denver, San Francisco, New York, and New Orleans), the field is seemingly poised for additional rigorous studies of the impact of public preschool.

Our lottery sample findings drive home the importance of understanding the characteristics of students in a city-based school lottery versus all students receiving the program and the lottery-induced treatment contrast, especially within naturally occurring randomized trials. This may be particularly important in contexts with prekindergarten programs that, like Boston's, are open to families of all income levels. Families with higher social capital are likely to be better at navigating choice and lottery systems than other families and may be over-represented in prekindergarten lottery studies similar to ours. In addition, as explained earlier, our study's lottery-induced treatment contrast amounted to comparing sample members first choice Boston prekindergarten programs to other mostly private and other public preschool options. This is generally not the policy question of interest to policymakers seeking to expand access to preschool. Given that most 4-year-olds now attend some form of center-based care, future lottery studies may likely to encounter this situation as well. Depending on the context, future lottery studies may be better poised to compare different preschool programs to each other than to answer the preschool versus none question.

There are several limitations that should be highlighted. The measures in our study were limited to those available via administrative records. Measures of other important school readiness and success skills such as children's socioemotional and executive function skills were not available. Our knowledge of the program's effects is accordingly more limited than we would like. Also, as explained in the robustness check section, one assumption underlying our CACE analysis—that always-takers in both the treatment and control groups (i.e., children who would have enrolled in Boston prekindergarten regardless of their treatment assignment status) experienced the same effect of enrollment—is difficult to evaluate.

In closing, unpacking the preschool convergence phenomenon is one of the most pressing issues facing the field of early education research (Phillips et al., 2017). Rigorous research on today's programs is beginning to catch up to the rapid pace of preschool expansion nationally, through efforts like the present article; efforts in North Carolina (Dodge, Bai, Ladd, & Muschkin, 2016), Tulsa (Hill et al., 2015; Phillips et al., 2017), and Tennessee (Lipsey et al., 2018); and the five place-based teams (including in Boston; McCormick et al., 2019) tracking children from preschool to third grade in the Institute of Education Sciences (2016) Early Learning Network. In addition, the field is potentially poised for additional rigorous lottery-based studies that permit longitudinal analysis like in this article—though it remains to be seen what policy questions these studies will be able to answer. Our lottery-based findings, combined with our analysis of the relevant counterfactual and our quasi-experimental work on the full sample, contribute to the new generation of public preschool studies that will hopefully help point the way to ensuring a stronger, lasting boost for all children.

References

Abdulkadiroğlu, A., Angrist, A., Dynarski, S., Kane, T., & Pathak, P. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics*, *126*, 649–748. https://doi.org/10.1093/qje/qjr017

Abdulkadiroglu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2015). *Research design meets market design: Using*

centralized assignment for impact evaluation (No. w21705). Cambridge, MA: National Bureau of Economic Research.

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders. NCEE 2011–4001*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*, 275–318. https://doi.org/10.1086/683665

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

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

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*, 7–39. https://doi.org/10.1080/19345747.2016.1232459

Barnett, W. S., Friedman-Krauss, A. H., Weisenfeld, G. G., Horowitz, M., Kasmin, R., & Squires, J. H. (2017). *The state of preschool 2016: state preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Bassok, D., Finch, J. E., Lee, R., Reardon, S. F., & Waldfogel, J. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *AERA Open*, *2*. https://doi.org/10.1177/2332858416653924

Bassok, D., Gibbs, C. R., & Latham, S. (2018). Preschool and children's outcomes in elementary school: Have patterns changed nationwide between 1998 and 2010? *Child Development*. https://doi.org/10.1111/cdev.13067

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 one year later in different kindergarten contexts. *Child Development*, *85*, 140–159. https://doi.org/10.1111/cdev.12117

Bloom, H. S., & Unterman, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management*, *33*, 290–319. https://doi.org/10.1002/pam.21748

Bloom, H., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socioemotional skills using data from the National Head Start Impact Study*. MDRC Working Paper. New York, NY: MDRC.

Bloom, H. S., Zhu, P., & Unlu, F. (2010). *Finite sample bias from instrumental variables analysis in randomized trials*. MDRC Working Paper. New York, NY: MDRC.

Boston Public Schools. (2017). *Focus on K2*. Retrieved from https://sites.google.com/bostonpublicschools.org/earlychildhood/focus-on-k2?authuser=0

Bound, J., Jaeger, A., & Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*, 443–450. https://doi.org/10.1080/01621459.1995.10476536

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2017). *Cradle to kindergarten: A new plan to combat inequality*. New York, NY: Russell Sage.

Chyn, E. (2018). Moved to opportunity: The long-run effect of public housing demolition on children. *American Economic Review*, *108*, 3028–3056.

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

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

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

Clements, D. H., Sarama, J. H., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *4*, 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*, 812–850. https://doi.org/10.3102/0002831212469270

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

Dobbie, W., & Fryer, Jr., R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3*, 158–87. https://doi.org/10.1257/app.3.3.158

Dodge, K. A., Bai, Y., Ladd, H. F., & Muschkin, C. G. (2016). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*, *88*, 996–1014. https://doi.org/10.1111/cdev.12645

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

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. https://doi.org/10.3102/0162373712461850

Feller, A., Grindal, T., Miratrix, L., & Page, L. (2016). Compared to what? Variation in the impact of early childhood education by alternative care type. *Annals of Applied Statistics*, *110*, 1245–1285.

Gatti, G. G., & Petrochenkov, K. (2010). *Pearson reading street efficacy study 2009–10 final report*. Pittsburgh, PA. https://www.pearsoned.com/wp-content/uploads/reading-street-efficacy-study-2009-2010_final.pdf

Gennetian, L., Morris, P., Bos, J., & Bloom, H. (2005). Using instrumental variables analysis to learn more from social policy experiments. In H. Bloom (Ed.), *Learning more from social experiments: evolving analytic approaches* (pp. 75–114). New York, NY: Russell Sage.

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). *Does Head Start do any lasting good?* (NBER Working Paper No. 17452). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w17452

Gormley, Jr., W. T., Phillips, D., & Anderson, S. (2018). The effects of Tulsa's pre-K program on middle school student performance. *Journal of Policy Analysis and Management*, *37*, 63–87.

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

Hehir, T., Grindal, T., & Eidelman, E. (2012). *Review of special education in the commonwealth of Massachusetts*. Report for the Massachusetts Department of Elementary and Secondary Education, Boston, MA. Retrieved from http://www.doe.mass.edu/sped/2012/0412sped.pdf

Hill, C. J., Gormley, W. T., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly*, *32*, 60–79. https://doi.org/10.1016/j.ecresq.2014.12.005

Hojman, A. (2015). *Evidence on the fade-out of IQ gains from early childhood interventions: A skill formation perspective*. Working paper, University of Chicago, Center for the Economics of Human Development, Chicago, IL.

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

Institute of Education Sciences. (2016). *IES launches research network on early childhood education*. Retrieved from https://ies.ed.gov/whatsnew/pressreleases/01_19_2016.asp

Jenkins, J. V. M., Watts, T., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. (2018). Do high quality kindergarten and first grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness*, *11*, 339–374. https://doi.org/10.1080/19345747.2018.1441347

Johnson, R. (2013). *School quality and the long-run effects of Head Start*. Working paper, University of California, Berkeley, CA.

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

Ladnier-Hicks, J., McNeese, R. M., & Johnson, J. T. (2010). Third grade reading performance and teacher perceptions of the Scott Foresman Reading Street program in Title I schools in South Mobile County. *Journal of Curriculum and Instruction*, *4*, 51–70. https://doi.org/10.3776/joci.2010.v4n2p51-70

Li, W., Leak, J., Duncan, G. J., Magnuson, K., Schindler, H., & Yoshikawa, H. (2016). *Is timing everything? How early childhood education program impacts vary by starting age, program duration and time since the end of the program*. Working Paper.

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

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

Ludwig, J., & Kling, J. (2007). Is crime contagious? *Journal of Law and Economics*, *50*, 491–518. https://doi.org/10.1086/519807

Massachusetts Department of Elementary and Secondary Education. (2013). *Appendix B: Retention rates by district and school by Grade: 2011–12*. Retrieved from http://www.doe.mass.edu/infoservices/reports/retention/

Massachusetts Department of Elementary and Secondary Education. (2014). *2014 MCAS report (DISTRICT) for grade 03 all students*. Retrieved from http://profiles.doe.mass.edu/state_report/mcas.aspx

Massachusetts Department of Elementary and Secondary Education. (2015). *Spring 2015 district assessment decision update*. Retrieved from http://www.doe.mass.edu/news/news.aspx?xml:id=13541

Massachusetts Department of Elementary and Secondary Education. (2016). *Working with 2015 Massachusetts assessment data: Advisory from the Office of Planning and Research and the Office of Student Assessment Services*. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2017). *The Massachusetts framework for educator evaluation*. Retrieved from http://www.doe.mass.edu/edeval/

Massachusetts Department of Elementary and Secondary Education. (n.d.). *About the data*. Retrieved from http://profiles.doe.mass.edu/help/data.aspx?section=students#selectedpop

Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*, 491–495. https://doi.org/10.1017/S0954579410000222

McCormick, M. P., Weiland, C., Hsueh, J., Maier, M., Hagos, R., Snow, C., . . . Schick, L. (2019). Promoting content-enriched alignment across the early grades: A study of policies & practices in the Boston Public Schools. *Early Childhood Research Quarterly*. https://doi.org/10.1016/j.ecresq.2019.06.012

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., . . . Shonkoff, J. P. (2017). Impacts of early childhood education on medium-and long-term educational outcomes. *Educational Researcher*, *46*, 474–487. https://doi.org/10.3102/0013189X17737739

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

National Center for Education Statistics. (2013). *NAEP state comparisons*. Retrieved from http://nces.ed.gov/nationsreportcard/statecomparisons/

Phillips, D., 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. https://doi.org/10.1037/dev0000151

Phillips, D., Lipsey, M., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., . . . Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects*. Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2017/04/consensus-statement_final.pdf.

Puma, M., Bell, S. H., Cook, 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.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan, & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–116). New York, NY: Russell Sage Foundation.

Reardon, S. F. (2017). *Studying education inequality and opportunity with big data*. Plenary address at the Society for Research in Educational Effectiveness spring conference, Washington, DC. Retrieved from https://www.sree.org/video/index.php?fullScreen=Yes&item=2017SBall2

Sameroff, A. (2009). *The transactional model of development: How children and contexts shape each other*. Washington, DC: American Psychological Association. https://doi.org/10.1037/11877-000

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

Shapiro, A., Martin, E., Weiland, C., & Unterman, R. (2019). If you offer it, will they come? Patterns of application and enrollment behavior in a universal prekindergarten context. *AERA Open*, *5*, https://doi.org/10.1177/2332858419848442

Slote, D. R., & Kelly, J. (2015). *Agreement will bring affordable housing to neighborhoods across Seattle*. Retrieved from http://www.seattle.gov/news/newsdetail_council.asp?ID=15140

Snow, C. E., & Matthew, T. J. (2016). Reading and language in the early grades. *The Future of Children*, *26*, 57–74. https://doi.org/10.1353/foc.2016.0012

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*, 475–485. https://doi.org/10.1007/s11121-014-0513-z

Swain, W. A., Springer, M. G., & Hofer, K. G. (2015). Early grade teacher effectiveness and pre-K effect persistence: Evidence from Tennessee. *AERA Open*, *1*. https://doi.org/10.1177/2332858415612751

Tipton, E. (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics*, *39*, 478–501. https://doi.org/10.3102/1076998614558486

U.S. Department of Education. (2017). *District of Columbia revised state template for the consolidated state plan: The Elementary and Secondary Education Act of 1965, as amended by the Every Student Succeeds Act*. Retrieved from https://osse.dc.gov/sites/default/files/dc/sites/osse/documents/OSSE%20ESSA%20State%20Plan_%20May%202%202017.pdf

Warren, J. R., Hoffman, E., & Andrew, M. (2014). Patterns and trends in grade retention rates in the United States, 1995–2010. *Educational Researcher*, *43*, 433–443. https://doi.org/10.3102/0013189X14563599

Weiland, C. (2016). Launching Preschool 2.0: A road map to high-quality public programs at scale. *Behavioral Science & Policy*, *2*, 37–46. https://doi.org/10.1353/bsp.2016.0005

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*, 199–209. https://doi.org/10.1016/j.ecresq.2012.12.002

Weiland, C., & Unterman, R. (2019). *By what factors do parents of young children rank schools? Evidence from Boston*. Manuscript in preparation.

Weiland, C., & Yoshikawa, H. (2013). The impacts of an urban public prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills: Evidence from Boston. *Child Development*, *84*, 2112–2130. https://doi.org/10.1111/cdev.12099

Weinstein, R. S. (2004). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press.

Weixler, L. B., Lincove, J. A., & Gerry, A. (2017). *The provision of public pre-k in the absence of centralized school management*. Manuscript in preparation.

What Works Clearinghouse. (2013). *WWC intervention report: Investigations in number, data, and space*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_investigations_021213.pdf

What Works Clearinghouse. (2014). *WWC procedures and standards handbook (Version 3.0)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse.

Whitehurst, G., & Klein, E. (2015). *Do we already have universal preschool?*. Washington, DC: Economic Studies at Brookings.

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, *26*, 21–35. https://doi.org/10.1353/foc.2016.0010

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. New York, NY: Foundation for Child Development, Society for Research in Child Development.

Zhai, F., Raver, C. C., & Jones, S. (2012). Academic performance of subsequent schools and impacts of early interventions: Evidence from a randomized controlled trial in Head Start settings. *Children and Youth Services Review*, *34*, 946–95. https://doi.org/10.1016/j.childyouth.2012.01.026

Supporting Information

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

Appendix S1. Additional Key Figures and Analytic Tables

Appendix S2. Identifying Lotteries in the Boston Public School's School Assignment Process

Appendix S3. Additional Measures Details

Appendix S4. Attrition Analysis

Appendix S5. Robustness Checks

Appendix S6. Details on the Full Applicant Sample and Inverse Probability Weight Work

APPENDIX A: Additional key figures and analytic tables

Figure 1: Application process for the full analytic sample

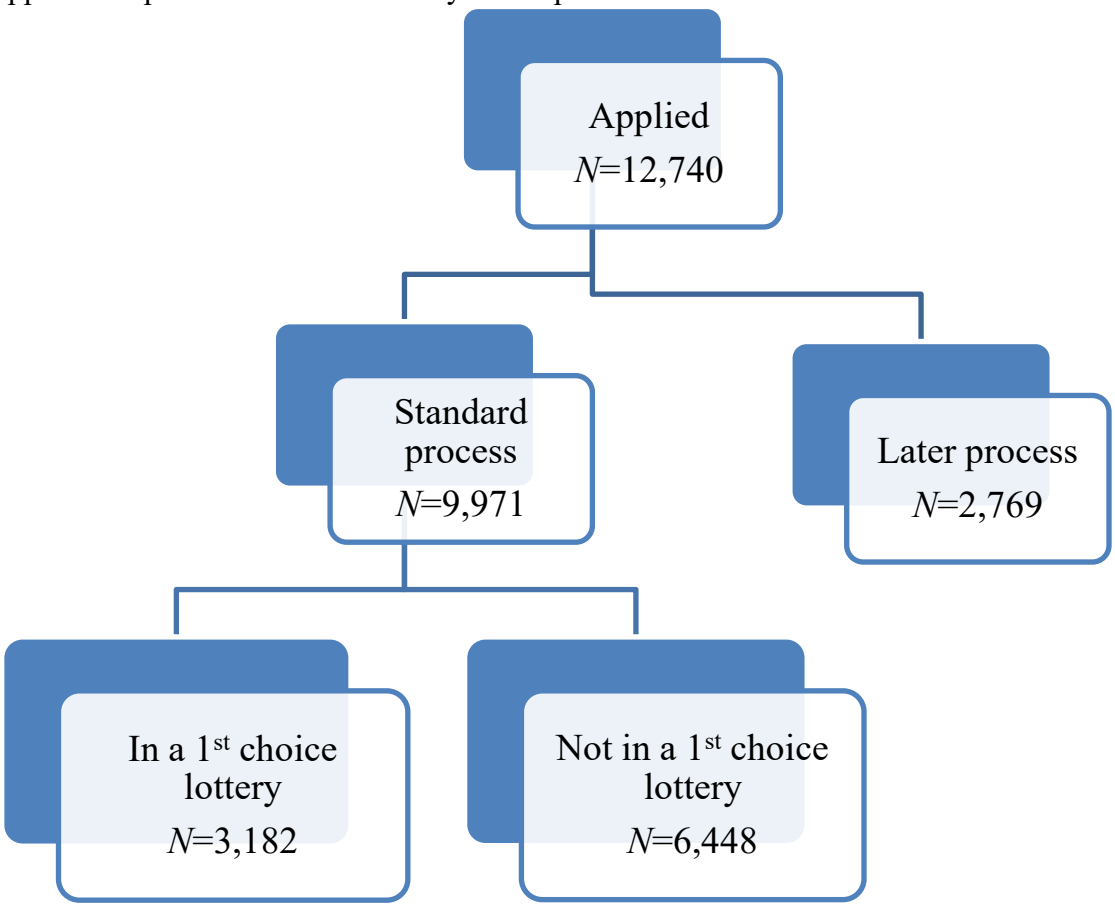


Table 1: Boston prekindergarten vs. K-3 classroom quality scores

	Prekindergarten (*N*=83 classrooms)			K-3 (*N*=84 classrooms)				
	Mean	SD	Range	Mean	SD	Range	Diff	Std diff
CLASS								
Emotional Support	5.63	0.60	4.00-6.83	5.11	1.05	1.75-6.69	-0.52	-0.87
Instructional Support	4.30	0.84	2.22-5.67	3.85	1.14	1.82-6.17	-0.45	-0.54
Classroom organization	5.10	0.68	2.75-6.22	4.94	1.16	1.75-6.75	-0.16	-0.23
ELLCO								
Language, Literacy, and Curriculum	3.53	0.45	2.50-4.50	3.14	0.74	1.25-4.5	-0.39	-0.87

Note: Standardized differences were calculated by dividing the prekindergarten and K-3 difference in scores by the prekindergarten standard deviation. The ELLCO used in prekindergarten was the Early Language and Literacy Classroom Observation tool (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002) while the ELLCO used in K-3 was the K-3 edition (Smith, Dickinson, & Sangeorge, 2002). To make the two version more comparable, for prekindergarten we used the eight items in the language, literacy, and curriculum scale and for K-3, the eight items most similar to the items in the prekindergarten scale (e.g., prekindergarten scale included oral language facilitation, while the K-3 scale included opportunities for extended conversations and effort to build vocabulary).

Table 2: Lottery sample by year

Cohort year	N Applicants	N Lottery Participants	% of Lottery Participants	N Schools	N Schools with Lotteries	% of Schools with Lotteries	N (%) schools 75% of lottery students competed for
Spring 2007	2,182	651	30	60	40	67	10 (17%)
Spring 2008	2,333	765	33	64	46	72	12 (19%)
Spring 2009	2,613	859	33	63	45	71	16 (25%)
Spring 2010	2,843	907	32	64	53	83	19 (30%)

Note: The total number of applicants across the four years was 9,971. The total number of lottery participants was 3,182 (25% of the total number of applicants).

Table 3: K-3 school context differences between schools over-represented in the lotteries versus other district schools

	Over-rep. schools (50% threshold)	Other schools	Diff	Over-rep. schools (75% threshold)	Other schools	Diff
Student Background Characteristics						
% Low-income	60.88	75.16	-14.28**	63.94	77.32	-13.38***
% ELL	21.89	31.77	-9.88	26.28	32.35	-6.07
% non-Eng. home lang.	34.04	42.29	-8.25	36.75	43.16	-6.40
% Disabilities	16.99	18.87	-1.88	18.28	18.80	-0.52
% African-American	27.07	31.06	-3.99	26.05	32.42	-6.37
% Asian	6.11	6.99	-0.89	7.67	6.57	1.09
% Hispanic	41.24	44.07	-2.83	39.93	45.27	-5.34
% White	23.00	14.80	8.19	23.59	12.61	10.98**
% Female	48.42	47.85	0.57	47.84	47.95	-0.11
Student Performance – % Proficient in 3rd grade						
ELA	49.55	35.51	14.04*	47.01	33.29	13.72***
Math	55.53	43.61	11.92	54.33	41.40	12.93**
Teacher and school characteristics						
% Licensed Ts	97.90	98.07	-0.17	98.12	98.02	0.10
Student-T ratio	13.66	13.28	0.38	13.54	13.23	0.30
% Exemplary Ts	23.11	15.62	7.49	22.58	13.95	8.62*
% Proficient Ts	72.74	78.76	-6.02	73.10	80.13	-7.03*
% T retention	80.63	79.16	1.47	80.04	79.05	0.99
Stability	91.04	86.32	4.72*	90.91	85.25	5.66***
Avg class size	18.66	18.05	0.61	18.60	17.93	0.67
N	6	56		17	45	

Notes: School-level averages were calculated using school context data from all available years between 2007-2008 and 2014-2015. Thresholds were determined by examining the distribution of lottery sample members across district schools. School context data was missing for 4 of the 62 schools represented in the lottery sample (1 over-represented school and 3 other district schools). Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent.

Appendix B: Identifying Lotteries in the Boston Public School's School Assignment Process

In this Appendix, we provide additional details on how the BPS assignment process works and how we identify lotteries within the process.

Since 1989 Boston Public Schools has used a centrally-administered school choice plan to assign students to schools. While the school choice plan has evolved over time,[1] in the winter and spring of 2007 (the year preceding the full implementation of the BPS prekindergarten program), students (and their families) submitted an application to the district with a ranking of up to 10 of their top schools.[2] There were four different rounds of this process and students could re-apply to other schools in later rounds if they did not like their assigned site from an earlier round or if they relocated to a different attendance zone.[3] Each round was administered in the same way, and the majority (roughly 70%) of the students applying to prekindergarten were assigned in Round 1. For simplicity we focus on Round 1 for the remainder of this section.[4]

In the BPS school choice plan schools also prioritized students. Specifically, in the study years prekindergarten schools ranked students using the following priority groups (in order of preference): Siblings, Walk zone priority[5], No priority. Almost all schools were open to "within-zone" families, as the district is divided into three attendance zones. Students could attend schools across zones only if the school had "district-wide" designation (1 elementary school) or if they lived within the walk zone of a school that is in an attendance zone other than the one in

[1] Choice systems based on the same algorithm Boston used in our focal years are also in use in other cities around the country, including in New York City (Abdulkadiroğlu, Pathak, & Roth, 2009; Bloom & Unterman, 2014).
[2] A student that did not rank any schools was assigned to a school with available seats and that school may or may not have been near the student's home. These students were not included in our analysis, as they were not involved in lotteries for particular schools.
[3] There are three attendance zones in total – North, West, and East.
[4] When analyzing data from later rounds, we identified lotteries using the method outlined in this section among students who were participating in the assignment process for the first time.
[5] A student is defined as "walk zone" if they live within 1 mile of the school.

which they lived. The assignment process was complicated further in prekindergarten by bilingual classroom assignments, as students who were deemed as eligible for bilingual services via a bilingual screening process competed not just for slots in a school but for slots in the school's bilingual classrooms. In some schools, students also could have competed for slots in regular education and inclusion classrooms (e.g., intentional mix of children with and without diagnosed special needs; see Weiland, 2016 for details).

Identifying lotteries

BPS uses an advanced algorithm to compute student assignments, taking the student preferences and school priorities described above into account (technically this algorithm is called a "student-proposing deferred acceptance algorithm"[6]). A critical first step of the algorithm is that it assigns every student a random number. Then, when a school's priority category is oversubscribed (i.e., there are more students with the given priority than the school can serve), the algorithm uses the students' random numbers to determine which students will be offered seats. This is the statistical equivalent of a lottery being held to assign students within the oversubscribed priority to the school.

Figure 1 illustrates what this process looks like for a hypothetical student. In the example, the student loses a lottery for her first-choice school based on her random number. This student becomes a control-group member in our analytic sample. This student is assigned (without a lottery) to her second choice school and enrolls there the following fall. As we discuss in the analytic methods section, this student will likely become a "control group cross-over" in our analysis of the effect of *enrolling* in Boston's prekindergarten program. Note,

[6] At the end of the "student-proposing deferred acceptance algorithm" it must be the case that if a student prefers another school to his/her final assignment, all of the seats at the other school are assigned to students with higher priority or with the same priority, but better random numbers.

whether this student was assigned to her second choice school through a lottery or not is irrelevant – the first lottery a student competes in determines her assignment to either the treatment group or control group and whether or not she enrolls the following fall determines how she will be classified in our enrollment effect analysis.

Figure 2 illustrates what this process looks like for a hypothetical school that can offer seats to 22 prekindergartners. BPS's algorithm attempts to assign to the school all students who list it as one of their choices and are not assigned to a more-preferred choice. The hypothetical school has 37 students who list it as a choice and do not receive a more preferred choice. Two students are siblings of students already at the school and receive a seat; 10 students are walk zone and receive a seat. The non-walk zone priority for this school is oversubscribed and a lottery (based on the students' random number) determines which students will be assigned to the school's 10 remaining seats.

All instances of oversubscription like the one described above form the basis of the study's analytic sample. However, there is one final constraint. Theoretically, if a student is competing in any lottery other than her first lottery, her probability of being assigned to a prior choice (and thus her probability of being assigned to the choice she is competing for) could be correlated with her characteristics, which would a pose a threat to randomization.[7] Accordingly, we restricted our lottery sample to students' first-choice schools.[8]

---

[7] For example, as stated in Bloom and Unterman, 2014, assume that academically weak students choose unpopular schools whose lotteries are weakly competitive (9 out of 10 participants win). Assume that all other students choose more popular schools with more competitive lotteries (1 out of 10 participants wins). Because weak students are in weakly competitive prior lotteries, those who lose these lotteries must have especially "poor" random numbers (otherwise they would have won). In contrast, because other students are in more competitive prior lotteries, students who lose these lotteries can have a mix of "superior" and "poor" random numbers. Carrying this example forward, in their next lottery, academically weak students will be more likely to lose than academically strong students.

[8] Others have been able to broaden their analytic sample past first lottery participants by employing a "risk set" approach. Unfortunately, given that students can list 10 choices, and within which they are

Kindergarten lotteries

The BPS also conducts lotteries for kindergarten, using the same algorithm and choice system as it uses for prekindergarten. Children who attend prekindergarten in the Boston Public Schools have the option to continue on in their prekindergarten school or apply to new schools for their kindergarten year. Important for our purposes, across all 4 years and all schools in our sample, the average school added 27 seats in kindergarten (range= -55-127; SD of 32). In percentage terms, this means the average school increased their capacity by about 150% between prekindergarten and kindergarten in our study years. The treatment identified through our lottery-based strategy thus is assignment to prekindergarten and not assignment to prekindergarten through third grade.

nested in 4 or more priorities, the number of risk sets required would be extremely large. Specifically, in the first round of the 2007 process, over 1,338 risk sets would be created and only 73 contain more than one student.

Figure 1.

BPS Assignment Process for a Hypothetical Student.

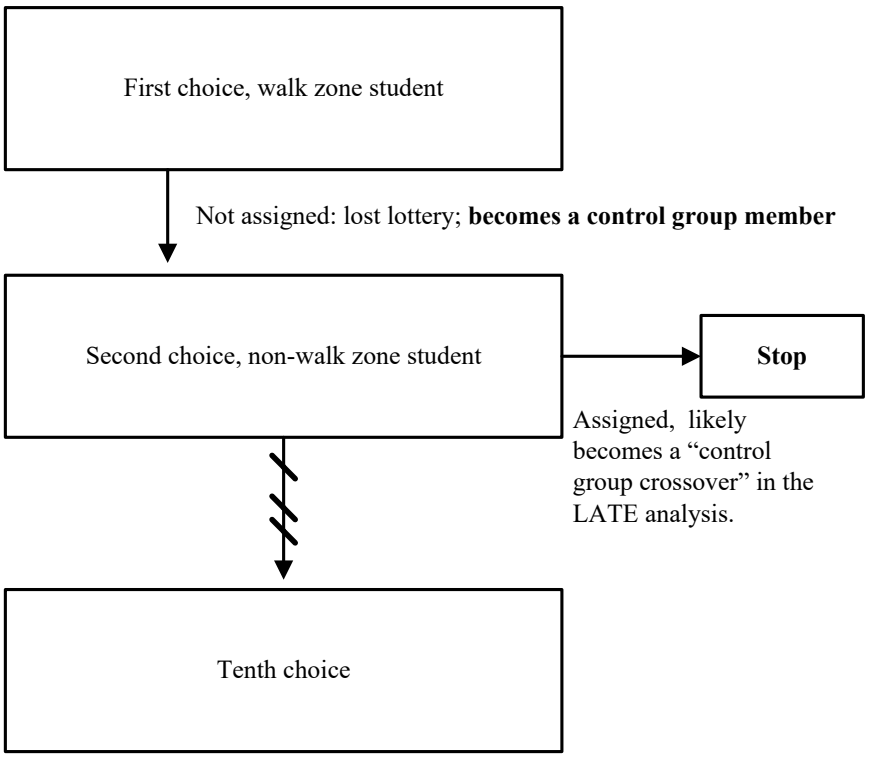
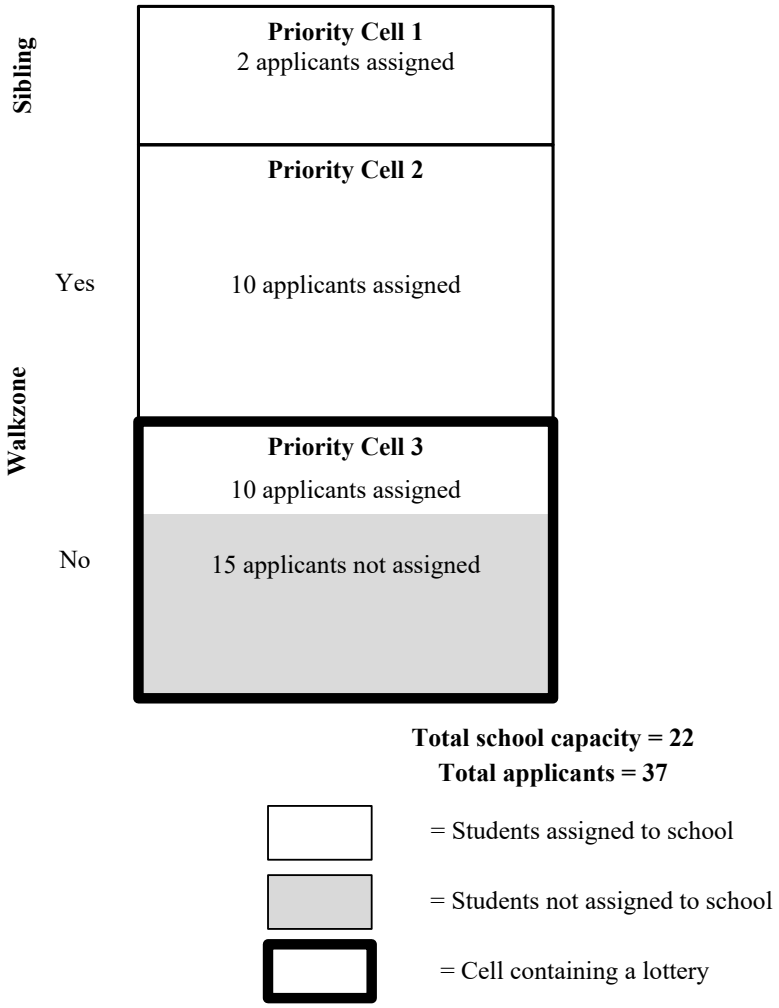


Figure 2.

BPS Assignment Process for a Hypothetical School



APPENDIX C: Additional measures details

Third grade standardized tests. The MCAS – the test taken by cohorts 1-3 and some of cohort 4 – is equated from year to year within year using an anchor test that is embedded within the actual MCAS test (Massachusetts Department of Elementary and Secondary Education, 2008). These items link performance standards on the original and subsequent MCAS tests, putting the within-grade scores on the same scale from year to year. Further, considerable care is taken in the equating process. Each year, psychometricians from two independent contractors independently and simultaneously equate the MCAS tests. The MA Department of Elementary and Secondary Education, with assistance as needed from its national Technical Advisory Committee, analyzes the results of the two independent equating analyses prior to reporting any MCAS test results. MCAS scores have been used in previous rigorous studies of educational interventions and have been shown to be sensitive to intervention effects (Abdulkadiroğlu et al., 2011; Angrist, Dynarski, Kane, Pathak, & Walters, 2010). Importantly, the results of the Massachusetts MCAS test regarding student proficiency match estimates of student proficiency as measured by an external test, the National Assessment of Educational Progress (NAEP; Bandeira de Mello, Blankenship, & McLaughlin, 2009) – the MCAS appears to be a valid assessment of students' reading and mathematics skills.

School context – percentage low income. For one of our school context variables – percentage low-income – the MA Department of Elementary and Secondary Education changed its definition in 2014-2015 (our last study) year. The prior definition counted as low-income any student who: 1) was eligible for free or reduced price lunch, 2) received Transitional Aid to Families benefits, and/or 3) was eligible for Supplemental Nutrition Assistance Program (SNAP). The updated measure added to this list students' foster care program status and

Medicaid status (called MassHealth; MA Department of Elementary and Secondary Education, n.d.). To keep the definition consistent in all years, we used schools' previous low income score (2013-2014) as our measure of schools' 2014-2015 low-income status.

Prekindergarten year care settings. For our first two cohorts, when students applied to the Boston Public Schools, their parents answered a set of questions about the child's last childcare experience, including the name of the provider, and were asked to choose one from the following care types: Head Start, private preschool, public preschool, licensed family daycare, family daycare, and other/none. As noted in Weiland and Yoshikawa (2013), parents often disagreed about program type for the same program name so we cleaned and recoded these data extensively, confirming the type for each named program so that codes are consistent across children. We verified the program type via extensive web searches and through lists of programs and types obtained from the Massachusetts Department of Early Education and Care, the Boston Early Education Quality Improvement Project, and the National Association for the Education of Young Children. Information was often unavailable regarding whether a family daycare provider was licensed and parents frequently disagreed regarding the same provider's licensing status. Thus, we collapsed licensed family daycare and family daycare into one category in our analysis. Other/none almost always refers to relative care, such as parental care or care by an immediate relative.

We also used a second source – state administrative records. These records detailed whether children were enrolled in non-BPS public preschool programs in traditional public schools or charter schools in the year before kindergarten entry. If this second source contradicted the parent report, we set children's care setting equal to the second source as we viewed the second source as more reliable than the first source. Together, these two data sources

allowed us to estimate the counterfactual against which the Boston program is being compared in our analyses.

Note that for care setting analyses for the full sample (see Appendix F), we also drew on data from the time of their prekindergarten application (e.g., data from the winter/spring before the start of the four-year-old prekindergarten year) in cases where kindergarten information was missing. We did so because these data were missing at a higher rate than in our lottery-based sample. This approach represents a mix of reporting on counterfactuals in the year before prekindergarten and the year before kindergarten, which mixes a time when non-parental care tends to be less common/available (age 3) with a time when it is relatively more common/available (age 4; Chaudry et al., 2017). However, these analyses are nonetheless are informative for providing a lower bound on estimating the counterfactual, especially as we hypothesize that children who received non-parental care in the year before prekindergarten (age 3) were unlikely to revert to parental care in the year before kindergarten (age 4).

APPENDIX D: Attrition analysis

Outcome data availability. As shown in Table 1, outcome data were available for the majority of students in our lottery sample. Data overall were missing at relatively low rates (3-16%) and the differences in outcome missingness by treatment status were relatively small (3-5 percentage points more likely to be missing for the control group, across outcomes). Two of these four differences, for retention and special education placement, were statistically significant ($p<.0001$).

Table 2 presents the background characteristics across groups (lottery winners vs control group) for students with non-missing third grade standardized tests data. There were minor differences in race/ethnicity – lottery winners were more likely to Asian (3 percentage points, $p<.05$) and less likely to be Hispanic (5 percentage points, $p<.05$). The resulting F value from a joint F-test was not statistically significant ($p=.231$).

Overall, the characteristics of students with available outcome data and their distribution across the two groups remain similar to the original random assignment sample presented in Table 1 in the main text, which suggests that availability of outcome data likely poses a minimal threat to the internal validity of estimates based on this sample.

Table 1: Percent of missing data on key outcomes

	Lottery winners	Control group	Estimated difference	P-value
Retention	2.81	8.04	-5.23***	<.0001
Special Education	3.63	7.27	-3.63***	0.0008
ELA	12.81	15.55	-2.74	0.082
Mathematics	12.53	15.91	-3.38	0.072
N children	1,101	2,081		

Note: ELA=English Language Arts. Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent.

Table 2: Balance on baseline characteristics for students with test score data

	Lottery winners	Control group	Estimated difference	P-value
Race/ethnicity (%)				
Hispanic	35.24	40.13	-4.89*	0.016
Black	25.16	23.08	2.08	0.220
White	26.33	25.37	0.96	0.588
Asian	10.40	7.60	2.81*	0.014
Other	2.87	3.83	-0.96	0.300
Male (%)	49.36	47.40	1.96	0.411
Eligible for free/reduced lunch (%)	59.24	61.36	-2.12	
Age	4.52	4.54	-0.02	0.115
Country of origin USA (%)	95.65	95.29	0.36	0.699
Home language (%)				
English	25.58	25.48	0.11	0.951
Spanish	52.65	55.11	-2.46	0.245
Other	21.76	19.41	2.36	0.185
N children	942	1594		

Note: Nine students were missing free/reduced price lunch information; all other data was available for all students. Values for lottery winners are the simple means for each requisite group. Values for the difference between lottery winners and control group members are obtained from a regression of a given baseline characteristic on a series of indicator variables that identify each lottery plus an indicator variable that equals 1 for lottery winners and 0 for lottery losers. The coefficient on the lottery indicator equals the difference in the mean baseline characteristic between lottery winners and control group members. The value for control group members equals the corresponding value for lottery winners minus the estimated difference between lottery winners and control group members. A two-tailed t-test was applied to the estimated difference. Statistical significance levels are indicated as: * = 5 percent. An F-test was used to assess the statistical significance of the overall difference between lottery winners and control group members reflected by the full set of baseline characteristics in the table. The resulting F value was not statistically significant (p=.231).

APPENDIX E: Robustness checks

Table 1: Primary third grade outcome results (CACE) across different imputation for missing data decisions

	Estimated difference imputing covariates	P-value	Estimated difference imputing covariates and outcomes	P-value	Estimated difference without Imputation	P-value
Retained in grade (%)						
Retained in K	0.14	0.944	-0.04	0.990	0.20	0.929
Retained in 1st grade	4.28	0.122	4.27	0.112	4.66	0.124
Retained in 2nd grade	-0.09	0.967	-0.14	0.943	-0.07	0.975
Ever retained	4.20	0.269	4.22	0.250	4.64	0.263
Special Education Classification (%)						
SPED in K	-1.23	0.777	-0.68	0.873	-1.32	0.783
SPED in 1st grade	0.73	0.880	0.79	0.862	0.92	0.864
SPED in 2nd grade	5.18	0.326	7.32	0.141	5.35	0.354
SPED in 3rd grade	0.02	0.997	1.93	0.721	-0.20	0.975
Ever SPED	0.84	0.883	2.13	0.707	0.82	0.896
MCAS & PARCC						
English Language Arts	0.02	0.882	-0.19	0.679	0.02	0.870
Math	-0.18	0.210	0.05	0.142	-0.20	0.197

Note: There were no data missing on enrollment variables (and thus those outcomes were excluded from this robustness check). Other outcomes were missing data as follows: grade retention 11-15% across variables; special education 9-14% across variables; and test scores, 12%. All models included the full set of covariates. Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent

Table 1b: Intent-to-treat (ITT) findings for the MCAS/PARCC English Language Arts and Math outcomes without imputing covariates or outcomes

	Lottery winners included in the sensitivity analysis		Control group students included in the sensitivity analysis			
	Number of students	Standard deviation of the outcome	Number of students	Standard deviation of the outcome	ITT impact estimate	*p*-value
English Language Arts	942	0.91	1594	0.91	0.01	0.870
Math	942	0.99	1594	0.95	-0.05	0.197

Note: ITT impact models included the full set of covariates. This table was added to the Appendix in July 2022, for the purposes of additional review by the What Works Clearinghouse.

Table 2: CACE estimates of kindergarten and third grade school context differences between lottery winner compliers and control group compliers

	K only				Third grade only			
	Lottery winner compliers	Control group compliers	Estimated difference	P-value	Lottery winner compliers	Control group compliers	Estimated difference	P-value
Student Background Characteristic								
% Low-income	67.42	79.15	-11.73***	<.0001	64.58	66.36	-1.77	0.618
% ELL	26.13	27.55	-1.41	0.462	28.99	24.17	4.81*	0.040
% non-Eng. home lang.	37.93	37.98	-0.05	0.983	40.48	33.48	7.00**	0.006
% Disabilities	17.08	19.89	-2.81***	0.000	17.73	16.42	1.31	0.215
% African-American	27.88	37.16	-9.28***	<.0001	24.98	35.85	-10.87***	<.0001
% Asian	7.87	10.07	-2.20*	0.049	8.43	11.00	-2.57	0.052
% Hispanic	41.53	38.28	3.25	0.188	38.94	27.91	11.03**	0.000
% White	7.87	-0.90	8.77**	0.001	8.43	4.56	3.87	0.268
% Female	48.07	48.00	0.07	0.860	48.44	49.34	-0.91	0.054
Student Performance – % Proficient in 3rd grade								
ELA	43.40	31.99	11.41***	<.0001	45.84	46.11	-0.26	0.926
Math	49.08	35.04	14.04***	<.0001	56.17	56.94	-0.77	0.792
Teacher and school characteristics								
% Licensed Ts	98.21	95.31	2.89*	0.018	96.63	92.95	3.68*	0.010
Student-T ratio	13.56	13.16	0.40	0.085	14.04	13.63	0.41	0.159
% Exemplary Ts	-	-	-	-	16.48	12.43	4.04	0.083
% Proficient Ts	-	-	-	-	77.31	78.68	-1.38	0.598
% T retention	82.73	76.16	6.57**	0.000	81.29	80.29	1.00	0.535
Stability	89.16	80.48	8.68***	<.0001	90.02	87.34	2.68*	0.025
Avg class size	18.72	17.56	1.16	0.119	19.28	19.68	-0.40	0.522

Note: Using publicly available data from the Massachusetts Department of Elementary and Secondary Education, we report school-level data for the school in which a student was enrolled for the longest period of time in kindergarten and in third grade. For kindergarten, across variables, data were missing for 7-11% of students overall and 4% of treatment students were missing data compared with 15% of their control group counterparts. For third grade, across variables, data were missing for 8-12% of students overall and 5% of treatment students were missing data compared with 14% of their control group counterparts. For kindergarten, average class size was only available for cohorts 3 and 4. For third grade, percentage of teachers scoring proficient or exemplary on state ratings was available for cohorts 3 and 4 only. ELA=English Language Arts; Ts=teachers. Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent.

Table 3: Impacts of enrolling in Boston prekindergarten (CACE) on children's K-3 outcomes using prekindergarten enrollment thresholds of 1 and 150 days

	Estimated difference (1 day threshold)	P-value	Estimated difference (150 days threshold)	P-value
Enrolled in BPS (%)				
Kindergarten	23.81***	<.0001	26.54***	<.0001
First grade	34.41***	<.0001	38.34***	<.0001
Second grade	28.80***	<.0001	32.11***	<.0001
Third grade	25.57***	0.0002	28.50**	0.0002
Enrolled K-3	27.97***	<.0001	38.00***	<.0001
Ever enrolled	34.10***	<.0001	31.18***	<.0001
Retained in grade (%)				
Retained in K	0.14	0.9444	0.16	0.9444
Retained in 1st grade	4.28	0.1223	4.77	0.1223
Retained in 2nd grade	-0.09	0.9672	-0.10	0.9672
Ever retained	4.20	0.2685	4.68	0.2685
Special Education Classification (%)				
SPED in Kindergarten	-1.23	0.7770	-1.38	0.7770
SPED in first grade	0.73	0.8800	0.82	0.8800
SPED in second grade	5.18	0.3260	5.77	0.3260
SPED in third grade	0.02	0.9966	0.03	0.9966
Ever SPED	0.84	0.8833	0.94	0.8833
MCAS & PARCC				
English Language Arts	0.02	0.8822	0.02	0.8822
Math	-0.18	0.2104	-0.20	0.2104

Note: Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent.

Appendix F: Details on the full applicant sample and IPW work

Full sample balance on observables. In the full sample, there were statistically significant differences between enrollee and non-enrollee children in terms of race/ethnicity, gender, free-reduced lunch status, and home language (see Table 2). For example, enrollees were less likely to be Hispanic than non-enrollees (42% vs. 48%, respectively, $p<.001$) and more likely to be Asian (9% vs. 5%, $p<.001$). Enrollees were also more likely to be free-reduced-lunch eligible than non-enrollees (69% vs. 60%, respectively, $p<.001$) and less likely to speak English at home (48% vs. 54%, respectively, $p<.001$). Our joint F-test found evidence of non-equivalence of the two groups in their background characteristics ($p<.0001$).

Outcome data availability. In the full sample, as shown in Table 3, third grade outcome data were 15-19 percentage points more likely to be missing for non-enrollees than enrollees ($p<.0001$). Table 4 presents the background characteristics across groups (enrolled vs non-enrolled) for students with non-missing third grade standardized tests data. There were differences between enrollees and non-enrollees in terms of race/ethnicity (e.g., enrollees were seven percentage points less likely to be Hispanic, $p<.0001$), country of origin (e.g., enrollees were four percentage points more likely to report the U.S. as their country of origin, $p<.0001$), and home language (e.g., enrollees were five percentage points less likely to speak English as their home language, $p<.0001$). The F value from a joint F-test in the full sample was statistically significant ($p<.000$). Data availability and students characteristic differences between enrollee and non-enrollee sample underscore cautions in the main text about interpreting estimates from this sample as *associations* only, meant to give some information about generalizability of the lottery-based ITT/CACE estimates.

Covariate balance in IPW work. Table 5 displays differences in covariates in the ITT

replication and full samples before and after IPW reweighting. In both samples, re-weighting resulted in smaller differences between relevant groups.

Table 1: Baseline characteristics for children in the lottery sample only versus in the full sample

	Lottery sample	Full applicant sample	Difference
Race/ethnicity (%)			
Hispanic	39.21	43.87	-4.69
Black	21.48	28.43	-6.92
White	28.27	17.06	11.17
Asian	7.13	7.58	-0.47
Other	3.91	3.06	0.81
Male (%)	49.24	51.72	-2.46
Eligible for free/reduced lunch (%)	50.60	65.07	-14.50
Age	4.51	4.52	0.01
Country of origin USA (%)	95.05	93.33	1.75
Home language (%)			
English	56.68	50.24	6.48
Spanish	24.36	29.01	-4.64
Other	18.95	20.75	-1.75
N children	3,182	12,740	--

Note: In the lottery sample, there was a small amount of missing data on all baseline characteristics except age: 12 children (0.4%) were missing race/ethnicity and male information, 34 (1.1%) were missing male and free/reduced lunch information, 113 (4.2%) were missing country of origin information, and 5 (0.2%) were missing home language information. In the full applicant sample, there likewise was a small amount of missing data on all covariates except age: 33 children (0.3%) were missing race/ethnicity information, 185 (1.5%) were missing male and free/reduced lunch information, 514 (4.0%) were missing country of origin information, and 499 (3.9%) were missing home language information. Means in the table were computed using non-missing data.

Table 2: Balance on observables in the full applicant sample

	Enrollee	Non-enrollee	Estimated association	P-value
Race/ethnicity (%)				
Hispanic	41.85	47.85	-6.00***	0.000
Black	28.64	28.65	-0.01	0.986
White	17.26	15.34	1.92**	0.002
Asian	9.13	5.10	4.03***	0.000
Other	3.12	3.06	0.06	0.848
Male (%)	51.04	52.83	-1.79	0.054
Eligible for free/reduced lunch (%)	68.79	59.55	9.24***	0.000
Age	4.53	4.52	0.01	0.080
Country of origin USA (%)	94.70	90.49	4.21***	0.000
Home language (%)				
English	47.75	54.28	-6.53***	0.000
Spanish	30.35	27.03	3.32***	0.000
Other	21.90	18.69	3.21***	0.000
N children	8,115	4,625		

Note: There was a small amount of missing data on covariates, as explained in the Appendix F Table 1 note. Means in the table were computed using non-missing data. Values for enrollees are the simple means for each requisite group. The coefficient on the enrollment indicator equals the difference in the mean baseline characteristic between enrollees and non-enrollees. The value for non-enrollees members equals the corresponding value for non-enrollees minus the estimated difference between enrollees and non-enrollees. Values for the difference between enrollees and non-enrollees were obtained from a regression of a given baseline characteristics and a vector of neighborhood level fixed effects on the enrollment indicator. A two-tailed t-test was applied to the estimated difference. Statistical significance levels are indicated as: ***=.1 percent * = 5 percent. An F-test was used to assess the statistical significance of the overall difference between enrollees and non-enrollees reflected by the full set of baseline characteristics in the table. For the enrollee vs. non-enrollee comparison, the resulting F value was statistically significant (p<.0001).

Table 3: Percent of missing data on key outcomes in the full applicant sample

	Enrollee	Non-enrollee	Estimated association	P-value
Retention	4.18	19.08	-14.94***	0.000
Special Education	1.36	15.76	-14.36***	0.000
ELA	13.56	32.16	-18.64***	0.000
Mathematics	13.46	31.96	-18.50***	0.000
N children	8,115	4,625		

Note: ELA=English Language Arts. Statistical significance levels are indicated as: ***=.1 percent ** = 1 percent; * = 5 percent.

Table 4: Balance on baseline characteristics for students with test score data in the full applicant sample

	Enrollee	Non-enrollee	Estimated association	P-value
Race/ethnicity (%)				
Hispanic	42.41	49.44	-7.03***	0.000
Black	28.44	28.65	-0.21	0.814
White	16.61	13.46	3.15***	0.000
Asian	9.37	5.46	3.91***	0.000
Other	3.16	2.99	0.17	0.646
Male (%)	50.89	51.42	-0.53	0.626
Eligible for free/reduced lunch (%)	70.85	70.94	-0.08	0.929
Age	4.53	4.52	0.01	0.106
Country of origin USA (%)	95.11	91.22	3.89***	0.000
Home language (%)				
English	47.83	53.13	-5.30***	0.000
Spanish	30.67	28.03	2.64**	0.005
Other	21.51	18.85	2.66**	0.002
N children	6,989	3,126		

Note: For children with non-missing math and ELA test scores, there was no missing data on race/ethnicity, gender, or age. There was a small amount of missing data on other variables: 95 (0.9%) were missing free/reduced lunch information, 1 (0.0%) was missing country of origin information, and 225 (2.3%) were missing home language information. Values for enrollees are the simple means for each requisite group. Values for the difference between enrollees and non-enrollees were obtained from a regression of a given baseline characteristic and a vector of neighborhood level fixed effects on the enrollment indicator. The coefficient on the enrollment indicator equals the difference in the mean baseline characteristic between enrollees and non-enrollees. The value for non-enrollees equals the corresponding value for enrollees minus the estimated difference between enrollees and non-enrollees, respectively. A two-tailed t-test was applied to the estimated difference. Statistical significance levels are indicated as: ***=.1 percent. An F-test was used to assess the statistical significance of the overall difference between enrollees and non-enrollees reflected by the full set of baseline characteristics in the table. The resulting F value was statistically significant (p<.000).

Table 5: Example covariate differences between treatment and control members in the lottery replication sample and the full sample before and after IPW

	ITT Replication Sample		Full Sample	
	Unweighted	Reweighted	Unweighted	Reweighted
Race/ethnicity (%)				
Hispanic	-17.52	1.43	-15.91	0.36
Black	2.06	-0.46	1.33	-0.29
White	13.47	-0.78	7.66	-0.40
Asian	9.07	-1.22	16.23	0.96
Other	-2.12	0.73	2.04	-0.92
Male (%)	-0.13	0.70	-3.79	-0.19
Eligible for free/reduced lunch (%)	-12.18	0.68	5.85	0.16
Age	0.02	0.00	0.03	0.00
Country of origin USA (%)	-4.65	0.39	2.72	0.31
Home language (%)				
English	3.44	0.60	-13.19	-0.17
Spanish	-10.45	0.20	6.87	-0.42
Other	7.11	-0.97	8.65	0.69
N children				
Treatment	1,667	4,634	7,829	5,799
Control	7,589	4,622	3,778	5,807

Notes: Results shown for the "enrolled in BPS in kindergarten" outcome. Neighborhood fixed effects corresponding to the neighborhood a child lived in during his or her preschool-eligible year (as proxied by a fixed effect for the public elementary closest to the child's home) were also included in IPW models but are not shown here for parsimony. Full results and results for other outcomes available upon request.

Table 6: Children's care settings (cohorts 1 and 2) in the prekindergarten year for the full sample

	Enrolled	Did not enroll
Age 4 Care Setting		
Any center-based preschool	100.00	76.10
BPS	100.00	0.00
Non-BPS center-based preschool	0.00	76.10
Private	*0.00*	*37.05*
Head Start	*0.00*	*25.81*
Public	*0.00*	*6.33*
Charter	*0.00*	*6.91*
Family daycare	0.00	5.58
At home	0.00	18.32
Total	100.00	100.00
Age 4 Care Setting or Age 3 setting (if age 4 care setting data was missing)		
Any center-based preschool	100.00	69.66
BPS	100.00	0.00
Non-BPS center-based preschool	0.00	69.66
Private	*0.00*	*39.22*
Head Start	*0.00*	*20.53*
Public	*0.00*	*5.44*
Charter	*0.00*	*4.47*
Family daycare	0.00	7.65
At home	0.00	22.68
Total	100.00	100.00

Note: Age 4 care setting types were reported by parents at the time of application to Boston kindergarten (e.g., the winter, spring, or summer preceding kindergarten fall), were pulled from Boston prekindergarten enrollment records, or were pulled from age 4 state administrative records on traditional public school or charter school enrollment. Age 3 care setting types were reported by parents at the time of application to Boston prekindergarten (e.g., the winter, spring, or summer preceding prekindergarten fall). Mean percentages are shown in the table. In all, 11% of full sample students were missing any care setting data (age 3 and 4) and 42% were missing age 4 care setting data.