EdWorkingPaper No. 20-321

Higher-Quality Elementary Schools Sustain the Prekindergarten Boost: Evidence from an Exploration of Variation in the Boston Prekindergarten Program's Impacts

Rebecca Unterman
MDRC

Christina Weiland
University of Michigan

While there is a consensus that attending preschool better prepares children for kindergarten, evidence on the factors that sustain the preschool boost into the early elementary years is still emerging. To add to this literature, we use lottery data from applicants to oversubscribed schools in Boston Public Schools (BPS) prekindergarten program to estimate variation in the effects of the program across school sites through the end of third grade. Student outcomes include children's kindergarten-through-second-grade retention, kindergarten-through-third-grade special education placement, and third-grade state English Language Arts and math test scores. We find statistically significant variation in effects in all student outcomes and we predict this variation with multiple proxies for early elementary school quality. We find that the academic proficiency of third-graders within the schools for which prekindergarten children competed is most strongly associated with prekindergarten program effects. Prekindergarten gains persisted if students applied to and won a seat in a higher-quality elementary school. Our findings appear to be driven by the schools themselves and not by student selection in higher-scoring schools, nor by the counterfactual. These findings imply that policymakers and practitioners interested in sustained gains may need to also invest in improving the quality of children's K-3 experience.

VERSION: November 2020

Suggested citation: Unterman, Rebecca, and Christina Weiland. (2020). Higher-Quality Elementary Schools Sustain the Prekindergarten Boost: Evidence from an Exploration of Variation in the Boston Prekindergarten Program's Impacts. (EdWorkingPaper: 20-321). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/hkv0-wj29

Higher-Quality Elementary Schools Sustain the Prekindergarten Boost:

Evidence from an Exploration of Variation in the Boston Prekindergarten Program's Impacts

Rebecca Unterman[a]

Christina Weiland[b]

[a]MDRC [b]University of Michigan

Date: November 5, 2020

Authors' note:

This study is funded by the Institute of Education Sciences RFA R305A140059.
Correspondence concerning this article should be addressed to Rebecca Unterman, MDRC
email: Rebecca.Unterman@mdrc.org

Thanks to the Boston Public Schools, 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, the UM research team (Anna Shapiro, Shana Rochester, and Eleanor Martin), Sara Staszak, and the Wellesley Centers for Women. Special thanks also to Hirokazu Yoshikawa, Howard Bloom, Richard Murnane, David Deming, Catherine Snow, Caroline Ebanks, and Gina Biancarosa.

Abstract

While there is a consensus that attending preschool better prepares children for kindergarten, evidence on the factors that sustain the preschool boost into the early elementary years is still emerging. To add to this literature, we use lottery data from applicants to oversubscribed schools in Boston Public Schools (BPS) prekindergarten program to estimate variation in the effects of the program across school sites through the end of third grade. Student outcomes include children's kindergarten-through-second-grade retention, kindergarten-through-third-grade special education placement, and third-grade state English Language Arts and math test scores. We find statistically significant variation in effects in all student outcomes and we predict this variation with multiple proxies for early elementary school quality. We find that the academic proficiency of third-graders within the schools for which prekindergarten children competed is most strongly associated with prekindergarten program effects. Prekindergarten gains persisted if students applied to and won a seat in a higher-quality elementary school. Our findings appear to be driven by *the schools themselves* and not by student selection in higher-scoring schools, nor by the counterfactual. These findings imply that policymakers and practitioners interested in sustained gains may need to also invest in improving the quality of children's K-3 experience.

Higher-Quality Elementary Schools Sustain the Prekindergarten Boost: Evidence from an Exploration of Variation in the Boston Prekindergarten Program's Impacts[1]

The evidence is clear that a wide range of preschool programs, operated across diverse settings and models, improve children's cognitive and socio-emotional readiness for kindergarten (Duncan & Magnuson, 2013; Phillips et al., 2017; Yoshikawa et al., 2013). Motivated in part by such evidence, 44 states now offer state-funded preschool programs. In all, 47% of U.S. 4 year olds were enrolled in public preschool in 2019 (Friedman-Krauss et al., 2019).

However, the evidence is more mixed regarding how long the preschool boost lasts. In studies conducted decades ago, the language, literacy, and mathematics test scores of preschool participants and nonparticipants tended to converge in the medium term (i.e., during elementary school). But in adulthood, preschool participants tend to outperform nonparticipants on a variety of behavioral, health, and educational outcomes. Children in today's large-scale preschool programs have not yet reached adulthood, but so far, the medium-term evidence from (Friedman-Krauss et al., 2019) these programs largely mirrors the medium-term pattern of the older studies (Phillips et al., 2017; Yoshikawa, Weiland, & Brooks-Gunn, 2016).

The mechanisms behind the medium-term convergence pattern are not well understood. One of the leading hypotheses is called the "sustaining environments" hypothesis, which posits that the quality (broadly defined) of children's educational settings after preschool is critical in sustaining the preschool boost (Bailey, Duncan, Odgers, & Yu, 2017). Specifically, this hypothesis holds that high-quality environments will build on preschool attenders' strong foundational skills, thereby sustaining the preschool advantage. Low-quality environments will do the opposite, essentially keeping higher-

[1] This paper is an update of a prior working paper. The original working paper is available here: https://www.mdrc.org/publication/quantifying-and-predicting-variation-medium-term-effects-oversubscribed-prekindergarten

skilled preschool attenders in place while nonattenders catch up. Notably, inklings of evidence supporting this hypothesis are present in the older literature. Garces, Thomas, and Currie (2002), for example, argued that the Head Start boost faded more quickly for Black children than White children because after Head Start, the former were likely to attend schools of lower quality, as measured by school-level test scores. In more recent programs, as we detail further in the next section, the evidence supporting this hypothesis has been mixed and is still emerging (Ansari & Pianta, 2018; Bassok et al., 2016; Bierman et al., 2014; Clements, Sarama, Wolfe, & Spitler, 2013; Jenkins et al., 2017; Johnson & Jackson, 2019; Kitchens, Gormley, & Anderson, 2020; Swain, Springer, and Hofer, 2015; Zhai, Raver, & Jones, 2012).

Solving the convergence puzzle is one of the chief challenges facing the field of early childhood education. Policymakers and practitioners' ability to create conditions in which preschool benefits can last — particularly in large-scale programs — is currently limited by the lack of empirical evidence. In the present paper, we help advance the science of early childhood education by exploring variation in the medium-term effects of prekindergarten within a unique sample — children who participated in oversubscribed lotteries for the Boston Public Schools (BPS) prekindergarten program.[2] BPS has unusually high instructional quality in prekindergarten compared with other large-scale U.S. programs (Chaudry, Morrissey, Weiland, & Yoshikawa, 2017; Weiland, Ulvestad, Sachs, & Yoshikawa, 2013), but in our study years, its kindergarten-through-third-grade (K-3) quality was weaker than its prekindergarten program (Weiland et al., 2019).

Recently, for the subset of program applicants who participated in oversubscribed lotteries for the program, we found no statistically significant differences between prekindergarten enrollees

[2] The Boston Public Schools (BPS) Department of Early Childhood refers to its public preschool program for four-year olds as "prekindergarten." When describing the BPS program in this paper, we similarly use the term "prekindergarten." However, when discussing the broader literature on early childhood care and education, we use the term preschool.

and non-enrollees on K-2 retention, K-3 special education placement, and third grade English Language Arts and math standardized test scores (Weiland et al., 2019). In the present study, we build on this work and test whether the impact of the Boston prekindergarten program on students' medium-term outcomes differs across school settings, and if so, whether prekindergarten programs located within higher-quality elementary schools better sustain effects than those located within lower-quality elementary schools. We find substantial variation in all examined student outcomes across sites, from substantially negative to substantially positive. One specific aspect of early elementary schools seems to matter more than others in sustaining the prekindergarten boost – the average test score proficiency levels of schools at the time of students' application for prekindergarten. Other proxies for school quality such as the percent of free/reduced lunch students in the school and demand for prekindergarten were not consistent predictors of a lasting boost. As we show, our findings appear to be driven by children's schooling experiences and not by student selection in higher-scoring schools, nor by the counterfactual. Our findings lend additional weight to the sustaining environments hypothesis and suggest that stakeholders seeking a lasting preschool boost may need to invest in improving K-3 quality as well.

The Mechanisms of Convergence

To date, there has been little empirical work on the mechanisms explaining the medium-term convergence phenomenon. The work that does exist has largely focused on the quality (variously defined) of educational settings in kindergarten and beyond, often referred to as the "sustaining environments" hypothesis (Bailey, Duncan, Odgers, & Yu, 2017). Intuitively, this focus makes sense. For children who attended preschool, an elementary school with high levels of instructional quality may build on their progress, amplifying the program's impacts. An elementary school with low levels of instructional quality may fail to individualize instruction and meet all students' needs, effectively stifling preschool students' growth and essentially holding them in place. For students who did not attend preschool, the school experience may have the opposite effect — a high-quality elementary school may help these students "catch up," while an elementary school with low levels of instructional quality may pull them down further, increasing the longer-term preschool impact.

The weight of the existing empirical evidence overall does lend support to the sustaining environments hypothesis. Seven recent studies found that preschool effects were more likely to be sustained if students subsequently experienced higher-quality early elementary school environments, as measured by spending (Johnson & Jackson, 2019); by school-level third-grade standardized test scores (Zhai, Raver, & Jones, 2012); by a multidimensional measure of school resources, organization, and social processes (Ansari & Pianta, 2018); by state ratings of first-grade teacher overall effectiveness (Swain, Springer, & Hofer, 2015); by alignment between preschool and kindergarten curriculum (Clements, Sarama, Wolfe, & Spitler, 2013; Mattera, Jacob, & Morris, 2018); or by the number of years from K-3 children had access to higher-quality K-3 experiences, as defined by teacher quality ratings and school's contribution to improving child test scores (Pearman et al., 2019). In the older literature, in their study of Head Start, Currie and Thomas (1998) found that benefits lasted to third grade only for White children. In a subsequent analysis, they posited that the explanation may have been that Black children attended lower-quality elementary schools (as defined by student test scores) than White children.

In addition, peers may be another potential component of a higher quality, sustaining early elementary school environment. If a larger percentage of children's peers attended preschool and thus enter kindergarten with stronger skills, this could also potentially lead to a sustained boost (Bailey, Duncan, Odgers, & Yu, 2017). Correlational work in preschool does support the existence of peer effects in these years (for example, Henry & Rickman, 2007; Justice, Petscher, Schatschneider,

& Mashburn, 2011; Weiland & Yoshikawa, 2014). Further, one quasi-experimental study found positive spillover effects of having a higher percentage of peers in kindergarten who attended preschool on children's literacy and math gains (Neidell & Waldfogel, 2010). The mechanism in this case could be direct peer effects, given that children do learn directly from one another, especially in in preschool classrooms which tend to emphasize frequent interactions between children. Or it could be due to indirect teacher effects. Having more preschool attenders with strong skills could spark teachers to increase rigor, to increase expectations for children, and/or to spend more time on higher performers (if there are fewer struggling students in the class who require more teacher support; Sameroff, 2009).

However, in contrast to these findings that lend support to the sustaining environments hypothesis, another recent study found persistence of effects of a preschool intervention only for children enrolled in kindergarten classrooms with a relatively low-quality emotional climate (Bierman et al., 2014). Another found more persistence for children in kindergarten and first-grade classrooms with lower teacher-reported levels of academic instruction (Magnuson, Ruhm, & Waldfogel, 2007). Two others that examined a host of kindergarten classroom features like class size, use of transition practices, and teacher-reported measures of instructional practices found largely null results (Bassok et al., 2016; Jenkins et al., 2017). Accordingly, to date, the evidence does not universally point in favor of the sustaining environments hypothesis, nor does it offer clarity on which specific post-preschool elements matter most.

Notably, while we focus on the sustaining environments hypothesis, there are also several other hypotheses about medium-term convergence. For example, Bailey, Duncan, Odgers, and Yu (2017) identified a second "foot-in-the-door" pathway by which they hypothesize preschool effects may or may not be sustained. 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 harm. An example would be clearing a bar into gifted education (access to a benefit) or away from grade retention (avoiding a potential harm). In support of this hypothesis, 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 (Yoshikawa, Weiland, & Brooks-Gunn, 2016). A recent meta-analysis of 18 studies from 1962 to 2003 found average reductions of 0.29 SD or 10.1 percentage points for grade retention and 0.40 SD or 12.5 percentage points for special education placement avoidance (McCoy et al., 2016). However, in our recent Boston study, we found no effects on these outcomes for the lottery sample, though there were such effects in our less rigorous propensity score analysis of the full prekindergarten applicant sample (Weiland et al., 2019).

Some (McCormick, Hsueh, Weiland, & Bangser, 2017)have hypothesized that another key to convergence could be which skills are emphasized and measured in the period from prekindergarten through third grade (McCormick, Hsueh, Weiland, & Bangser, 2017). The boost from a prekindergarten program that focuses on constrained literacy skills — for example, the discrete set of basic skills that almost all children master by third grade, such as letter sounds and basic spelling (Paris, 2005; Snow & Matthews, 2016) — is likely to be less enduring than the boost from a program that focuses on students' deeper unconstrained skills, meaning more broadly and skills like world knowledge, vocabulary, and conceptual thinking. Both kinds of skills are important for children's early literacy learning, but prekindergarten-through-third-grade (P-3) assessments in classroom settings tend to privilege measuring students' constrained skills, thereby leading teachers to neglect the unconstrained skills critical to students' longer-term success (Snow & Matthews, 2016). To date, the role of unconstrained versus constrained literacy skills teaching is largely unexplored in the preschool persistence literature. Because the Boston prekindergarten program focused on building both kinds of skills and because we can decompose children's third grade English Language Arts tests by skill type, we are able to provide some of the first evidence on this possibility as well.

The Boston Public Schools Prekindergarten and Kindergarten-Through-Third-Grade Programs

The Boston Public Schools prekindergarten program is a relatively large-scale program that was based entirely in the public schools in our study years, paid teachers on the same scale as kindergarten-through-twelfth-grade (K-12) teachers, subjected teachers to the same educational requirements as K-12 teachers (masters degree within 5 years), and was open to any child in the city, regardless of income. In our study years, the program implemented the language and literacy-focused curriculum Opening the World of Learning (OWL), which targets children's early language and literacy skills and includes a social skills component embedded in each unit, in which teachers discuss socio-emotional issues with children and integrate emotion-related vocabulary words. It also implemented Building Blocks, an early childhood mathematics curriculum that covers both numeracy and geometry and has a heavy focus on verbal mathematical reasoning. Both curricula have shown positive effects on children's outcomes in other studies (Ashe et al., 2009; Clements & Sarama, 2007; Clements et al., 2011), though the evidence base for Building Blocks is stronger than that for OWL.

In 2007 to 2009, covering two of the four cohorts in our sample, 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 (see Weiland & Yoshikawa, 2013, for additional details). In 2009 to 2011 (covering the other two cohorts in our sample), as a result of 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.

Taken together, Boston's structural and programmatic choices make it fairly unusual among public programs nationally, which tend not to require masters' degrees; usually do not pay prekindergarten teachers on the same scale as K-12 teachers; tend to target slots to children from low-income families or with other risk factors; do not require a proven, consistent curriculum; and do not employ coaching (Barnett et al., 2017).

Empirically, the Boston Prekindergarten program has been shown to have the highest average instructional quality of a large-scale program to date on the CLASS observational quality measure (Chaudry et al., 2017). It also showed strong effects on children's language, literacy, mathematics, and executive function skills at kindergarten entry 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). Effects were particularly pronounced for Hispanic students, low-income students, and children with special needs (Weiland, 2016).

In our study years, Boston K-3 offerings were not as strong as in prekindergarten (Weiland et al., 2019). District kindergarten-through-third-grade (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 with the prekindergarten curricula used in the district (Agodini et al., 2010; Gatti & Petrochenkov, 2010; Ladnier-Hicks, McNeese, & Johnson, 2010; What Works Clearinghouse, 2013; Wilkerson, Shannon, & Herman, 2006), nor were the supports for implementing them as systematic or rich as for the prekindergarten program. Data on a widely used measure of the quality of teacher-child interactions (Classroom Assessment Scoring System (CLASS); Pianta, La Paro, & Hamre, 2008) 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 show that prekindergarten classroom quality was higher on average than K-3 quality (Weiland et al., 2019).[3]

In recent work in Boston, we followed four cohorts of students who competed in naturally occurring lotteries for a prekindergarten slot (about 25 percent of all appliers). We found no differences in K-2 retention, K-3 special education placement, or third grade standardized ELA and math test scores for enrollees versus non-enrollees (Weiland et al., 2019). We build on this evidence in the present study by estimating variation in impacts across sites and exploring whether, consistent with the sustaining environments hypothesis, attending a higher-quality elementary school helps to sustain the boost. Understanding if there is variation in effects across program sites and the predictors of this variation may provide useful information to policy-makers and practitioners seeking to learn from the implementation of the program. For example, if some sites experienced sustained positive effects, their characteristics can provide insight into the mechanisms through which BPS prekindergarten impacts students over time, as well as help other districts decide if they may experience similar effects.

Present Study

Using data from four cohorts of students whose families listed oversubscribed Boston prekindergarten sites as their first choice, we build on our previous work and address the following research questions:

1. Does the impact of the BPS prekindergarten program on students' grade retention, special education identification, and third-grade state standardized mathematics and English language arts (ELA) test scores differ across program sites?

[3] For example, prekindergarten classrooms scored 5.6 on the CLASS emotional support and 4.3 on instructional support subscales, compared with 5.1 and 4.1, respectively, for K-3 classrooms. The standardized differences between prekindergarten and K-3 classroom quality were 0.2 (organizational support), 0.5 (instructional support), and 0.9 (emotional support).

2. Do BPS prekindergarten programs located within higher-quality elementary schools produce different impacts from those located within lower-quality elementary schools? Does the answer depend on how elementary school quality is measured? Can any differences in impacts between these two groups be explained by differences in the K-3 settings children experienced, i.e. the treatment contrast?

Method

Sample

Our sample comes from the population of students who applied to the Boston Prekindergarten program for four year olds between 2007 and 2010. As shown in Figure 1, in all, 12,740 families applied to the program in our focal years. Nearly 10,000 of these families applied through the four rounds of the district's school choice lottery, in the spring before their children were age-eligible for the program. This is what we call the "standard process," from which we identified naturally occurring lotteries for students' first-choice school (labeled "first-choice lottery" in Figure 1) involving 3,182 students, or 25 percent of all appliers and 32 percent of those who applied through the standard process.

School Assignment Process Details

Under BPS's choice plan, in the winter and spring, families could apply for up to 10 schools they wanted their child to attend for prekindergarten the following fall. Families were assigned different priorities to different schools based on criteria set by the district (e.g., sibling, walk zone, etc). When there was more demand than supply for a given school, the assignment algorithm used family choice lists, priorities, and a random number to randomly assign some students (and not others) to a given choice. Weiland and colleagues (2019) provide further details on this process.

Our experimental sample students are drawn primarily from the first round of the assignment process (as are most students who attend BPS prekindergarten) and are distributed relatively equally across all four years of the study sample. As discussed in detail in Weiland and colleagues (2019) and presented in Appendix B, a joint F-test used to assess the statistical significance of the overall baseline (or pre-existing) difference between the lottery winners and control group members in the experimental sample could not reject the null hypothesis that there was no difference between the two groups ($p = 0.200$). The internal validity of the sample was maintained throughout the follow-up period (e.g., relatively low overall attrition and little differential attrition by treatment status; Appendix B).

In Table 1, we present the baseline characteristics of the experimental sample lottery alongside those of all students who applied to the Boston prekindergarten program during the study period. While the two samples appeared to be similar in age, country of origin, and gender, there were some noticeable differences between them. For example, white students accounted for 28 percent of the experimental sample versus 17 percent of all BPS prekindergarten appliers and about 51 percent of the experimental sample qualified for free or reduced-price lunch, while 65 percent of all BPS appliers did. Further, while the percentage of BPS elementary schools with prekindergarten represented by lotteries ranged from 67 to 83 across years, the lottery sample students are highly concentrated in a subset of the BPS prekindergarten sites (e.g., 75% of lottery sample students competed for about a quarter of eligible district schools; Weiland et al., 2019).

BPS prekindergarten programs are located within elementary school buildings; when lottery participants win a seat in a prekindergarten program, they are automatically enrolled in for the same school for kindergarten.[4] For the purposes of this paper, students who competed for an oversubscribed prekindergarten program actually competed for a school experience encompassing prekindergarten through third grade. Said differently, we used these lotteries to estimate the causal effect of winning the opportunity to attend a specific P-3 program and to explore whether features of this program (measured at baseline, before student enrollment in BPS) predicted differences in students' medium-term outcomes.

Unfortunately, because measures of BPS *prekindergarten* quality are not available for the full study sample during this period, we cannot disentangle the relationship between prekindergarten quality and elementary school quality in this paper. For example, when we find higher lottery-based impacts for programs located in higher-quality elementary school sites, we cannot know whether this is the case because these prekindergarten sites produced higher impacts or whether the students' K-3 experience did a better job of sustaining them. We return to this design limitation in the discussion.

Measures

Student outcomes. We examined variation in BPS prekindergarten effects on four primary student outcomes: 1) retention in grades K-2; 2) placement in special education in grades K-3 (defined as having an IEP); 3) third-grade mathematics scores on the state standardized test; and 4) third-grade English Language Arts (ELA) scores on the state standardized test. We standardized children's test scores on the mean and standard deviation of all third graders within BPS taking the given exam in that year so that test score data can be interpreted as a given group's performance compared with that of the average BPS third grader. See Appendix A for more details.

[4] As reported in Weiland et al. (2019), 91 percent of the lottery winners who enrolled in BPS prekindergarten enrolled in BPS for kindergarten. Of these students, roughly 90 percent stayed enrolled in the school they attended for prekindergarten.

We also decomposed children's performance on ELA test score items into "more constrained" and "unconstrained" subscores. This measurement decision follows a consensus among literacy experts that there is continuum in the degree to which reading skills are constrained (i.e., skills for which there is a ceiling most students reach in early elementary school, like alphabet knowledge) versus unconstrained (i.e., skills for which there is always room for improvement, like vocabulary and comprehension; Snow & Matthews, 2016). See Appendix A for more details.

Covariates. From administrative records, we created variables that described whether a student was Asian, Black, Latino, White, or other; whether the students' home language was English only, English and Spanish, or English and another language; students' age as of September 1 in the year they were applying to prekindergarten; whether the student was eligible for free-or-reduced-price lunch; whether the student was male, and whether the student's country of birth was the United States.

Predictors of impact variation. We drew on administrative records to create a parsimonious set of site characteristics that we used to predict variation in program impacts. For each students, site characteristics are from *the year they competed in a lottery for the school*; they do not overlap with the years students attended the school and are not affected by the students in our lottery sample. We list our five types of site characteristics below, succinctly summarize why we included each, and include 25th, 50th, and 75th percentile values in parenthesis:

- Demand, defined as the number of applicants per available seat for each prekindergarten program (percentile values: 4.2, 6.23, and 8.8). Parent preferences are dynamic, complex, and heterogeneous by family characteristics. But following both theory and some empirical evidence, higher demand might signal a school that holistically is better poised to offer a higher-quality P-3 experience and build on

(Hastings, Kane, & Staiger, 2005) (Hastings, Kane, & Staiger, 2006) children's prekindergarten gains (Hastings, Kane & Staiger, 2005;(Hastings, Van Weelden, & Weinstein, 2007); (Hoxby, 2007));

- Third-grade test score proficiency, calculated as an average of the percentages of students scoring proficient on third-grade math and ELA exams (percentile values: 30.5, 44.5, 58.5). A study in Chicago found that the effects of a preschool enhancement intervention lasted into elementary school only for those who subsequently entered higher-quality elementary schools (Zhai et al., 2012). Third test scores before students enroll in the school serve a proxy for the school's efficacy from P-3 in educating students, as well as for the types of students attracted to the school.
- Student growth percentile (SGP) in math in third grade, a state-created metric of yearly changes in a students' third-grade test scores relative to the yearly changes of students with similar characteristics.[5] This measure was available for cohorts two through four (percentile values: 42.0, 50.0, 58.0). We include this measure because it serves as school value-added proxy – i.e., it at least partially removes the role of student selection into a given school.
- Proportion of students from low-income families (percentile values: 61.0, 75.0, and 80.2). We include this measure as a proxy for the resources available to a school - schools with lower proportions of students from lower-income families might have additional resources to support the quality of students' P-3 experiences and if so, may better sustain the prekindergarten boost.

[5] Math and ELA SGP scores were highly correlated for the schools in our sample (r = 0.91, p-value = 0.000). We focus on math for parsimony. Results based on schools' ELA SGP measures were similar and are available upon request.

- School climate, constructed from surveys of BPS students (grades 3-11) and teachers (K-12) and scored on a four-point Likert scale (1 = strongly disagree to 4 = strongly agree). Following Rochester, Weiland, Unterman, and McCormick (2019), we used three subscale scores: positive emotional climate (percentile values: 2.80, 2.82, 3.00), student engagement/teacher effectiveness (percentile values: 2.72, 3.21, 3.34), and principal effectiveness subscale scores (percentile values: 3.24, 3.56, 3.63).[6] This measure was available for cohorts two through four. We include these measures because research has consistently linked a more positive school climate to a host of benefits for students and teachers, including higher student academic achievement (e.g., Cornell, Shukla, & Konold, 2016; Kraft, Marinell, & Yee, 2015; Sherblom, Marshall, & Sherblom, 2006);
- Percentage of kindergarten peers who attended BPS prekindergarten in the prior year (percentile values: 28.87, 50.77, 73.33). As reviewed earlier, peer effects have been documented in the early years (Henry & Rickman, 2007; Justice et al., 2011; Neidell & Waldfogel, 2010; Weiland & Yoshikawa, 2014). Further, kindergarten teachers might respond to having a larger percentage of children with higher skills by increasing rigor. More advanced content in kindergarten has been shown to predict children's gains (Engel, Claessens, Watts, & Farkas, 2016).

Correlations between our predictors of impact variation ranged from -0.04 to -0.80 (see Appendix C Table C.1). The weakest association (*r* = -0.04) is between principal effectiveness and demand, while the strongest association is a negative one, between the proportion of low-income students in a school and the school's third-grade academic proficiency scores (*r* = -0.80).

[6] Correlations between school climate dimensions ranged from 0.16 to 0.68 (results available upon request).

Treatment contrast measures (measures of students' school experiences). Using publicly available school-level data from Massachusetts, we created measures of students' K-3 schooling experiences, for the school the student was reported *being enrolled in for the longest time during his or her kindergarten, first-, second-, and third-grade school years*. We averaged each yearly value separately for each measure to capture students' K-3 exposure to a given school characteristic and ultimately, to estimate the treatment-control contrast. In other words, we used these measures to understand whether differences in K-3 outcomes between students who attended higher-quality versus lower-quality elementary schools might be have been due to differences in students' school experiences.

Our measures included several school-level peer characteristics: the proportion of English language learners; the proportion of students with disabilities; the proportion of students who qualify for free or reduced-price lunch; racial make-up; and the percentage of students who scored proficient or higher on third-grade state standardized ELA and math tests. We also included school-level measures of the proportion of licensed teachers within the school, teacher-to-student ratio, percent of teachers retained in the school from the prior year, average class size, percentage of teachers rated as proficient or exemplary by administrators under the state's teacher evaluation system (available for cohorts 3 and 4 only; Massachusetts Department of Elementary & Secondary Education, 2017), and the percentage of students who remain in the school throughout the school year (stability rate). Finally, using BPS administrative records of Boston prekindergarten attendance and the school and classroom a student was enrolled in for the longest time during his or her K-3 school years, we calculated the percentage of each student's classmates who had attended the Boston prekindergarten program.

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 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 daycare, 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 D.

Data Analytic Plan

We relied on the naturally occurring lotteries created by the BPS school assignment process to identify the causal effect of the BPS prekindergarten program on students who are offered the opportunity to enroll. Many researchers have utilized this experimental, lottery-based approach (Abdulkadiroğlu et al., 2011; Bloom & Unterman, 2014; Dobbie & Fryer, 2011). Within our research design, a set of students randomly "win" the opportunity to attend the BPS prekindergarten program, and a set of students randomly "lose" the opportunity to attend the BPS prekindergarten program. Those who "win" make up our treatment group, and those who "lose" make up our control group. Like in a randomized controlled trial, students in the treatment and control groups are, in expectation, equivalent in all measurable and unmeasurable characteristics. As students are followed over time, the only difference between the two groups is the causal effect of being offered the opportunity to attend the BPS program. We estimated, for each lottery, differences in mean outcomes for winners and control group members, and averaged the results across lotteries.

Estimating the distribution of Intent-to-Treat effects across sites. As a first step in our analysis of impact variation and to answer Research Question 1, we quantified and illustrated the distribution of intent-to-treat (ITT) effects across sites using the framework created by Bloom, Raudenbush, Weiss, and Porter (2017) and applied by Weiss et al. (2017). As they suggest, we assumed that our study sites are a sample drawn from a "super population" of prekindergarten sites, and our goal is to generalize to the larger population from which we have drawn. We estimated key statistics for these distributions using a two-level hierarchical linear model and illustrated the distributions using site-level constrained empirical Bayes impact estimates, which, as shown in Bloom and colleagues (2017), constrain the cross-site variance to match that estimated by the model below. This is preferable to an empirical Bayes model, which may slightly underestimate cross-site variation.

Some students who lose a lottery win a subsequent lottery and enroll in the program; the average enrollment rate difference between treatment and control group members (that is, compliance rate difference) was 0.29. The lottery-induced BPS prekindergarten enrollment rate differences did not vary statistically significantly across sites ($\tau = 0.04$, $p = 0.213$). This apparently constant compliance rate difference permitted us to analyze variation in site-level effects using intent-to-treat impact estimates rather than complier average causal effects. Said differently, since the compliance rate difference does not differ across sites, we infer that differences in intent-to-treat impact estimates across sites are not driven by differences in compliance rates across sites. At various points below, to approximate the *effect of enrolling* in the BPS prekindergarten program, we computed a Wald estimate by dividing the estimated treatment effect by the average compliance rate difference.

As described in Weiss and colleagues (2017) and using their notation and definitions below, we estimated the distribution of the treatment effects focusing on the cross-site grand mean of the distribution (β) and the cross-site standard deviation of the distribution (τ).[7] To estimate β and τ, we used the following two-level hierarchical linear model:

Level 1: Lottery Participants

$$Y_{ij} = \sum_{r=1}^{R} \alpha_r I_{rij} + B_j T_{ij} + \sum_{l=1}^{L} \gamma_l X_{lij} + e_{ij} \qquad (1)$$

Level 2: Sites

$$\beta_j = \beta + b_j \qquad (2)$$

where:

$e_{ij} \sim N(0, \sigma^2{}_{|\text{X,Lottery\_Block}}(T)$

$b_j \sim N(0, \tau^2)$

$Cov(e_{ij}, b_j) = 0$

In this model, Y_{ij} is the value of the outcome measure for individual *i* in site *j*, $\alpha_r I_{rij}$ equals 1 if individual *i* in site *j* belongs to lottery block *r* and zero otherwise, and T_{ij} equals 1 if individual *i* in site *j* was assigned to treatment and zero otherwise. We also include baseline covariates (race/ethnicity, gender, eligibility for free or reduced-price lunch, age, country of origin, and home language status), X_{lij}, to improve the precision of parameter estimates. The fixed random assignment block intercepts (α_r) account for the fact that individuals were randomly assigned within lottery blocks and that the proportion of sample members randomized to treatment can differ across lottery

[7] See Raudenbush and Bloom (2015) for discussions of related estimands.

blocks. The model allows for site-specific program-effect coefficients (β_j) that can differ randomly across sites. The β_j's are modeled as representing a cross-site population distribution with a mean value of β and a standard deviation of τ. Hence, the site-level random error term, b_j, has a mean of zero and a standard deviation of τ . Finally, the model allows for the variability of level-1 residuals to differ by treatment group. The individual-level random error term, e_{ij}, is assumed to have a mean of zero and a variance of $\sigma^2{}_{|\text{X,Lottery\_Block}}(T)$, which can be different for treatment group members and control group members.[8] To assess the statistical significance of τ, we used a chi-square test on a Q statistic. The Q statistic is widely used in meta-analysis to test for heterogeneity of effects (Hedges & Olkin, 1985).

Differences in ITT effects across sites. To estimate whether key site characteristics predict variation in impacts (Research Question 2), we selected a parsimonious set of prekindergarten program and elementary school characteristics measured at baseline (before school assignment). For each of these site characteristics, we estimated whether the main effect of treatment is moderated — that is, whether it is affected in direction or strength by its values. A simple presentation of this model is:

$$Y_{ij} = \beta_0 T_{ij} + \beta_1 T_{ij} * C_{ij} + \textstyle\sum_{i=1}^{I} \pi_i I_{ij} + \theta X_{ij} + \varepsilon_{ij} \qquad (3)$$

where the terms are defined as in equation 2, with the addition of C_{ji} which is the characteristic of interest for a given lottery (also called site), and ε_{ji} is a random error for student *i* that is clustered by the prekindergarten school that students entered after their lottery.[9] To help interpret the findings, for each site characteristic, after reporting β_0 and β_1, we used the site characteristic's values (reported in

[8] Bloom and colleagues (2017) provide further information about this model and Raudenbush and Bloom (2015) explore its properties.

[9] This information is only available for students who enroll in BPS prekindergarten. For students who lost a lottery, we assume that they attend independent settings.

the measures section) to estimate the magnitude of the treatment effect at the 25th, 50th, and 75th percentile sites. When doing so, we used a generalized linear hypothesis [GLH] test to test whether the differences between percentiles were statistically significant at the 0.05 level.

Note that Equation 3 assumes a linear relationship between treatment effects and the values of the site characteristic. When building our final analytic model, we did not rely on this assumption; rather, we followed the recommendations of Singer and Willett (2003) and fit a systemic sequence of models to determine the best estimation model for the data. Our first model imposed the fewest constraints on the relationship between the treatment by site effect and the outcome (using a general specification), and we then moved to the more constrained linear model (when appropriate). Specifically, we estimated the first model (a general specification) by dividing the sample into quintiles based on the values of each site characteristic and estimating quintile-by-quintile treatment effects. We then estimated a linear specification of the relationship and used a GLH test to determine whether the change in the goodness of fit imposed by the linear constraint was counterbalanced by the degrees of freedom gained. For all site characteristics, the linear model proved superior for the data.

We used Equation 3 to estimate the relationship between site characteristics and the effect of being *assigned* to a Boston prekindergarten program (i.e., ITT). When there was a clear pattern in ITT effects for a predictor, in an attempt to understand what the effect of *enrolling* was for students experiencing particularly high and particularly low values of the site characteristic, we divided our sample into subsamples of students who competed in a lottery for the bottom and top quartiles of the site characteristic distribution and estimated the Complier Average Causal Effect (CACE) using a standard application of instrumental variables analysis (Gennetian, Morris, Bos, & Bloom, 2005). The first stage was specified as:

$$E_{ij} = \beta_0 T_{ij} + \sum_{j=1}^{J} \pi I_{ij} + \theta X_{ij} + w_{ij} \qquad (4)$$

where E_{ji} is a BPS prekindergarten enrollment indicator equal to 1 if student *i* ever enrolled in BPS prekindergarten and zero otherwise, and all other terms are defined as in Equation 3. The second-stage equation was specified as:

$$Y_{ij} = \delta \hat{E}_{ij} + \sum_{j=1}^{J} \alpha I_{ij} + \theta X_{ij} + e_{ij} \qquad (5)$$

where $\hat{E}_{ji}$ equals the fitted value of the enrollment outcome from the first-stage equation, and e_{ij} is a random error that is clustered by the prekindergarten school that students entered after their lottery. The estimated value of δ is a consistent estimate of the average effect of enrolling in BPS prekindergarten for target BPS prekindergarten enrollees.

Results

RQ 1: Does the impact of BPS prekindergarten differ across sites?

As reported in Weiland and colleagues (2019), on measures of grade retention, special education identification, and third-grade ELA and mathematics achievement, BPS prekindergarten had an estimated grand mean (that is, average) effect that is not statistically significantly different from zero. However, we found that for all outcomes, there was small to moderate variation in the treatment effect *across sites* that was statistically significant at the 0.05 level. This variation was captured in the $\hat{\tau}$ statistic, a statistic commonly interpreted as the standard deviation of the site-specific treatment effect estimates as they vary randomly around the grand mean treatment effect. For example, if a sample produced an estimated grand mean treatment effect of zero and a $\hat{\tau}$ statistic of 1, 68 percent of the sites would have estimated treatment effects between -1 and 1, and 95 percent of the sites would have estimated treatment effects between -2 and 2.

Figures 2 and 3 illustrate the distribution of treatment effects across BPS prekindergarten sites for two measures of students' academic progress: ever retained in grade and ever identified as special education.[10] The distribution of prekindergarten effects on students' probability of being retained in grade is fairly narrow; roughly two-thirds of the sites produced effects relatively close to zero (between -1.72 and 1.72 percentage points). It is worth noting that this analysis captures the effects of being *assigned* to the program. The distribution of *enrollment* effects from a standard Wald adjustment is wider; roughly two-thirds of the sites produced enrollment effects with between -6 percentage points and 6 percentage points. Similarly, the distribution of prekindergarten effects on students' probability of being identified for special education is wider than the distribution of retained-in-grade effects: 68 percent of the sites have produced assignment effects between -4.53 and 4.53 (which would translate into enrollment effects ranging from roughly -20 to 20 percentage points).

Figures 4 and 5 illustrate the distribution of treatment effects across BPS prekindergarten sites for the two measures of students' academic achievement. Both of these measures are standardized on the district's third-grade mean and standard deviation for the testing year. The site-level estimated effects of assignment to BPS prekindergarten on students' academic achievement range from -0.3 to 0.2. Both distributions have estimated tau statistics close to 0.1, meaning that 68 percent of the sites produced effects between -0.1 and 0.1 on students' academic achievement measures. Again, it is worth noting that the distribution in the effect of enrolling in BPS prekindergarten is wider, and with a simple Wald adjustment, roughly 68 percent of the sites would have approximate enrollment effects ranging from -0.34 to 0.34. The typical third- to fourth-grade reading gains are 0.36 standard deviations; thus, this range of effects is quite large and covers approximately two years of typical growth (Hill, Bloom, Black, & Lipsey, 2008).

[10] Like in Weiss and colleagues (2017), the cross-site estimates are constrained to ensure a cross-site variance equal to estimate tau (see Bloom, Raudenbush, Weiss, & Porter, 2017). This constraint adjusts for the fact that conventional empirical Bayes estimates tend to understate true variability across estimates (Raudenbush & Bryk, 2002).

We also estimated the effect of being assigned to Boston prekindergarten on students' constrained and unconstrained ELA skills. The average effect size (ES) of treatment assignment on both student outcomes was not statistically significantly different from zero (ES = -0.0004, *p* = 0.962 and ES = 0.008, *p* = 0.418 for constrained and unconstrained, respectively). For both outcomes, variation in the cross-site distribution of effects was not statistically significantly different from zero ($\hat{\tau}$ could not be estimated and $\hat{\tau}$ was 0.001, *p*-value = 0.1870 for constrained and unconstrained outcomes, respectively) and the distributions could not be illustrated.

In re-analyses of the nationally representative randomized trial of Head Start, Bloom and Weiland (2015) and Walters (2015) reported variation across sites of between 0.12 and 0.17 standard deviations on end-of-preschool outcomes. While the estimated effects reported here are slightly less than 0.15 standard deviations, they are produced within one district, so one may expect a narrower distribution. (Variation in the counterfactual condition across sites may also contribute to variation in effects; we discuss this in greater detail in the next section.) Weiss and colleagues (2017) found that effects from highly specific interventions vary less across sites ($\hat{\tau}$=0.03 SD), while less specific interventions had more variation across sites ($\hat{\tau}$=0.12 SD); per their definition, the specific curricula and coaching model of the BPS prekindergarten program qualifies it as a relatively specific intervention. In this context, the variation we found of roughly 0.10 SD is promising for the purposes of predicting variation across sites.

RQ 2: Do BPS prekindergarten programs located within higher-quality elementary schools produce different impacts than those located within lower-quality elementary schools?

Per our study's goals, we also examined whether measures of school quality predicted the variation in site impacts that we observed. When discussing all findings in this section, we focus primarily on the estimated treatment effects for students who competed for schools at the 25th, 50th, and 75th percentiles, as these findings have a clear (or clearer) substantive interpretation compared with the estimated coefficients.

Demand for program. Demand for a given program was the only site characteristic for which a linear model was not appropriate. A visual inspection of the quintile-by-quintile treatment effects and a GLH test using the model fit statistics confirmed that a quadratic specification of demand (adding an interaction between demand squared and treatment to Model 4) was the superior model.[11]

The first panel of Table 2 presents the estimated effect of winning a lottery as a function of treatment, an interaction between treatment and the popularity of each site at baseline, and an interaction between treatment and the popularity of each site at baseline squared. There was a statistically significant relationship between the demand for a site and treatment effects for one outcome — the likelihood that a student will be classified as special education. As shown in the far right columns of Table 2, in sites scoring at the 25th percentile, winning a lottery increased students' risk of special education identification by 2.54 percentage points, while in sites at the 75th percentile, winning a lottery slightly *decreased* students' risk of special education identification by 0.829 percentage point. Both of these effects were relatively small. Across all other four student outcomes, there was little relationship between the demand for a given program in the school assignment process and estimated treatment effects.

School-level third-grade ELA and Math proficiency. Aggregate third-grade ELA and math proficiency levels can also serve as a proxy for the quality of the school. While there was no effect on the probability of a student ever being retained in a grade, for all other student outcomes, as the average school proficiency level becomes larger, the treatment effect's benefit to students also

[11] Results available upon request.

increases. Specifically, as shown in Table 2, on average, in sites scoring at the 25th percentile, winning a lottery increased students' risk of special education identification (4.84 percentage points), but in sites at the 75th percentile, winning a lottery *decreased* students' risk of special education identification by 1.5 percentage points. Similarly, in sites scoring at the 25th percentile, winning a lottery negatively affected students' math and ELA scores (ES = -0.05 SD, and ES = -0.08, respectively), but in sites at the 75th percentile, winning a lottery *positively* affected students' math and ELA scores (ES = 0.08 and ES = 0.04, respectively).

Given the systematic relationship between aggregate third-grade ELA and math proficiency levels and treatment effects presented above, Table 3 presents *enrollment effects* for the students who competed for sites in the bottom quartile of schools and the top quartile of schools. For example, the second row shows that 13.54 percent of the students who won a lottery for a prekindergarten program in the bottom quartile and enrolled were ever retained in early elementary school, while 1 percent of their control group counterparts had the same experience (ES = 12.58, *p*-value = 0.011). Of the students who won a lottery for a prekindergarten program in the top quartile and enrolled, 4 percent were retained, while 9 percent of their control group counterparts had the same experience (ES = -5.19, *p*-value = 0.433). The estimated enrollment effects on students' academic achievement are also striking. Students who won a lottery and enrolled in a bottom-quartile prekindergarten program experienced negative effects of roughly 0.50 and 0.36 standard deviations (for ELA and math, respectively), while students who won a lottery and enrolled in a top-quartile prekindergarten program experienced positive effects of roughly 0.45 and 0.66 standard deviations (for ELA and math, respectively). These enrollment effects (both positive and negative) are large in magnitude and represent three-fourths to a year of typical third- to fourth-grade growth (Hill, Bloom, Black, & Lipsey, 2008). Although these students represent those at the tails of the distribution, their experience indicates that the kindergarten-through-third-grade environment into which students are randomized is associated with whether they will experience either positive or negative effects of prekindergarten that are sustained through early elementary school.

Median Student Growth Percentile. The findings above suggest a relationship between the third-grade academic proficiency of students in a school and treatment effects, which may be due to the type of students the school attracts or the school's contribution to students' academic achievement (or a combination). In contrast, Massachusetts state data on every school's SGP ranking attempt to control for students' background characteristics and capture the school's contribution. Since the state began releasing these data in 2007, they were only available as baseline characteristics of the site (moderators) for cohorts 2 through 4.

Though our smaller sample for this analysis has less power, the third panel of Table 2 presents associations between this site characteristic and treatment effects that are similar, though weakened, to our third grade proficiency findings. For example, on average, in sites scoring at the 75th percentile on the SGP measure, winning a lottery improved students' ELA scores by 0.06 SDs. For sites scoring at the 75th percentile on the average proficiency measure, the corresponding improvement was 0.08 SDs. While limited, this analysis provides an important context for the third-grade academic proficiency associations reported above, and suggests that they are not solely driven by strong students self-selecting into schools with higher test scores.

Proportion of low-income students. While the proportion of low-income students within a school is not a proxy for school quality, it is a potential indicator of additional resources and, along with the SGP findings, can help put the third-grade academic proficiency findings in context. There was a statistically significant relationship between the proportion of low-income students within a school and treatment effects for one outcome — students’ scores on their third-grade ELA exams. On average, in sites scoring at the 25th percentile (having a low proportion of low-income students), winning a lottery had no effect on students’ ELA scores; in sites at the 75th percentile (having a high proportion of low-income students), winning a lottery had a negative effect of 0.07 standard deviations. Like the SGP analysis above, while limited, this analysis suggests that the relationship between third-grade academic proficiency levels and treatment effects reported above are not solely driven by students with different levels of family resources self-selecting into schools with higher test scores.

Percentage of kindergarten peers who received BPS prekindergarten. We found a statistically significant relationship between the percentage of kindergartners who received BPS prekindergarten within a site and treatment effects for one outcome — the likelihood that a student would be classified as special education. In sites scoring at the 25th percentile, winning a lottery increased students’ risk of special education identification by 3.31 percentage points, while in sites at the 75th percentile, winning a lottery had a lower increased risk of special education identification of 1.43 percentage points. Across all four other student outcomes, there was little relationship between the percentage of students attending kindergarten at the site who enrolled in BPS prekindergarten and estimated treatment effects.

Measures of the school climate. Using BPS teacher and student surveys, we constructed three alternative measures of school quality following Rochester and colleagues (2019) — reports of the school’s positive emotional climate, reports of teacher effectiveness and student engagement, and reports of principal effectiveness. As seen in Table 4, across all three measures, there were very few relationships to treatment effects. The one consistent finding comes from students’ academic achievement in mathematics: The effects of winning a lottery on this outcome were larger for schools with higher reports of teacher effectiveness and student engagement and schools with higher reports of principal effectiveness. However, as no other student outcomes showed a similar pattern of effects for these predictors, we view these findings as suggestive only.

Schools with High and Low Third-Grade Academic Proficiency Scores: Exploring the Treatment Contrast.

Differences in students' school experience. The most consistent predictors of a sustained boost were third-grade average proficiency scores. In Table 5, to better understand these findings, we examined the treatment-control contrast for students who competed for schools in the bottom and top quartiles of the school-level test score distribution. The means in Table 5 represent levels on key characteristics of the schools students actually experienced in their K-3 settings (i.e., enrollment coefficients; ITT coefficients are presented in Appendix Table D.1.) While there were a few key lottery-induced student experience differences, overall these differences in a magnitude appear too small to explain the effects reported in Table 3. For example, lottery winner enrollees who competed for sites in the 25th percentile of the distribution enrolled in schools with very similar proportions of low-income students as their control group counterparts. At the 75th percentile, lottery winner enrollees enrolled in schools with 6 percent fewer low-income students than their control group counterparts. In addition, the racial distribution of the student bodies experienced by lottery winner enrollees was very similar to that experienced by their control group counterparts in the 25th percentile of the distribution, while at the 75th percentile of the distribution, lottery winner enrollees enrolled in schools with 6 percent fewer African-American students and 9 percent more white students.

Differences in students' counterfactual. Another possible explanation is that our school-level test scores results are due not to school quality but to a differential counterfactual for control group children in lower- versus higher- performing schools. Using parent-reported data for two cohorts of students (2007 and 2008), we found that 41 percent of the control group members who did not enroll in BPS prekindergarten reported attending private prekindergarten, 18 percent reported attending Head Start, 13 percent received family-based day care, and the remaining 28 percent were at home with a parent or guardian. With data available for only two cohorts of students, we are unable to rigorously explore variation in the counterfactual setting across sites, though a descriptive look at average levels of setting exposure within these limited data shows no evidence of variation by setting type. Although the counterfactual setting was strong, it was strong for all students across all sites, and thus variation in the P-3 quality must be at least partially responsible for the observed variation in effects.

Discussion

While there is a consensus that attending preschool better prepares children for kindergarten (Phillip et al., 2017), evidence on the factors that sustain the preschool boost into the early elementary years is still emerging. Using a unique sample of students who competed for the Boston prekindergarten program and new methods for estimating and predicting impact variation (Bloom et al., 2017), we find that effects on prekindergarten enrollees' K-2 grade retention, K-3 special education placement, and math and ELA third grade test scores vary across schools, from substantially negative to substantially positive. We also found evidence aligned with the sustaining environment hypothesis, which posits that children's experiences after prekindergarten matter for sustaining the boost (Baily et al., 2017). Specifically, we found that one specific aspect of early elementary schools seems to matter more than others in sustaining the prekindergarten boost– the average test score proficiency levels of schools at the time of students' application for prekindergarten. Other proxies for school quality such as the percent of free/reduced lunch students in the school and demand for prekindergarten were not consistent predictors of a lasting boost.

Importantly, ours is the first empirical investigation of how the impacts of public prekindergarten programs vary across sites in the early elementary years, making the amount of variation across sites we detected hard to gauge. That said, the variation appears quite substantial when compared with some existing benchmarks in the field. For example, the range of enrollment effects on students' academic achievement in ELA was relatively large — sites one standard deviation below the mean produced negative effects roughly equal to one year of the typical third- to fourth-grade growth, and sites one standard deviation above the mean produced positive effects of the same size (Hill et al., 2008).

Our variation findings are driven by the quality of the BPS prekindergarten program and the students' subsequent elementary school experience, by student selection into higher test score schools, or the services received by control group members in the absence of the program, or a combination. Overall, three pieces of evidence in our study point to the former – that the quality of the schools themselves appear to be drivers of a lasting boost (or not). First, our findings for predicting a lasting boost were similar for the average test score proficiency levels of schools at the time of students' application for prekindergarten and for an alternative measure that attempts to control for students' background characteristics and capture the school's contribution to student growth. Accordingly, our findings that test scores are the best predictor of a lasting boost does not appear to be solely driven by higher-resource students self-selecting into higher-resourced schools. Second, in our treatment-contrast work, we found that prekindergarten program recipients in sites with positive third-grade effects were slightly more likely than their control group counterparts to have more economically advantaged and White early elementary school peers, but these differences do not appear large enough to be the sole explanation for the effects. And finally, differences in the counterfactual experienced by control group students who competed for bottom versus top quintile seats does not appear to be a driver of variation in impacts across sites. This was an important possibility to investigate, given evidence that the counterfactual matters greatly in preschool studies (Feller et al., 2016). We found in our context that the counterfactual was strong but it was strong for all students across all sites. Accordingly, variation in the P-3 quality must be at least partially responsible for the observed variation in effects.

Notably, our finding that the average test score proficiency levels of schools are the best predictor of a lasting boost is not the first such finding in the literature. Using a propensity scores approach, Zhai and colleagues (2012) found that the effects of a preschool enhancement intervention lasted into elementary school only for those who subsequently entered higher-quality elementary schools in Chicago. In our case, we hypothesize that when prekindergarten programs are nested within lower-performing elementary schools, they also struggle to positively affect prekindergarten students – e.g., that effects were not sustained in such contexts because they were small to begin with. Logically, if an elementary school is struggling to implement high-quality K-3 programming, it may not have resources to help the new prekindergarten program on its campus get off the ground and/or may struggle more than higher-performing schools to attract strong prekindergarten teachers. Alternatively, it could be the case that the prekindergarten boost was not maintained because of the quality of K-3 schooling – i.e, that lower-quality K-3 schooling did not build on a strong prekindergarten experience; or it could be a combination of these two processes. More investigation, including qualitative work, is needed to better understand processes *inside* schools with higher versus lower third grade test scores that led to a sustained prekindergarten boost versus not.

As another possibility, recent work on the sustaining environments hypothesis shows the importance of aligned instruction and content across students' preschool and early elementary experiences. For example, a recent randomized trial study of a preschool math curriculum found that the early preschool math effects were sustained more so when students experienced an aligned preschool and kindergarten math curriculum than just an enhanced prekindergarten math curriculum alone (Clements et al., 2013). It may be that higher-quality BPS schools fostered communication between their prekindergarten staff and early elementary school staff and naturally aligned their curricula to benefit students, more so than was the case in lower-performing BPS schools.

Finally, we found no support for the constrained versus unconstrained literacy skills hypothesis (McCormick et al., 2017; Snow & Matthews, 2016) – e.g., the idea that the preschool boost might depend on skill type, with more enduring effects expected on the latter than the former. For both skill types, the average effect of treatment did not differ from zero and the variation in the cross-site distribution of effects was not statistically significantly different from zero. Our findings could be due to measurement limitations, given that we created our measures from students' third grade ELA test items. Or, perhaps, this hypothesis does not explain why preschool effects persist. More work examining this hypothesis, with better measures, is needed.

This study has several important limitations. First, it is exploratory and noncausal in nature. Second, as mentioned earlier in the paper, our sample includes only students in over-subscribed schools, or about 25 percent of all applicants to the program. A propensity scores analysis on the full applicant sample found prekindergarten enrollment was associated with small benefits in K-3 on all examined outcomes (Weiland et al., 2019). As such, it is difficult to gauge the external validity of the current findings. Third, we are limited by the measurement of both our outcomes and moderators. A richer set of outcome measures covering the full range of relevant skills, collected each year students were in school, would have enhanced our study. Likewise, fine-grained measures of children's experiences in their classrooms, rather than school-level proxies for quality as well as their home learning experiences (Han, O'Connor, & McCormick, 2019; McCormick et al., 2019), might also have pointed to more specific factors more relevant for practice and policy. Finally, because measures of BPS *prekindergarten* quality are not available for the full study sample during this period, we cannot disentangle the relationship between prekindergarten quality and elementary school quality in this paper. For example, when we find higher lottery-based impacts for programs located in higher-quality elementary school sites, we cannot know whether this is because these prekindergarten sites produced larger impacts or whether the students' K-3 experience did a better job of sustaining them (or some combination).

Taken together, consistent with the weight of the empirical evidence to date (Ansari & Pianta, 2018; Clements et al., 2013; Currie & Thomas, 1998; Johnson & Jackson, 2019; Mattera et al., 2018; Pearman et al., 2019; Swain et al., 2015; Zhai et al., 2012), our exploratory results suggest that the quality of a student's early elementary school experience is an important piece of sustaining the prekindergarten boost. And importantly, descriptive statistics show that the post-prekindergarten schooling environments of our lottery sample children had room for improvement during this time period. Relative to other districts in the state, BPS in our focal years had relatively weak third-grade performance, around the bottom 11 percent of districts on the state third-grade standardized math test and the bottom 5 percent of districts for third-grade English Language Arts (Massachusetts Department of Elementary and Secondary Education, n.d.). As stakeholders consider how to best sustain the gains from preschool, investments in improving the quality of children's K-3 experiences and in aligning children's P-3 experiences so that later grades intentionally build on prior ones may be required.

References

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics, 126*(2), 699-748.

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

Ansari, A., & Pianta, R. C. J. D. P. (2018). Variation in the long-term benefits of child care: The role of classroom quality in elementary school. *54*(10), 1854.

Ashe, M. K., Reed, S., Dickinson, D. K., Morse, A. B., & Wilson, S. J. J. E. C. S. A. I. J. o. E. (2009). Opening the world of learning: Features, effectiveness, and implementation strategies.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. J. J. o. r. o. e. e. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *10*(1), 7-39.

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

Bassok, D., Finch, J. E., Lee, R., Reardon, S. F., & Waldfogel, J. J. A. O. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *2*(3), 2332858416653924.

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. J. C. d. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *85*(1), 140-159.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient. *Journal of Research on Educational Effectiveness, 10*(4), 817-842. doi:10.1080/19345747.2016.1264518

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*(2), 290-319.

Bloom, H. S., & Weiland, C. (2015). Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study. *Available at SSRN 2594430*.

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

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

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. J. J. f. R. i. M. E. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *42*(2), 127-166.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. J. A. E. R. J. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *50*(4), 812-850.

Clements, D. H., & Sarama, J. J. J. f. r. i. M. E. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. 136-163.

Cornell, D., Shukla, K., & Konold, T. R. J. A. O. (2016). Authoritative school climate and student academic engagement, grades, and aspirations in middle and high schools. *2*(2), 2332858416633184.

Currie, J., and Thomas, D. (1998). *School quality and the longer-term effects of Head Start*. NBER Working Paper No. 6362. Cambridge, MA: National Bureau of Economic Research.

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(3)*, 158-187.

Duncan, G. J., & Magnuson, K. J. J. o. e. p. (2013). Investing in preschool programs. *27*(2), 109-132.

Engel, M., Claessens, A., Watts, T., & Farkas, G. J. E. r. (2016). Mathematics content coverage and student learning in kindergarten. *45*(5), 293-300.

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G., & DiCrecchio, N. J. N. I. f. E. E. R. (2019). The State of Preschool 2018: State Preschool Yearbook.

Garces, E., Thomas, D., & Currie, J. J. A. e. r. (2002). Longer-term effects of Head Start. *92*(4), 999-1012.

Gatti, G., & Petrochenkov, K. J. P., PA: Gatti Evaluation Inc. (2010). Pearson SuccessMaker math efficacy study: 2009–10 final report.

Gennetian, L. A., Morris, P. A., Bos, J. M., & Bloom, H. S. (2005). Constructing Instrumental Variables from Experimental Data to Explore How Treatments Produce Effects. In H. S. Bloom (Ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Hastings, J. S., Kane, T. J., & Staiger, D. O. (2005). *Parental Preferences and School Competition: Evidence from a Public School Choice Program*. Retrieved from

Hastings, J. S., Kane, T. J., & Staiger, D. O. (2006). *Preferences and Heterogeneous Treatment Effects in a Public School Choice Lottery*. Retrieved from

Hastings, J. S., Van Weelden, R., & Weinstein, J. (2007). *Preferences, Information, and Parental Choice Behavior in Public School Choice*. Retrieved from Cambridge, MA:

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Boston: Academic Press, Inc.

Henry, G. T., & Rickman, D. K. J. E. o. e. r. (2007). Do peers influence children's skill development in preschool? *, 26*(1), 100-112.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Hoxby, C. M. (2007). *The economics of school choice*: University of Chicago Press.

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. J. J. o. r. o. e. e. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *, 11*(3), 339-374.

Johnson, R. J. M. i. p. (2013). School quality and the long-run effects of Head Start. 159-208.

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. J. C. d. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *, 82*(6), 1768-1777.

Kelly, D., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. J. N. B. o. E. R. (2013). Performance of US 15-Year-Old Students in Mathematics, Science, and Reading Literacy in an International Context. First Look at PISA 2012. NCES 2014-024.

Kraft, M. A., Marinell, W. H., & Shen-Wei Yee, D. J. A. E. R. J. (2016). School organizational contexts, teacher turnover, and student achievement: Evidence from panel data. *53*(5), 1411-1449.

Ladnier-Hicks, J., McNeese, R. M., Johnson, J. T. J. J. o. C., & Instruction. (2010). Third grade reading performance and teacher perceptions of the Scott Foresman Reading Street program in Title I schools in South Mobile County. *4*(2), 51.

Massachusetts Department of Elementary and Secondary Education. (2011). *MCAS student growth percentiles: Interpretive guide*. Retrieved from http://www.doe.mass.edu/mcas/growth/InterpretiveGuide.doc.

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

Mattera, S., Jacob, R., & Morris, P. J. N. Y. M., March. (2018). Strengthening children's math skills with enhanced instruction: The impacts of Making Pre-K Count and High 5s on kindergarten outcomes.

McCormick, M., Hsueh, J., Weiland, C., & Bangser, M. J. M. (2017). The Challenge of Sustaining Preschool Impacts: Introducing ExCEL P-3, a Study from the Expanding Children's Early Learning Network.

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., . . . Shonkoff, J. P. J. E. R. (2017). Impacts of early childhood education on medium-and long-term educational outcomes. *46*(8), 474-487.

National Assessment Governing Board, W., DC. (2008). *Reading framework for the 2009 national assessment of educational progress*: ERIC Clearinghouse.

Neidell, M., Waldfogel, J. J. T. R. o. E., & Statistics. (2010). Cognitive and noncognitive peer effects in early education. *92*(3), 562-576.

Paris, S. G. J. R. r. q. (2005). Reinterpreting the development of reading skills. *40*(2), 184-202.

Phillips, D., Gormley, W., & Anderson, S. J. D. P. (2016). The effects of Tulsa's CAP Head Start program on middle-school academic outcomes and progress. *52*(8), 1247.

Phillips, D., Lipsey, M., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., and 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.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., & Jenkins, F. J. N. C. f. E. S. (2012). Highlights from TIMSS 2011: Mathematics and Science Achievement of US Fourth-and Eighth-Grade Students in an International Context. NCES 2013-009.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. *American Journal of Evaluation*, 1098214015600515.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Rittle-Johnson, B., & Schneider, M. J. O. h. o. n. c. (2015). Developing conceptual and procedural knowledge of mathematics. 1118-1134.

Rochester, S. E., Weiland, C., Unterman, R., McCormick, M., & Moffett, L. J. E. C. R. Q. (2019). The little kids down the hall: Associations between school climate, pre-K classroom quality, and pre-K children's gains in receptive vocabulary and executive function. *48*, 84-97.

Sameroff, A. (2009). *The transactional model*: American Psychological Association.

Sherblom, S. A., Marshall, J. C., & Sherblom, J. C. J. J. o. R. i. C. E. (2006). The relationship between school climate and math and reading achievement. *4*(1-2), 19-31.

Singer, J. D., Willett, J. B., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*: Oxford university press.

Snow, C. E., & Matthews, T. J. J. T. F. o. C. (2016). Reading and language in the early grades. 57-74.

Swain, W. A., Springer, M. G., & Hofer, K. G. J. A. O. (2015). Early grade teacher effectiveness and pre-K effect persistence: Evidence from Tennessee. *1*(4), 2332858415612751.

Walters, C. R. J. A. E. J. A. E. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *7*(4), 76-102.

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. J. E. C. R. Q. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *28*(2), 199-209.

Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2019). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development*.

Weiland, C., & Yoshikawa, H. J. C. D. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *84*(6), 2112-2130.

Weiland, C., & Yoshikawa, H. J. J. o. A. D. P. (2014). Does higher peer socio-economic status predict children's language and executive function skills gains in prekindergarten? *, 35*(5), 422-432.

Weiland, C. J. B. S., & Policy. (2016). Launching Preschool 2.0: A road map to high-quality public programs at scale. *2*(1), 37-46.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials. *Journal of Research on Educational Effectiveness*, 1-34.

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.

Wilkerson, S., Shannon, L., & Herman, T. J. A., Texas: Magnolia Consulting. (2006). An efficacy study on Scott Foresman's Reading Street Program: Year one report.

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

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. J. T. F. o. C. (2016). When does preschool matter? , 21-35.

Zhai, F., Raver, C. C., Jones, S. M. J. C., & Review, Y. S. (2012). Academic performance of subsequent schools and impacts of early interventions: Evidence from a randomized controlled trial in Head Start settings. *34*(5), 946-954.

Figures

Figure 1
Application Process for the Full Analytic Sample

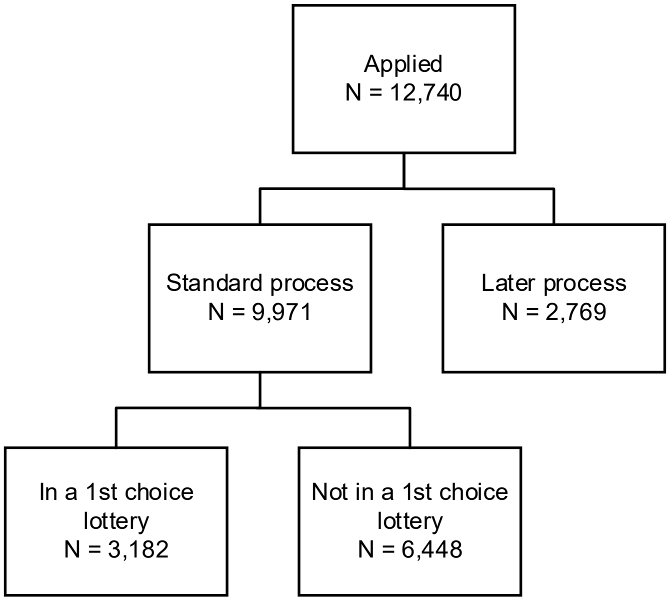
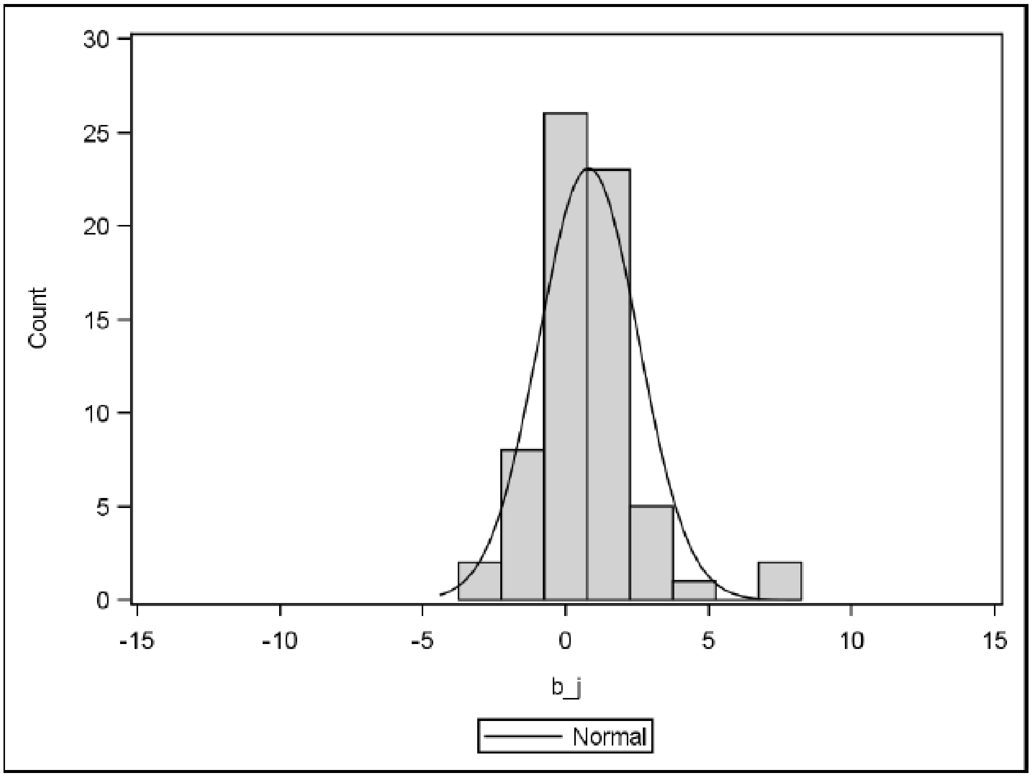


Figure 2

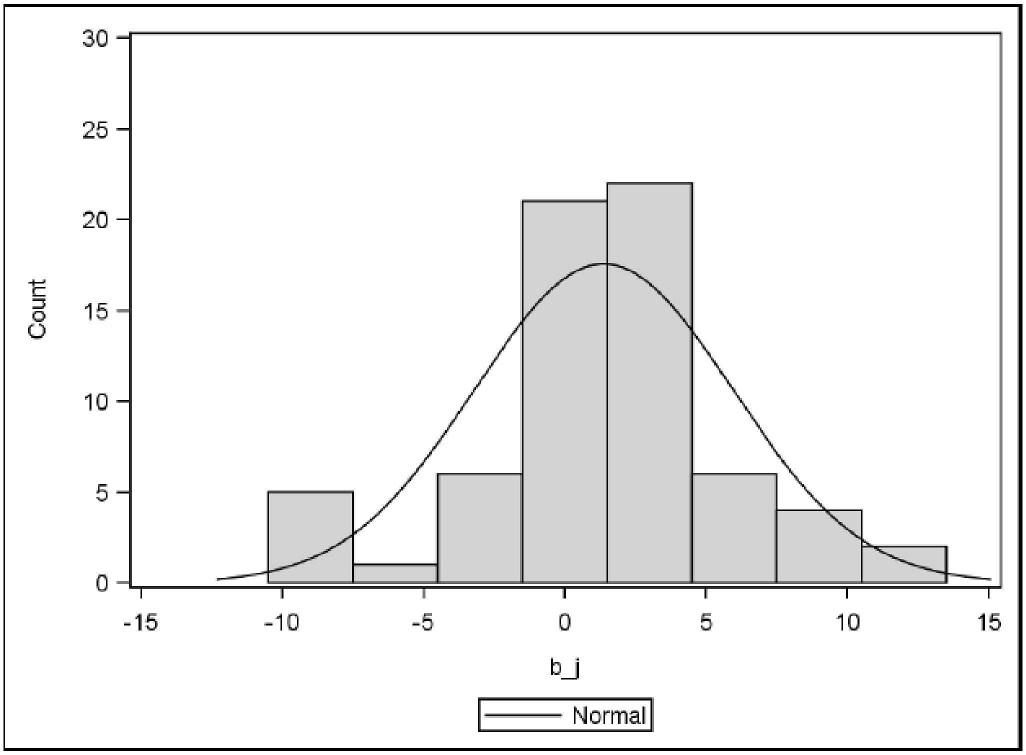
Histogram of Site-level Constrained Empirical-Bayes Impact Estimates on Ever Retained



Estimated tau=1.72, *p*-value on Q-statistic=0.043

Figure 3

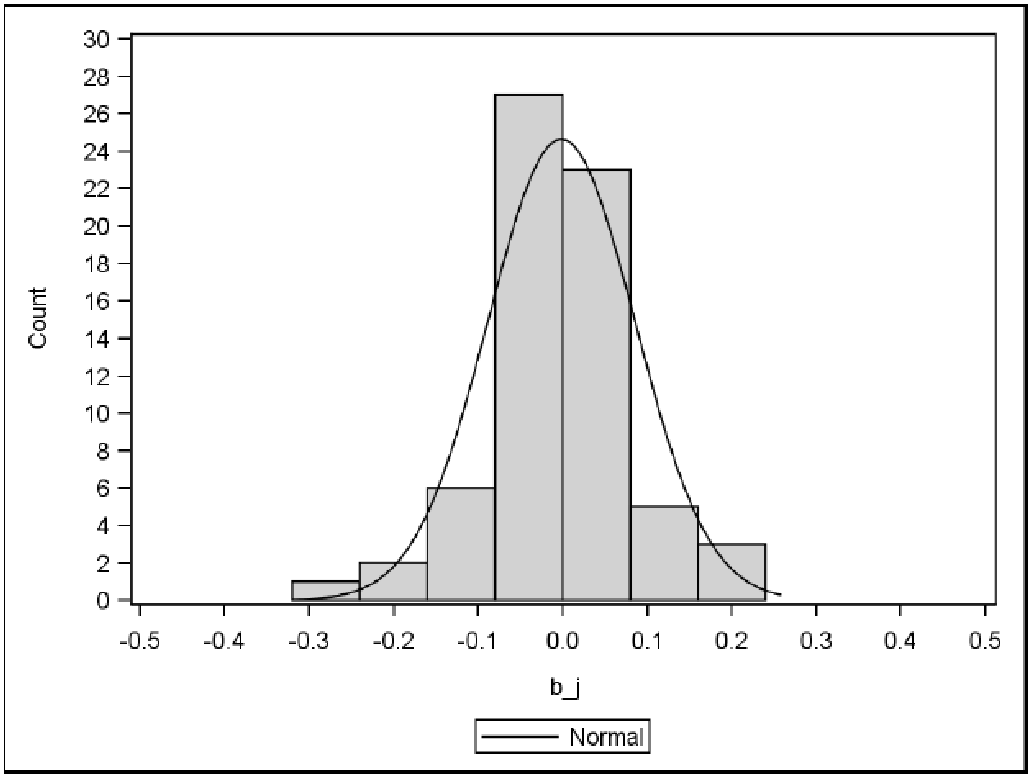
Histogram of Site-level Constrained Empirical-Bayes Impact Estimates on Ever Identified as Special Education



Estimated grand mean difference: 1.37, *p*-value= 0.376
Estimated tau=4.53, *p*-value on Q-statistic= 0.019

Figure 4

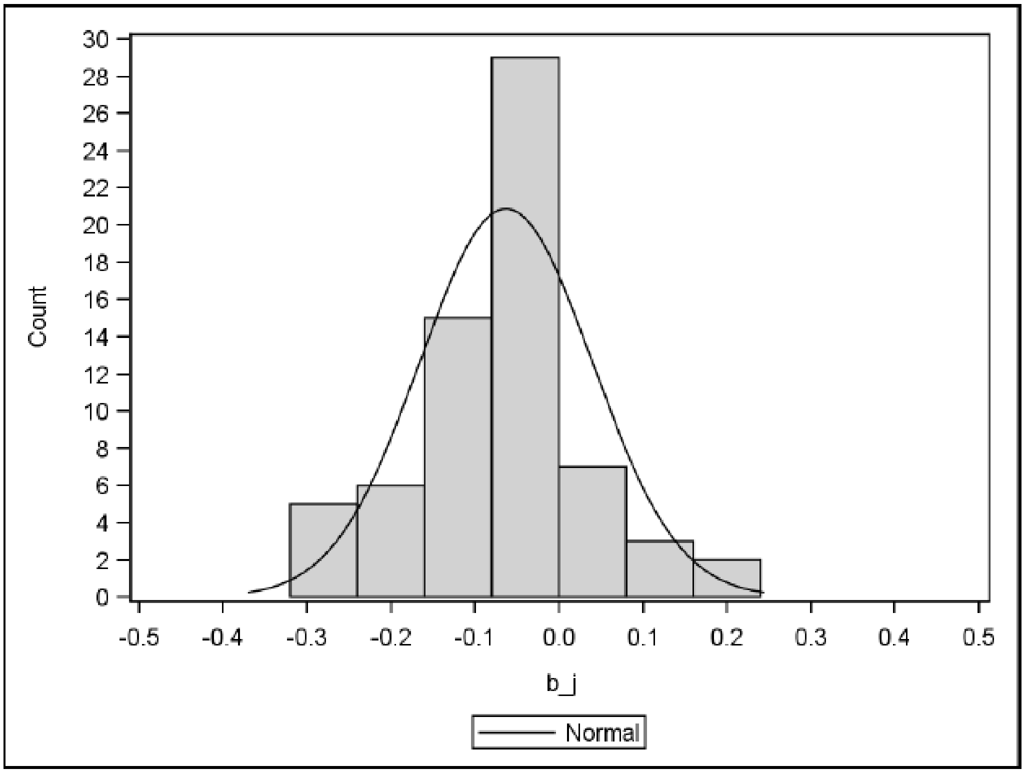
Histogram of Site-level Constrained Empirical-Bayes Impact Estimates on ELA test scores



Estimated grand mean difference: -0.002, *p*-value= 0.968
Estimated tau=0.095, *p*-value on Q statistic=0.043

Figure 5

Histogram of Site-level Constrained Empirical-Bayes Impact Estimates on Math scores



Estimated grand mean difference: -0.06, *p*-value= 0.1270
Estimated tau=0.102, *p*-value on Q statistic=0.032

Tables

Table 1

Baseline Characteristics of Lottery Sample

Characteristic	Lottery Sample	Full applicant sample	Difference
Race/ethnicity (%)			
Latino	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 lottery participants	3,182	12,740	-9,558

Note: In the lottery sample, there was a small amount of missing data on all covariates 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

Predictors of the Treatment Effect

Outcome	Treatment Coefficient	*P*-Value	Site Char. x Treatment Coefficient	*P*-Value	Site Char.2 x Treatment Coefficient	*P*-Value	Total Treatment Effect, by Site Char. Percentile		
							25th	50th	75th
Demand (applicants per seat)									
Ever retained in years 2-4 (%)	0.943	0.643	0.034	0.918	-0.003	0.669	1.040	1.052	1.037
Ever special education in years 1-4 (%)	6.925*	0.034	-1.186*	0.047	0.035*	0.080	2.536	0.878	-0.829†
English Language Arts	-0.060	0.443	0.013	0.342	-0.001	0.086	-0.018	-0.006	0.003
Math	-0.088	0.303	0.009	0.560	-0.001	0.215	-0.060	-0.054	-0.051
Average 3rd grade proficiency									
Ever retained in years 2-4 (%)	2.867	0.196	-0.071	0.153			0.704	-0.289	-1.282
Ever special education in years 1-4 (%)	11.687*	0.001	-0.225**	0.005			4.837	1.692	-1.452†
English Language Arts	-0.177*	0.047	0.004*	0.022			-0.044	0.017	0.078†
Math	-0.234*	0.008	0.005*	0.021			-0.094	-0.030	0.034†
Student growth percentile (cohorts 2-4)									
Ever retained in years 2-4 (%)	-0.805	0.874	0.039	0.680			0.829	1.140	1.451
Ever special education in years 1-4 (%)	10.395*	0.041	-0.165	0.055			3.482	2.165	0.848†
English Language Arts	-0.329	0.105	0.007*	0.049			-0.045	0.009	0.063†
Math	-0.163	0.457	0.002	0.548			-0.058	-0.039	-0.019
Low-income students in school (low to high)									
Ever retained in years 2-4 (%)	-3.559	0.444	0.071	0.255			0.802	1.734	2.101
Ever special education in years 1-4 (%)	-10.290	0.146	0.148	0.081			-1.113	0.847	1.619
English Language Arts	0.266*	0.047	-0.004*	0.040			0.009	-0.046	-0.067†
Math	0.215	0.203	-0.004	0.102			-0.020	-0.071	-0.090
Percent of kindergarten peers in BPS prekindergarten									
Ever retained in years 2-4 (%)	4.550	0.080	-4.916	0.179			2.626	0.555	1.600
Ever special education in years 1-4 (%)	-1.917	0.594	2.568	0.614			-0.912	0.170	-0.376
English language arts	-0.097	0.317	0.065	0.631			-0.072	-0.044	-0.058
Math	-0.054	0.562	0.081	0.523			-0.023	0.011	-0.006

NOTE: * P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.
† *P*-value < 0.05 for difference across percentiles.

Table 3

Effects of Enrollment in Bottom and Top Quartile Third-Grade Math Proficiency Site Subgroups

	Bottom Quartile of Site Characteristic				Top Quartile of Site Characteristic			
Outcome	Lottery Winner Compliers	Control Group Compliers	Estimated Difference	*P*-Value for Estimated Difference	Lottery Winner Compliers	Control Group Compliers	Estimated Difference	*P*-Value for Estimated Difference
Ever retained in years 2-4	13.54	0.96	12.58*	0.0109	4.04	9.23	-5.19	0.4326
Ever special education 1-4	18.5	6.05	12.45	0.0722	15.42	23.47	-8.05	0.4516
English Language Arts	0.22	0.72	-0.50*	0.0042	0.64	0.19	0.45*	0.0292
Math	0.25	0.61	-0.36*	0.034	0.76	0.10	0.66*	0.0162
N lottery participants	285	645			235	544		

NOTE: * *P*-value < 0.05 for impact estimates.

Table 4

Predictors of the Treatment Effect — School Climate Measures

Outcome	Treatment Coefficient	*P*-Value	Site Char. x Treatment Coefficient	*P*-Value	Total Treatment Effect, by Site Char. Percentile		
					25th	50th	75th
Positive emotional climate (cohorts 2-4)							
Ever retained in years 2-4 (%)	24.238	0.155	-8.178	0.162	1.337	1.175	-0.297
Ever special education in years 1-4 (%)	57.294*	0.040	-18.490	0.053	5.517	5.151	1.823
English Language Arts	-0.617	0.354	0.204	0.370	-0.046	-0.042	-0.005
Math	-1.048	0.153	0.346	0.170	-0.081	-0.074	-0.012
Teacher effectiveness and student engagement (cohorts 2-4)							
Ever retained in years 2-4 (%)	3.654	0.867	-0.976	0.886	0.999	0.520	0.393
Ever special education in years 1-4 (%)	44.338	0.164	-12.616	0.201	10.024	3.842	2.202
English Language Arts	-0.783	0.351	0.235	0.366	-0.143	-0.028	0.002
Math	-1.606*	0.044	0.484*	0.041	-0.289	-0.052	0.011†
Principal effectiveness (cohorts 2-4)							
Ever retained in years 2-4 (%)	0.093	0.996	0.121	0.982	0.485	0.509	0.524
Ever special education in years 1-4 (%)	4.421	0.872	-0.233	0.977	3.665	3.620	3.591
English Language Arts	-0.257	0.720	0.068	0.746	-0.036	-0.022	-0.014
Math	-1.478*	0.048	0.421*	0.046	-0.112	-0.030	0.023†

NOTES: Sample size is 1,101 for the treatment group, 2,081 for the control group.

* *P*-value < 0.05 for impact estimates.

† *P*-value < 0.05 for difference across percentiles.

Table 5

Effects of Enrollment on Students School Experience in Bottom and Top Quartile Third-Grade Math Proficiency Site Subgroups

	Top Quartile of Site Characteristic				Bottom Quartile of Site Characteristic			
Outcome	Lottery Winner Compliers	Control Group Compliers	Estimated Difference	*P*-Value for Estimated Difference	Lottery Winner Compliers	Control Group Compliers	Estimated Difference	*P*-Value for Estimated Difference
English language learners (%)	31.59	30.82	0.78	0.265	22.44	23.35	-0.91	0.3045
Students with disabilities (%)	17.15	17.28	-0.12	0.609	17.33	17.92	-0.59	0.1549
Low-income (%)	69.94	70.63	-0.69	0.422	51.31	57.49	-6.18**	< 0.0001
African-American (%)	30.71	31.55	-0.84	0.281	15.86	22.12	-6.26**	< 0.0001
Asian (%)	4.6	5.77	-1.17**	0.001	13.77	12.88	0.89	0.2107
Hispanic (%)	46.58	43.7	2.88**	0.001	26.4	29.66	-3.20**	0.0053
White (%)	15.36	15.94	-0.59	0.5	40.34	31.45	8.89**	< 0.0001
Licensed to teach (%)	96.89	96.09	0.80*	0.044	97.7	97.78	-0.07	0.8807
Teacher-student ratio	13.61	13.52	0.09	0.164	14.42	14.17	0.26*	0.0341
Teacher retained (%)	79.46	79.96	-0.51	0.179	84	81.65	2.35**	< 0.0001
Average class size (N)	19.13	18.71	0.42*	0.031	19.07	19.23	-0.16	0.4795
Average teachers proficient (%)	78.06	80.21	-2.16**	0.004	84.33	82.64	1.69	0.1109
Average teachers exemplary (%)	14.24	12.2	2.04**	0.003	11.32	12.82	-1.50	0.146
Student stability (%)	87.23	86.5	0.73	0.011	93.00	90.73	2.27**	< 0.0001
Sample size (all lottery participants)	285	645			235	544		

NOTES: * *P*-value < 0.05, ** *P*-value < 0.01.