Working Paper

Quantifying and Predicting Variation in the Medium-Term Effects of Oversubscribed Prekindergarten Programs

Rebecca Unterman
(MDRC)

Christina Weiland
(University of Michigan)

July 2019

This study is funded by the Institute of Education Sciences RFA R305A140059.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Arnold Ventures, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

Correspondence concerning this article should be addressed to Christina Weiland, School of Education, 610 E. University Ave, Ann Arbor, MI 48104; email: weilandc@umich.edu.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Abstract

In this paper we use data from students who participated in the oversubscribed Boston Public Schools (BPS) prekindergarten program as a window into variation in the program's medium-term effects. We first examine whether, for the sample of students who applied to oversubscribed BPS prekindergarten programs, there is variation in the effects of the Boston prekindergarten program on children's kindergarten-through-second-grade retention, kindergarten-through-third-grade special education placement, and third-grade state test scores. We find statistically significant variation in effects on 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. Students who won a lottery for a prekindergarten program in a school with third-grade academic proficiency scores in the bottom quartile of the distribution experience no or negative effects by third grade. In contrast, students who won a lottery for a prekindergarten program in a school with third-grade academic proficiency scores in the top quartile of the distribution experience positive effects by third grade. An exploration of how this quality measure is defined suggests that while a part of its predictive power may be related to the characteristics of the students who enroll in these schools (specifically, their family income level), it also appears that the schools themselves contribute to these effects. Prekindergarten gains persisted if kids applied to and won a seat in a higher-quality elementary school.

Contents

Abstract iii
List of Exhibits vii
Acknowledgments ix

The Mechanisms of Convergence 2
Method 5
Data Analytic Plan 12
Results 16
Discussion 26

Appendix

A Constrained and Unconstrained English Language Arts Measures 31

B Pearson Correlation Coefficients 37

C Average Third-Grade Academic Proficiency Treatment Contrast 41

References 45

List of Exhibits

Table

1	Baseline Characteristics	8
2	Predictors of the Treatment Effect	21
3	Effects of Enrollment for Third-Grade Math Proficiency Site Subgroups	23
4	Predictors of the Treatment Effect — School Climate Measures	25
5	Effects of Enrollment on Treatment Contrast for Third-Grade Math Proficiency Site Subgroups	27
A.1	Inter-Rater Reliability on Coding of Released MCAS ELA Items 2012-2014	34
A.2	Item Type and Classification Coding for Released MCAS ELA Items 2012	34
A.3	Item Type and Classification Coding for Released MCAS ELA Items 2013	35
A.4	Item Type and Classification Coding for Released MCAS ELA Items 2014	36
B.1	Correlations Between School-Level Predictors of Variation in Impacts Across Schools	39
C.1	Average Third-Grade Academic Proficiency Treatment Contrast	43

Figure

1	Application Process for the Full Analytic Sample	6
2	Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Ever Retained	17
3	Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Ever Identified as Special Education	18
4	Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — ELA	19
5	Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Math	20

Acknowledgments

This work greatly benefited from contributions and feedback from Anna Shapiro, Sara Staszak, Shana Rochester, Eleanor Martin, 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, and Carrie Conaway and the Massachusetts Department of Elementary and Secondary Education. Special thanks also to Hirokazu Yoshikawa, Howard Bloom, Richard Murnane, David Deming, Catherine Snow, Caroline Ebanks, and Gina Biancarosa.

At MDRC, Rebecca Bender (consultant) edited the paper, and Carolyn Thomas prepared it for publication.

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 and Magnuson, 2013; Phillips et al., 2017; Yoshikawa et al., 2013). However, the evidence is more mixed regarding how long the preschool advantage 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 (that is, 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 these programs mirrors the medium-term pattern of the older studies (Phillips et al., 2017; Yoshikawa, Weiland, and 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, and 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-skilled preschool attenders in place while nonattenders catch up. Notably, inklings of evidence for this hypothesis are present in the older literature as well. 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 and Pianta, 2018; Bassok et al., 2016; Bierman et al., 2014; Clements, Sarama, Wolfe, and Spitler, 2013; Jenkins et al., 2017; Zhai, Raver, and Jones, 2012).

Solving the convergence puzzle is one of the chief challenges facing the field of early childhood education. Stakeholders' 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. BPS has unusually high instructional quality in prekindergarten compared with other large-scale U.S. programs (Chaudry, Morrissey, Weiland, and Yoshikawa, 2017; Weiland, Ulvestad, Sachs, and Yoshikawa, 2013), but its kindergarten-through-third-grade (K-3) quality is weaker than its prekindergarten program (Weiland et al., in press). Recently, for the subset of district children who participated in oversubscribed lotteries for the program, we found that the large end-of-program gains experienced by BPS prekindergarten participants (Weiland and Yoshikawa, 2013) are, on average, not experienced by the lottery sample children as students move through early elementary school (though preschool attendance was associated with medium-term benefits within the full sample of applicants; Weiland et al., in press). The present study builds on this work and tests 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 high-quality elementary schools sustain effects at a greater rate than those located within low-quality elementary schools.

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 the educational setting in kindergarten and beyond, often referred to as the "sustaining environments" hypothesis (Bailey, Duncan, Odgers, and Yu, 2017). Intuitively, this focus makes sense. For children who received prekindergarten, 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 prekindergarten students' growth and eroding the prekindergarten program's impact. For students who did not enroll in prekindergarten, 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 prekindergarten impact.

There is some empirical evidence supporting the sustaining environments hypothesis. Four recent studies found that prekindergarten effects were more likely to be sustained if students subsequently experienced higher-quality early elementary school environments, as measured by spending (Johnson, 2013); by school-level third-grade standardized test scores (Zhai, Raver, and Jones, 2012); by a multidimensional measure of school resources, organization, and social processes (Ansari and Pianta, 2018); or, in Tennessee, by state ratings of first-grade teacher overall effectiveness (Swain, Springer, and Hofer, 2015). In the older literature, in their study of Head Start, Currie and Thomas (1998) found that effects lasted to Grade 3 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 (Currie and Thomas, 1998). Two recent studies examined post-preschool quality from the more proximal level of alignment of prekindergarten and kindergarten mathematics experiences. One of the two found persistent impacts on math for those in the aligned condition, but not in the enhanced prekindergarten math-only condition (Clements, Sarama, Wolfe, and Spitler, 2013). The other study found positive effects through the end of kindergarten for all children in the enhanced prekindergarten mathematics condition (versus business-as-usual prekindergarten) but considerably larger impacts for children in the prekindergarten and kindergarten alignment condition (Mattera, Jacob, and Morris, 2018).

In addition, peers are another potential avenue of stronger-quality elementary school experiences. 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, and Yu, 2017). Correlational work in preschool does support the existence of peer effects in these years (for example, Henry and Rickman, 2007; Justice, Petscher, Schatschnei-

der, and Mashburn, 2011; Weiland and 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 and Waldfogel, 2010), perhaps through direct peer effects or indirect teacher effects (for example, increased expectations for children or more time to focus on fewer struggling students).

In contrast to these findings, however, 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). In addition, two recent studies that examined the role of a host of early elementary school structural and process features in promoting a sustained prekindergarten boost found largely null results (Bassok et al., 2016; Jenkins et al., 2017). Accordingly, to date, the evidence provides no clear answer on whether what happens after prekindergarten holds the key to sustained effects, nor does it offer clarity on which specific post-prekindergarten elements matter most.

Notably, while we focus on the sustaining environments hypothesis, there are also several 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 prekindergarten effects may or may not be sustained. Attending prekindergarten 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, such positive effects are seen in the medium term — for example, 0.29 standard deviation (SD) or 10.1 percentage points for grade retention avoidance and 0.40 SD or 12.5 percentage points for special education placement avoidance (McCoy et al., 2015). 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 analysis of the full prekindergarten applicant sample (Weiland et al., in press).

Finally, another key to convergence could be which skills are emphasized and measured in the period from prekindergarten through third grade. The boost from a prekindergarten program that focuses on constrained skills — for example, the discrete set of basic literacy and mathematics skills that almost all children master by third grade, such as letter knowledge and simple counting — is likely to be less enduring than the boost from a program that focuses on students' deeper unconstrained skills, meaning more broadband skills like world knowledge, vocabulary, conceptual thinking, and problem solving. Both kinds of skills are important for children's early 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 and Matthews, 2016). This neglect could explain why success in second- or third-grade reading or math does not automatically translate into long-term academic achievement, as shown by the fact that U.S. students score relatively well in Grade 4 international comparisons but much more poorly in Grades 8 and 10 (Kelly et al., 2013; Provasnik et al., 2012).

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

The Boston Public Schools prekindergarten program is a relatively large-scale program that is based entirely in the public schools, pays teachers on the same scale as kindergarten-through-twelfth-grade (K-12) teachers, subjects teachers to the same educational requirements as K-12 teachers, and is 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 and Sarama, 2007; Clements et al., 2011), though the evidence base for Building Blocks is stronger than that for OWL.

In 2007 to 2009, curricula implementation was supported via trainings and regular coaching, meaning weekly to biweekly on-site support from an experienced early childhood coach trained in both curricula (see Weiland and Yoshikawa, 2013, for additional details). In 2009 to 2011, 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 master's degrees; usually do not pay prekindergarten teachers on the same scale as K-12 teachers; 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). The Boston 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, Morrissey, Weiland, and Yoshikawa, 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 and Yoshikawa, 2013). Effects on language and mathematics were the largest among programs examined using the age-cutoff approach. Effects were particularly pronounced for Hispanic students, low-income students, and children with special needs (Weiland, 2016). However, recent work in Boston has shown that the outcomes of preschool attenders and nonattenders in a lottery-based subsample across four years (the focal sample in the present study) converge over time (Weiland et al., in press).

Regardless of whether families are offered a seat in the prekindergarten program, all families are guaranteed a seat in kindergarten. In our study period, 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 and Petrochenkov, 2010; Ladnier-Hicks, McNeese, and Johnson, 2010; What Works Clearinghouse, 2013; Wilkerson, Shannon, and Herman, 2006), nor were the supports for implementing them as systematic or rich as for the prekindergarten program. Classroom-quality data collected by the Wellesley Centers for Women in spring 2012 on 84 K-3 classrooms in BPS and in spring 2010 on 83 prekindergarten classrooms and reanalyzed by our study team show that prekindergarten classroom quality was higher on average than K-3 quality (Weiland et al., in press). For example, prekindergarten classrooms scored 5.6 on the CLASS emotional support and 4.3 on instructional support, 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). The district responded to this and related evidence by subsequently (not in our study period) developing its own kindergarten-through-second-grade (K-2) curriculum and associated professional development program (Boston Public Schools, 2017).

Current 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?

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? If so, are there measurable features of the students' school experiences that account for the differences in impacts?[1]

Method

Data Set

We addressed our research questions using data made available to the project team by BPS and the Massachusetts Department of Education. We began with data on students' choices and baseline demographics during the BPS assignment process from the spring of the 2006-07 through 2009-10 school years (for enrollment in 2007-08 through 2010-11). We merged these data, by each student's unique identifier, with district and state administrative records covering the school years students were age-eligible for prekindergarten through third grade. School

[1]The majority of the students who compete for a BPS prekindergarten site stay at that school through early elementary school.

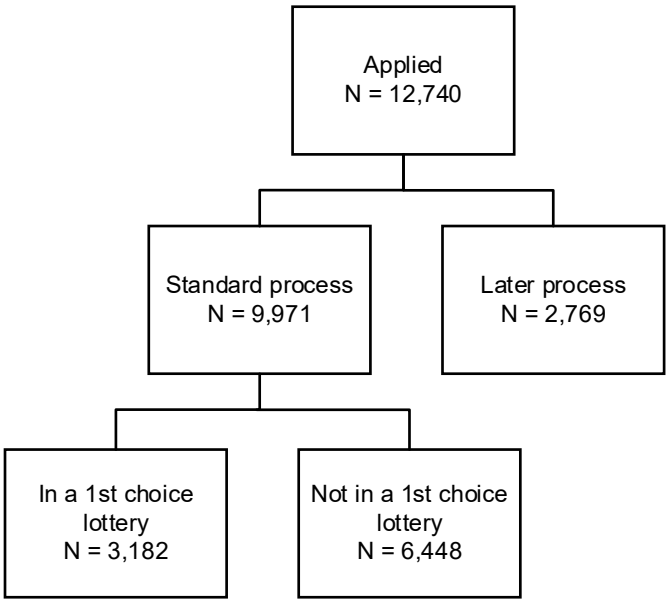
context data from the state of Massachusetts and school climate data from BPS exist at the school level; we merged these school-level data, by follow-up year, onto each student's file by the school identifier for the school he or she was enrolled in for the longest period of time during that school year.

Sample

Our sample comes from the population of students who applied to the Boston prekindergarten program for four-year-olds. 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 (described in greater detail in the next section), 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 the figure) involving 3,182 students, or 25 percent of all appliers and 32 percent of those who applied through the standard process.

Figure 1

Application Process for the Full Analytic Sample



Lottery Process Details

We used naturally occurring lotteries within the BPS school assignment process to identify an experimental sample consisting of 3,182 students who competed for a seat in an oversubscribed prekindergarten program. This process is described in detail in Weiland et al. (in press). The 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 et al. (in press), a joint F-test used to assess the statistical significance of the overall baseline (or preexisting) 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 (Weiland et al., in press).

In Table 1, we present the baseline characteristics of the experimental sample lottery alongside those of all students who applied to Boston's 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. Regarding students' race/ethnicity, Hispanic students made up 39 percent of the experimental sample versus 44 percent of all BPS prekindergarten appliers; white students accounted for 28 percent of the experimental sample versus 17 percent of all BPS prekindergarten appliers. About 21 percent of the experimental sample was black versus 28 percent of the full applicant sample. About 51 percent of the experimental sample qualified for free or reduced-price lunch, while 65 percent of all BPS appliers did. Fifty-seven percent of the experimental sample spoke English at home versus 50 percent of the full sample.

In each year, the proportion of BPS applicants who became lottery participants was between 30 and 32 percent. Across BPS, the percentage of schools represented by lotteries ranged from 67 to 83. However, as reported in Weiland et al. (in press), the lottery sample students are highly concentrated in a subset of the BPS prekindergarten sites. A less rigorous analysis of the full prekindergarten applicant sample suggests that effects for the lottery sample probably do not generalize to all Boston prekindergarten students. More specifically, a propensity score analysis on the full applicant sample found prekindergarten enrollment was associated with benefits in K-3 on all examined outcomes.

BPS prekindergarten programs are located within elementary school buildings; when lottery participants win a seat in a prekindergarten program, they are automatically enrolled in its attached BPS elementary school for kindergarten.[2] For the purposes of this paper, students who competed for an oversubscribed prekindergarten program actually competed for a school

[2]As reported in Weiland et al. (in press), 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.

Table 1
Baseline Characteristics

Characteristic	Lottery Sample	Full Applicant Sample	Difference
Race/ethnicity (%)			
Hispanic	39.21	43.87	-4.69
Black	21.48	28.43	-6.92
White	28.27	17.06	11.17
Asian	7.13	7.58	-0.47
Other	3.91	3.06	0.81
Other characteristic			
Male (%)	49.24	51.72	-2.46
Eligible for free/reduced-price lunch (%)	50.60	65.07	-14.50
Age (years)	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
Sample (N)	3,182	12,740	

NOTES: In the lottery sample, there was a small amount of missing data on all covariates except age: 12 children (0.4 percent) were missing race/ethnicity and gender information, 34 (1.1 percent) were missing gender and free/reduced-price lunch information, 113 (4.2 percent) were missing country of origin information, and 5 (0.2 percent) 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 percent) were missing race/ethnicity information, 185 (1.5 percent) were missing gender and free/reduced-price lunch information, 514 (4.0 percent) were missing country of origin information, and 499 (3.9 percent) were missing home language information. Means in the table were computed using nonmissing data.

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 attendance) 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.

Student Outcomes

Grade retention and special education identification. We used state administrative records to follow students' progressions through early elementary school and special education identification. (For more details, see Weiland et al., in press.)

Third-grade ELA and mathematics achievement. For third-grade ELA and mathematics analyses, we used students' statewide mathematics and ELA standardized tests. We standardized each student's theta score on the mean and standard deviation of all third-graders within BPS taking the given exam in that year. Test score data in this paper accordingly can be interpreted as a given group's performance compared with that of the average BPS third-grader. These data are from the state and are available for around 88 percent of the sample. Follow-up data availability is not a function of treatment assignment. (For more details, see Weiland et al., in press.)

Constrained ELA and unconstrained ELA skill development. There is a consensus among literacy experts that reading comprehension is an unconstrained skill — that is, there is always room for improvement (in contrast to constrained skills like letter knowledge, which have a ceiling; Snow and Matthews, 2016). However, the subskills of reading comprehension range in degree of constraint. Following the Reading Framework for the 2009 National Assessment of Educational Progress (National Assessment Governing Board, 2012), reading comprehension consists of three major components: students' ability to locate and recall key information, to integrate and interpret information to make meaning, and to critique and evaluate texts. In our view, the first of these skills — locate and recall — is relatively more constrained than the other two skills, which each require more integration of text and critical thinking for the student to make meaning from text.

We applied this definition of the subskills of reading comprehension and their relative degree of constraint in analyzing publicly available third-grade Massachusetts Comprehensive Assessment System (MCAS) ELA questions and answers. Each year from 2012 to 2014, the state of Massachusetts released a subset of third-grade MCAS ELA test items (17 items in 2012 and 18 items in 2013 and in 2014). We coded the released items into three categories, each tapping one of the three key components of reading comprehension described above. The first of these categories — students' ability to locate and recall key information — we considered "more constrained." The latter two we considered "unconstrained." We describe our coding of these two measures in greater detail in Appendix A.

Covariates

Using administrative records, we constructed a set of student-level covariates. We captured students' race/ethnicity using a set of dichotomous variables that identified whether a

student was Asian, black, Hispanic, white, or mixed/other. Similarly, we used a set of dichotomous variables to identify whether the students' home language was English only, English and Spanish, or English and another language. Using student birthdates, we calculated students' age as of September 1 in the year they were applying to prekindergarten. We also created dichotomous variables that identified whether the student was eligible for free or reduced-price lunch when he or she applied to prekindergarten, whether the student was male, and whether the student's country of birth was the United States.

Site Subgroup Characteristics (Potential Proxies for School Quality)

We drew on administrative records to create the following set of site characteristics that we used to predict variation in program impacts. Each of the following characteristics was merged onto students' records *by the year they competed in a lottery for the school* (so they do not overlap with the years students attended the school and are not affected by the students in the lottery sample). Correlations between these measures range from -0.04 to 0.80. The weakest association (r = -0.04) is between principal effectiveness and the number of applicants per seat, while the strongest association is between the percentage of low-income students in a school and the school's third-grade academic proficiency scores (r = 0.80). (Correlations between all site subgroup characteristics are presented in Appendix Table B.1.)

Demand for program. Within school choice settings, some hypothesize that program demand indicates program quality, while others posit that parents do not assess prekindergarten program quality well (Bassok et al., 2016). Using Round 1 school assignment data from the spring of 2007 through the spring of 2010, we constructed a measure of the number of applicants per available seat for each of the prekindergarten programs competed for by the study sample. Values for this measure among the study sample range from 1.76 (by definition, all programs in the study sample were oversubscribed) to 53.8. The 25th percentile value is 4.2, the 50th percentile value is 6.23, and the 75th percentile value is 8.8.

Average percentage proficient on third-grade math and ELA exams. The percentage of third-graders scoring at or above proficient on the state standardized ELA and mathematics exams provides a later look at the academic achievement level of the students in the school, a proxy for school quality. To compute this measure, we averaged each school's state-reported percentage proficient ELA and mathematics values for a given school year. Values for this measure among the study sample range from 1.0 to 84.5. The 25th percentile value is 30.5, the 50th percentile value is 44.5, and the 75th percentile value is 58.5.

Median school-level student growth percentile (math). In 2008, the state of Massachusetts began capturing student progress using a metric called the student growth percentile (SGP), which captures the yearly changes in a student's MCAS scores relative to the yearly changes of students with similar characteristics. As described by the state, "A student with a growth percentile of 90 in 5th grade mathematics grew as much [as] or more than 90 percent of her academic peers (students with similar score histories) from the 4th grade math MCAS to the 5th grade math MCAS" (Massachusetts Department of Elementary and Secondary Education,

2011). Beginning in 2008, the state began reporting median SGP scores for each school as an accountability metric meant to complement school-level average MCAS proficiency rankings, which do not take into account student growth or student peers. Values for this measure among the study sample range from 16.5 to 92.0. The 25th percentile value is 42.0, the 50th percentile value is 50.0, and the 75th percentile value is 58.0. As the math and ELA SGP scores were highly correlated for the schools in this sample (r = 0.91, p-value = 0.000), we chose to focus our analysis on schools' math SGP measures.[3]

Percentage of low-income students. The Massachusetts State Department of Education releases data annually on the percentage of students from low-income families within its schools. The state counts as low-income any student who (1) was eligible for free or reduced-price lunch, (2) received Transitional Aid to Families benefits, and/or (3) was eligible for the Supplemental Nutrition Assistance Program (SNAP). During the 2014-15 school year, the state changed its definition of "low-income" slightly. (See Weiland et al., in press, for details.) While this change does not affect our use of the measure as a moderator, in the treatment contrast comparisons when we measured a student's school experience in the 2014-15 school year, we used the school's previous (2013-14) low-income score. Values for this measure (the moderator) among the study sample range from 27.9 to 97.2. The 25th percentile value is 61.0, the 50th percentile value is 75.0, and the 75th percentile value is 80.2.

Average measure of school climate. The BPS school climate surveys were administered in the spring of each school year to students (Grades 3-11) and teachers (Grades K-12) in the 2009, 2010, and 2011 school years,[4] making these data available as moderators for students in cohorts 2 through 4. Approximately 53 percent of BPS teachers completed the survey and approximately 57.5 percent of all BPS students in Grades 3 to 11 completed the survey. The teacher and student surveys included a total of 94 items, organized by the district into 11 subscales. Psychometric work on this measure (Rochester, Weiland, Unterman, and McCormick, 2019) pointed to four relevant school climate dimensions (52 items from the teacher survey and 42 items from the student survey): positive emotional climate, student engagement, teacher effectiveness, and principal effectiveness. All items have the same four-point Likert scale (1 = strongly disagree to 4 = strongly agree). In the study sample, measures of student engagement and teacher effectiveness were highly correlated (Pearson correlation coefficient = 0.91, p-value = < 0.0001), so we have averaged these dimensions into one dimension, called teacher effectiveness and student engagement, for data reduction purposes. The correlations between all other dimensions range from 0.16 to 0.68 (results available upon request). For the measure of teacher effectiveness and student engagement, the 25th percentile value is 2.72, the 50th percentile value is 3.21, and the 75th percentile value is 3.34. For the measure of positive emotional climate, the percentile values are 2.80, 2.82, and 3.00, respectively. And for principal effectiveness, the percentile values are 3.24, 3.56, and 3.43, respectively.

[3]Results based on schools' ELA SGP measures were similar and are available upon request.

[4]Given the low rates of response from parents (13.5 percent), we used only the student and teacher survey responses in the present study.

Percentage of kindergarten peers who received BPS prekindergarten. Using BPS administrative records of BPS prekindergarten attendance, we calculated the percentage of kindergarten students who had attended the BPS prekindergarten program in the prior year. Values for this measure (the moderator) among the study sample range from 0 to 100 percent. The 25th percentile value is 28.87, the 50th percentile value is 50.77, and the 75th percentile value is 73.33.

School Context Measures (Measures of Students' School Experience)

The state of Massachusetts makes publicly available school-level data on every public school on an annual basis. We have merged these data sets with our students' files using the school identifier for the school the student is reported *being enrolled in for the longest time during his or her kindergarten, first-, second-, and third-grade school years*. These data capture the percentage of English language learners within the school, the percentage of students with disabilities within the school, the percentage of students who qualify for free or reduced-price lunch within the school, the racial make-up of the school's student body, the percentage of licensed teachers within the school, the teacher-to-student ratio, the percentage of teachers retained or remained working in the same position compared with the previous school year, the average class size, the percentage of teachers rated as exemplary or proficient in the state's rating system, and the percentage of students who remain in the school throughout the school year (stability rate). For each student, we averaged yearly values for each measure to capture his or her kindergarten-through-third-grade exposure to a given school characteristic. (See Weiland et al., in press, for greater detail.)

Data Analytic Plan

As mentioned above, during the study period, BPS relied on a district-wide school assignment process to place students in its prekindergarten programs. During this period, the program had more demand for prekindergarten seats than supply, particularly within some schools; seats were allocated via an algorithm within the assignment process that is used in several cities around the country. Effectively, for oversubscribed schools, the algorithm performed a "coin flip" to determine which children were offered a seat and which were not (explained in greater detail in Weiland et al., in press). We relied on these naturally occurring lotteries 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 and Unterman, 2014; Dobbie and 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 the treatment group of the analytic sample, and those who "lose" make up the 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. The basic approach for the analysis was thus to estimate, for each lottery, differences in mean outcomes for winners and control group members, and to average the results across lotteries.

Estimating the Distribution of Intent-to-Treat Effects Across Sites

As a first step in our analysis of impact variation, we quantified and illustrated the distribution of intent-to-treat (ITT) effects across sites using the framework set forth by Bloom, Raudenbush, Weiss, and Porter (2017) and applied by Weiss et al. (2017). In this method, 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, Raudenbush, Weiss, and Porter (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 (that is, compliance rate difference) was 0.29. The lottery-induced BPS prekindergarten enrollment rate differences did not vary statistically significantly across sites (τ = 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 et al. (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).[5] To estimate β and τ, we used the following two-level hierarchical linear model:

Level 1: Lottery Participants

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

Level 2: Sites

$$Bj=\beta+bj \quad (2)$$

where:

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

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

$bj \sim N(0, \tau 2)$

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

In this model, Y_{ij} is the value of the outcome measure for individual i in site j, $\alpha_r Lottery\_BLOCK_{rij}$ 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, X_{lij}, to improve the precision of parameter estimates. The model has a set of fixed random assignment block intercepts (α_r), which 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 blocks. The model allows for site-specific program-effect coefficients (Bj) that can differ randomly across sites. The Bj'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 $\sigma^2_{|X,Lottery\_Block}(T)$, which can be different for treatment group members and control group members.[6] 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 and Olkin, 1985).

Differences in ITT Impacts Across Sites (Moderator Analysis)

To estimate whether key site characteristics predict variation in impacts, 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:

$$Y_{ji} = \beta_0 T_{ji} + \beta_1 T_{ji} * C_{ji} + \sum_{j=1}^{J} \pi_j I_{ji} + \theta X_{ji} + \varepsilon_{ji} \quad (3)$$

where Y_{ji} is a relevant short- or medium-term outcome for student i; T is a lottery winner indicator equal to 1 if student i wins lottery j and zero otherwise; C_{ji} is the characteristic of interest for a given lottery (also called site); I is a vector of a lottery indicators equal to 1 for lottery j and zero otherwise; X_{ji} is a vector of student-level covariates (race/ethnicity, gender, eligibility for free or reduced-price lunch, age, country of origin, and home language status); and ε_{ji} is a random error for student i that is clustered by the prekindergarten school that students entered after their lottery.[7] To help interpret the findings, for each site characteristic,

[6]Bloom et al. (2017) provide further information about this model, and Raudenbush and Bloom (2015) explore its properties.

[7]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.

after reporting β_0 and β_1, we used the site characteristic's values (reported in 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.)

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 least 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 (Model 4) 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 the superior model 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. When there was a clear pattern in ITT effects by 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, and Bloom, 2005). The first stage was specified as:

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

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:

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

where $\hat{E}_{ji}$ equals the fitted value of the enrollment outcome from the first-stage equation, and e_i 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 et al. (in press), 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, 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.

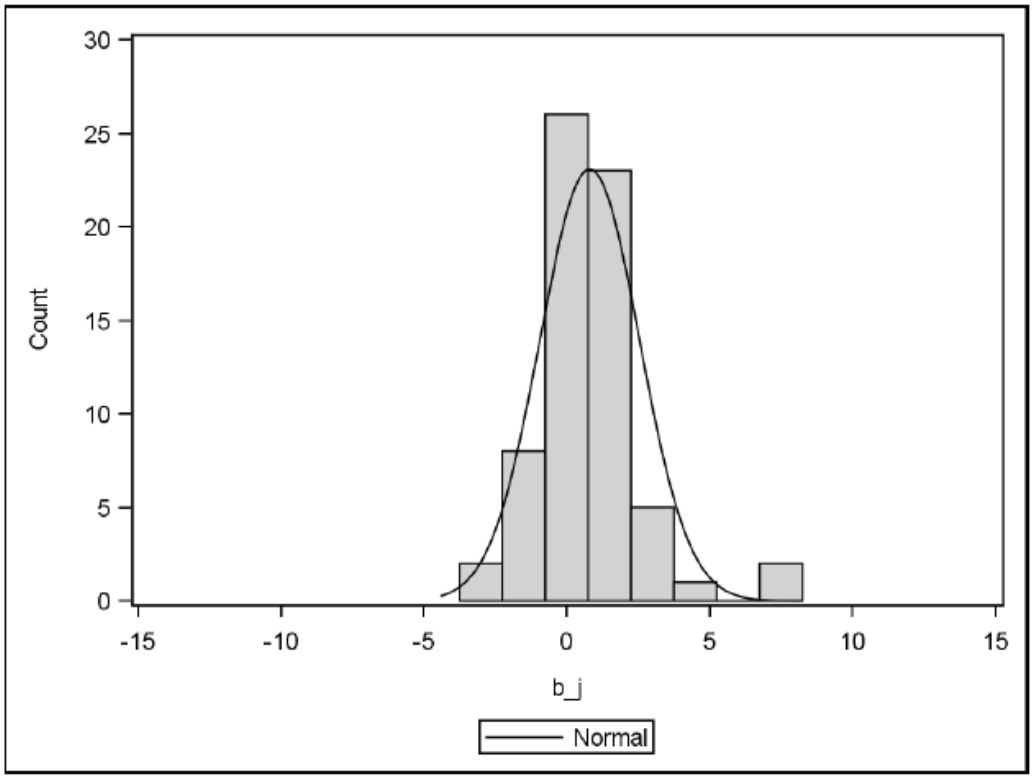
Figures 2 and 3 illustrate the distribution of treatment effects across BPS prekindergarten sites for two measures of students' academic progress: ever retained in grade and ever identified as special education.[8] The distribution of prekindergarten effects on students' probability of being retained 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 these effects capture 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. 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

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

Figure 2

Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Ever Retained



Estimated grand mean difference: 0.80, p = 0.395
Estimated tau = 1.72, p on Q-statistic = 0.043

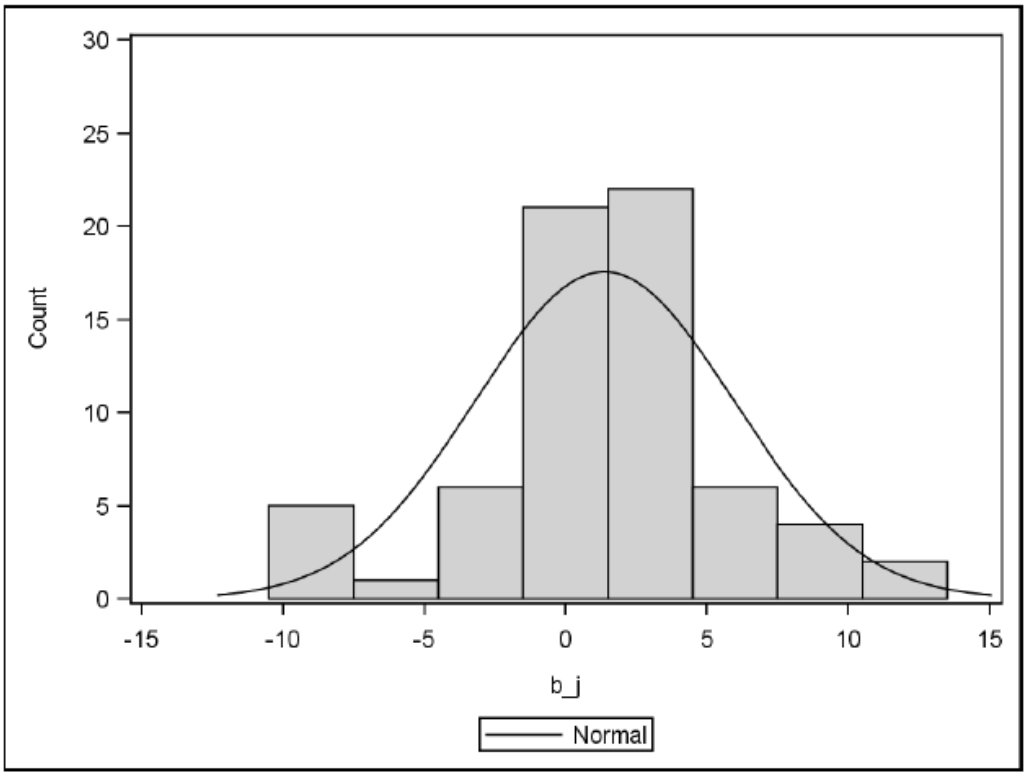
third- to fourth-grade reading gains are 0.36 standard deviations; thus, this range of effects is quite large and covers approximately two years of typical growth (Hill, Bloom, Black, and Lipsey, 2008).

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 $\widehat{\tau}$ was 0.001, p-value = 0.1870 for constrained and unconstrained outcomes, respectively) and the distributions could not be illustrated.

In an analysis of Head Start sites across the nation, Bloom and Weiland (2015) and Walters (2015) reported variation across sites of between 0.12 and 0.17 standard deviations. While the estimated effects reported here are slightly less than 0.15 standard deviations, they are

Figure 3

Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Ever Identified as Special Education



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

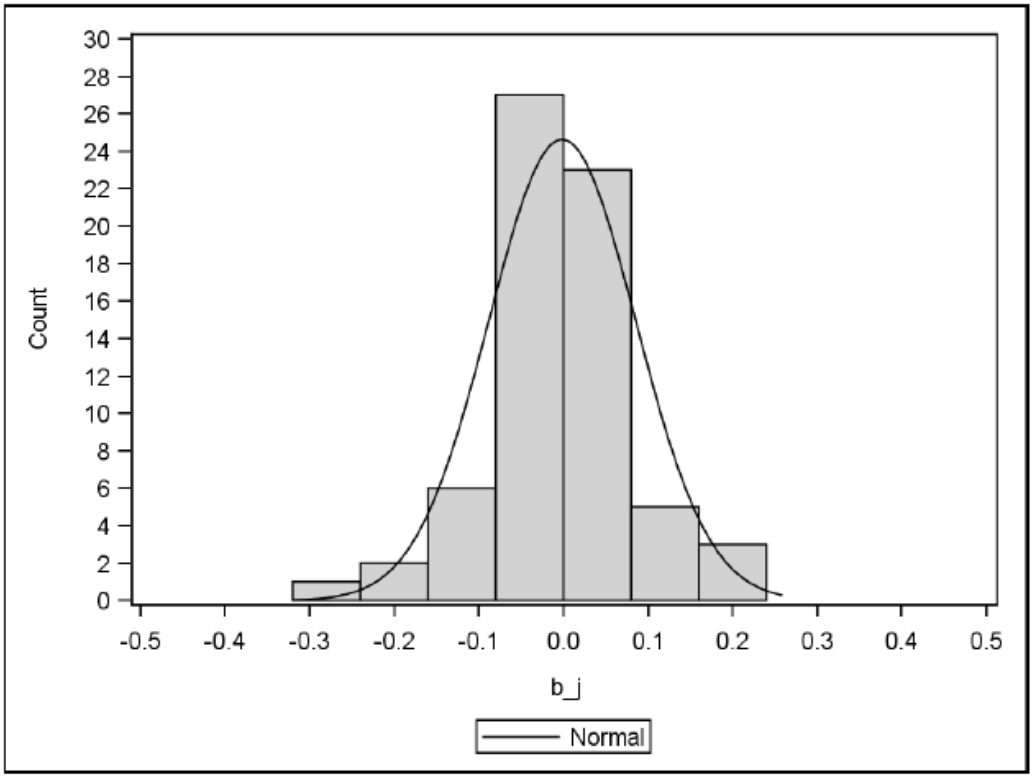
produced within one district by a central office implementing one curriculum, so one may expect a narrower distribution. (Variation in the counterfactual condition across sites may also contribute to variation in effects; this is discussed in greater detail in the next section.) Weiss et al. (2017) found 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 from 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.

Figure 4

Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — ELA



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

Demand for Program

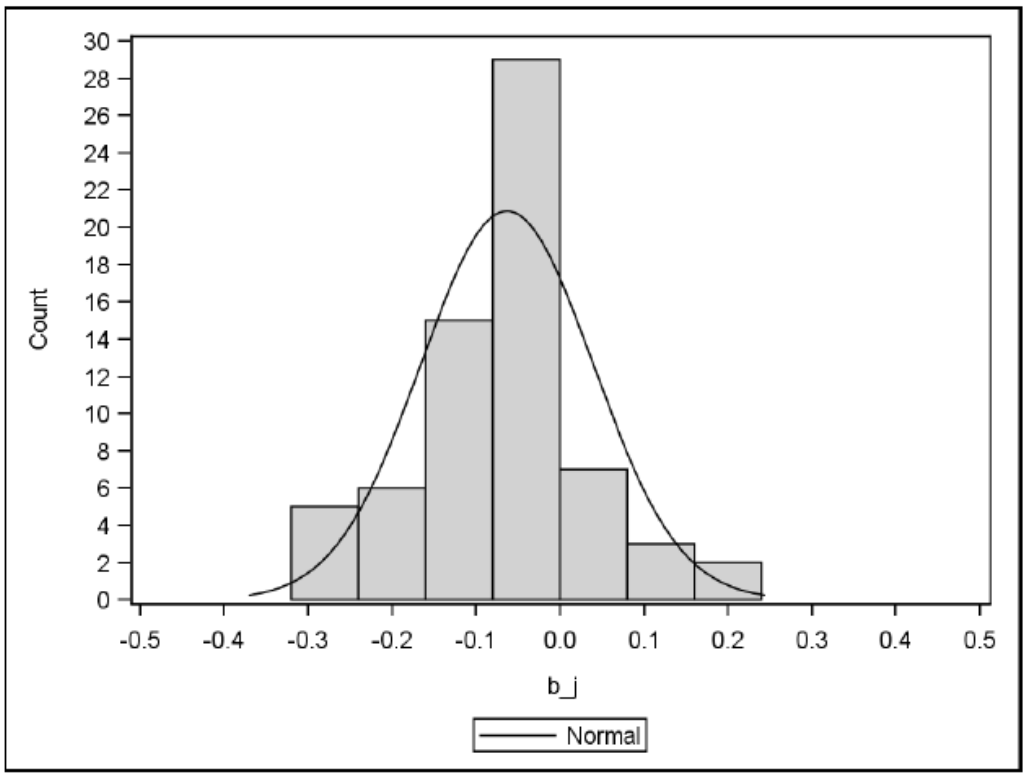
Demand for a given program was the only site characteristic for which a linear model was not appropriate. A visual inspection of the quintile-by-quintile treatment effects and a GLH test using the model fit statistics confirmed that a quadratic specification of demand (adding an interaction between demand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

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.[9] There was a statistically significant relationship between the demand for a site and treatment effects for one

[9]When discussing all findings in this section, we have focused on the estimated treatment effects for students who competed for schools at the 25th, 50th, and 75th percentiles, presented in the far right three columns of Table 2, as these findings have a clear (or clearer) substantive interpretation (compared with the estimated coefficients).

Figure 5

Histogram of Site-Level Constrained Empirical-Bayes Impact Estimates — Math



Estimated grand mean difference: -0.06, p = 0.1270
Estimated tau = 0.102, p on Q-statistic = 0.0324

outcome — the likelihood that a student will be classified as special education. 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 the other three 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 increases. Specifically, on average, in sites scoring at the 25th percentile, winning a lottery increased students' risk of

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 classified as 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 classified as 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 classified as 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 classified as 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

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

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 the typical third- to fourth-grade growth (Hill, Bloom, Black, and 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. 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 the small sample has less power, the fourth panel of Table 2 presents a similar, though weakened, association between this site characteristic and treatment effects. Specifically, on average, winning a lottery markedly increased students' risk of being classified as special education (ES = 10.40, $p = 0.041$), and the effect's interaction with each site's SGP level just missed the 0.05 standard of statistical significance ($p = 0.055$). Estimated

Table 3

Effects of Enrollment for 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 classified as special education	18.5	6.05	12.45	0.0722	15.42	23.47	-8.05	0.4516
English language arts	0.22	0.72	-0.5 *	0.0042	0.64	0.19	0.45 *	0.0292
Math	0.25	0.61	-0.36 *	0.034	0.76	0.1	0.66 *	0.0162
Sample size (all lottery participants)	285	645			235	544		

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

effects for the relationship between treatment effects and SGP levels for measures of students' academic achievement also appear similar to the pattern seen above but just missed conventional statistical significance levels. Though 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.

Percentage of Low-Income Students

While the percentage of low-income students within a school is not a proxy for school quality, it is an indicator of the additional resources students brought to the school 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 percentage 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 percentage of low-income students), winning a lottery had no effect on students' scores; in sites at the 75th percentile (having a high percentage 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

Using BPS administrative records on BPS prekindergarten participation, we constructed a measure of the percentage of students attending kindergarten at the site who enrolled in BPS prekindergarten. There was 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 will 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 — 1.43 percentage points. Both of these effects were relatively small. 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 — 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 were larger for schools with higher reports of teacher effectiveness and

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)	24.238	0.155	-8.178	0.162	1.337	1.175	-0.297
Ever retained in years 2-4 (%)	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 classified as 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 classified as 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.

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, these findings should be interpreted with caution.

Schools with High and Low Third-Grade Academic Proficiency Scores: Other Possible Explanations

School Context Analysis (Differences in Students' School Experience)

Table 5 returns to an analysis of *enrollment effects* for the students who competed for sites in the bottom quartile of schools and the top quartile of schools. (Model coefficients from the ITT model are presented in Appendix Table C.1.) While there were a few key lottery-induced student experience differences, overall there do not appear to be differences of a magnitude that could explain the effects reported in Table 3. For example, while lottery winner enrollees who competed for sites in the 25th percentile of the distribution enrolled in schools with very similar percentages of low-income students as their control group counterparts, lottery winner enrollees in the 75th percentile of the distribution 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.

Discussion

Although the BPS prekindergarten program was implemented in a district with policies and supports in place to promote consistent implementation across sites, there was statistically significant variation in program effects in our sample of students who competed in over-subscribed prekindergarten lotteries for nearly every key outcome, with effects that ranged from negative to positive. For example, the range of enrollment effects on students' academic achievement in ELA was relatively wide — 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). These findings are driven by either the quality of the BPS prekindergarten program and the students' subsequent elementary school experience or the services received by control group members in the absence of the program. After exploring the relationship between the quality of students' P-3 school experience and estimated effects, we found a moderate relationship between school quality, as measured by the percentage of students in the school scoring at or above proficient on third-grade state tests, and program effects. This relationship does not appear to be solely driven by higher-resource students self-selecting into high-test-score schools. An analysis of the measurable features of the lottery-induced differences in students' experiences across these sites indicates that prekindergarten program recipients in sites with

Table 5

Effects of Enrollment on Treatment Contrast for 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
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	-3.26 **	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	0.8 *	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	-1.5 **	0.146
Student stability (%)	87.23	86.5	0.73	0.011	93	90.73	2.27 *	< 0.0001
Sample size (all lottery participants)	285	645			235	544		

NOTES: * P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

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 this difference does not appear large enough to be the sole explanation for the effects.

With very little treatment contrast in K-3 settings, one might expect no effect of the program by third grade, but negative effects prompt questions about what lottery losers received during the prekindergarten year. Our project team has preregistration data from the school assignment process for two cohorts of students (2007 and 2008). In these data, 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 these limited data shows no evidence of variation by setting type — while 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.

It is possible that when prekindergarten programs are nested within low-performing elementary schools, they also struggle to positively affect students. Logically, if an elementary school is struggling to implement a high-quality K-3 curriculum, it may not have resources to help the new prekindergarten program on its campus get off the ground. It may be the case that new prekindergarten programs located in low-quality elementary schools need additional resources and professional development support to make up for what other prekindergarten programs receive from the professional community on their school campus. While the retrospective nature of this study prevented us from collecting systematic data on the full supports different prekindergarten programs received, in our future work we will attempt to systematically explore whether features of prekindergarten program implementation are related to their effects on students.

Finally, recent work on the sustaining environments hypothesis stresses the importance of aligned instruction and content across students' preschool and early elementary experiences. For example, a recent randomized controlled trial study of a preschool math curriculum found that the early prekindergarten math effects were sustained only when students received an aligned kindergarten and first-grade math curriculum (Jenkins et al., 2017). It may be that high-quality BPS schools fostered communication between their prekindergarten staff and early elementary school staff and naturally aligned their curricula to benefit students.

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 oversubscribed schools, or about 25 percent of all applicants to the program, and a propensity scores analysis on the full applicant sample found prekindergarten enrollment was associated with benefits in K-3 on all examined outcomes (Weiland et al., in press). As such, it is difficult to gauge the external validity of the current findings. Third, we are limited by the measurement of both the 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, 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 the case because these prekindergarten sites produced higher impacts or whether the students' K-3 experience did a better job of sustaining them.

Taken together, our exploratory results suggest that the quality of a student's early elementary school experience is an important piece of sustaining the prekindergarten boost. Descriptive statistics show that the post-prekindergarten schooling environments 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 reading (Massachusetts Department of Elementary and Secondary Education, n.d.). As others have suggested (Bailey, Duncan, Odgers, and Yu, 2017; Phillips, Gormley, and Anderson, 2016), sustaining the gains from prekindergarten may require investment in improving the quality of children's K-3 experiences and in aligning children's P-3 experiences so that prekindergarten attenders do not simply repeat the same material in kindergarten — work that Boston has already begun (Boston Public Schools, 2017).

Appendix A

Constrained and Unconstrained English Language Arts Measures

There is a consensus among literacy experts that reading comprehension is an unconstrained skill — that is, there is always room for improvement — versus more constrained skills like letter knowledge, where there is a ceiling (Snow and Matthews, 2016). However, the subskills of reading comprehension range in degree of constraint; following the Reading Framework for the 2009 National Assessment of Educational Progress (National Assessment Governing Board, 2008), reading comprehension consists of three major components: students' ability to locate and recall key information, to integrate and interpret information to make meaning, and to critique and evaluate texts. In our view, the first of these skills — locate and recall — is relatively more constrained than the other two skills, which each require more integration of text and critical thinking for the student to make meaning from text and which we consider "unconstrained."

We applied this definition of the subskills of reading comprehension and their relative degree of constraint using publicly available third-grade Massachusetts Comprehensive Assessment System (MCAS) English language arts (ELA) questions and answers. Each year from 2012 to 2014, the state of Massachusetts released a subset of third-grade MCAS ELA test items (17 items in 2012, and 18 items in 2013 and in 2014). Specifically, we coded the released items into three categories, each tapping one of the key components of reading comprehension delineated above.

Our item coding process had two steps. First, an advanced Ph.D. student specializing in language and literacy development among children 0 to 8 years of age coded MCAS items released by the state for the 2015 school year (that is, a non-analytic year) to develop the coding schema. Second, two Ph.D. students applied the schema to the 2014 items, calculated their inter-rater reliability (percentage agreement and kappa), reviewed coding disagreements, and resolved them to create final codes. This second step was then repeated for the 2013 and 2012 items. Percentage agreement was between 88 percent and 94 percent, and kappa was between 0.74 and 0.91 across the three years (as shown in Appendix Table A.1). Item types and classification coding by year are available in Appendix Tables A.2 through A.4.

Ultimately, we created simple unit-weighted averages of each student's total correct items, separately for "more constrained" and "unconstrained." Notably, we did not code the PARCC test taken by most of cohort 4 using this same schema because we did not want to conflate test content/construction differences with differences in skill types. Also, we coded only for ELA and not math. Massachusetts also releases mathematics MCAS items each year. However, procedural (that is, constrained) and conceptual (that is, unconstrained) knowledge in math are intertwined (Rittle-Johnson and Schneider, 2015) to a greater degree than in the literacy domain.

Appendix Table A.1

Inter-Rater Reliability on Coding of Released MCAS ELA Items, 2012-2014

	Classification	
Year	Simple Agreement	Kappa
2012	88.2%	0.810
2013	94.4%	0.743
2014	94.4%	0.909

Appendix Table A.2

Item Type and Classification Coding for Released MCAS ELA Items, 2012

Question Number	Item Classification
1	Locate and recall
2	Locate and recall
3	Integrate and interpret
4	Integrate and interpret
5	Integrate and interpret
6	Locate and recall
7	Integrate and interpret
8	Locate and recall
9	Integrate and interpret
10	Locate and recall
11	Critique and evaluate
12	Critique and evaluate
13	Integrate and interpret
14	Locate and recall
15	Critique and evaluate
16	Integrate and interpret
17	Integrate and interpret

Appendix Table A.3

Item Type and Classification Coding for Released MCAS ELA Items, 2013

Question Number	Item Classification
1	Locate and recall
2	Integrate and interpret
3	Locate and recall
4	Locate and recall
5	Integrate and interpret
6	Critique and evaluate
7	Locate and recall
8	Critique and evaluate
9	Integrate and interpret
10	Locate and recall
11	Integrate and interpret
12	Integrate and interpret
13	Integrate and interpret
14	Critique and evaluate
15	Critique and evaluate
16	Locate and recall
17	Locate and recall
18	Integrate and interpret

Appendix Table A.4

Item Type and Classification Coding for Released MCAS ELA Items, 2014

Question Number	Item Classification
1	Locate and recall
2	Integrate and interpret
3	Locate and recall
4	Locate and recall
5	Integrate and interpret
6	Locate and recall
7	Integrate and interpret
8	Critique and evaluate
9	Integrate and interpret
10	Locate and recall
11	Integrate and interpret
12	Integrate and interpret
13	Integrate and interpret
14	Integrate and interpret
15	Locate and recall
16	Locate and recall
17	Locate and recall
18	Critique and evaluate

Appendix B

Pearson Correlation Coefficients

Appendix Table B.1

Correlations Between School-Level Predictors of Variation in Impacts Across Schools

Variable	Low-Income Students (%)	Median School-Level Student Growth Percentile	Demand for Program	Average % Proficient on 3rd-Grade Math and ELA Exams	Teacher Effectiveness and Student Engagement	Principal Effectiveness	Positive Emotional Climate
Low-income students (%)	1						
Median school-level student growth percentile	0.217*** < 0.0001	1					
Demand for program	-0.180*** < 0.0001	-0.045* 0.0431	1				
Average % proficient on 3rd-grade math and ELA exams	-0.810*** < 0.0001	-0.126*** < 0.0001	0.327*** < 0.0001	1			
Teacher effectiveness and student engagement	0.154*** < 0.0001	0.0603 0.0118	-0.088*** < 0.0001	0.069*** 0.0005	1		
Principal effectiveness	0.187*** < 0.0001	0.104*** < 0.0001	-0.025 0.1968	0.007 0.7408	0.664*** < 0.0001	1	
Positive emotional climate	-0.356**** < 0.0001	-0.093*** < 0.0001	0.070*** 0.0003	0.279*** < 0.0001	0.418*** < 0.0001	0.163*** < 0.0001	1

NOTES: *** P < 0.001. *p < 0.05.

Appendix C

Average Third-Grade Academic Proficiency Treatment Contrast

Appendix Table C.1

Average Third-Grade Academic Proficiency Treatment Contrast

Outcome	Coeff. on Treatment	P-Value	Coeff. on Site x Treatment Interaction	P-Value
English language learners (%)	4.516 *	0.001	-0.093 **	0.000
Students with disabilities (%)	0.472	0.318	-0.005	0.665
Low-income (%)	4.701 *	0.002	0.151 **	< 0.0001
African-American (%)	-0.791	0.561	-0.066 *	0.015
Asian (%)	-1.396	0.036	0.030 *	0.044
Hispanic (%)	8.764 **	< 0.0001	-0.173 **	< 0.0001
White (%)	-6.066 **	0.000	0.204 **	< 0.0001
Licensed to teach (%)	2.391 **	0.000	-0.029 *	0.027
Teacher:student ratio	-0.075	0.596	0.004	0.254
Teacher retained (%)	-1.754 *	0.011	0.056 **	< 0.0001
Average class size (N)	0.243	0.525	-0.005	0.540
Average teachers proficient (%)	-1.466	0.269	0.043	0.132
Average teachers exemplary (%)	1.198	0.347	-0.027	0.327
Proficient in 3rd grade ELA (%)	-7.125 **	< 0.0001	0.189 **	< 0.0001
Proficient in 3rd grade math (%)	-6.700 **	< 0.0001	0.190 **	< 0.0001
Student stability (%)	-1.146 *	0.027	0.049	< 0.0001

NOTES: * P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

References

Abdulkadiroğlu, A., Angrist, A., Dynarski, S., Kane, T., and Pathak, P. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics*, *126*, 649-748.

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.* NCEE 2011-4001. National Center for Education Evaluation and Regional Assistance.

Ansari, A., and Pianta, R. C. (2018). Variation in the long-term benefits of child care: The role of classroom quality in elementary school. *Developmental Psychology, 54*(10), 1854-1867.

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

Bailey, D., Duncan, G. J., Odgers, C. L., and Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7-39.

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

Bassok, D., Finch, J. E., Lee, R., Reardon, S. F., and Waldfogel, J. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *AERA Open*, *2*(3). Retrieved from https://doi.org/10.1177/2332858416653924.

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

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., and 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.

Bloom, H. S., and 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., and 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*. MDRC working paper. New York: MDRC.

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., and Yoshikawa, H. (2017). *Cradle to kindergarten: A new plan to combat inequality*. New York: Russell Sage.

Clements, D. H., and Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*, 136-163.

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

Clements, D. H., Sarama, J., Wolfe, C. B., and Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812-850. Retrieved from https://doi.org/10.3102/0002831212469270.

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., and Fryer, R. G., Jr. (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., and Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*, 109-132.

Garces, E., Thomas, D., and Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review*, *92*(4), 999-1012.

Gatti, G. G., and Petrochenkov, K. (2010). *Pearson Reading Street efficacy study 2009-10 final report*. Retrieved from https://www.pearsoned.com/wp-content/uploads/reading-street-efficacy-study-2009-2010_final.pdf.

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

Hedges, L., and Olkin, I. (1985). *Statistical models for meta-analysis*. Orlando, FL: Academic Press.

Henry, G. T., and Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review*, *26*, 100-112.

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

Jenkins, J., Watts, T. W., Magnuson, K., Gershoff, E., Clements, D. H., Sarama, J., and Duncan, G. J. (2018). Do high-quality kindergarten and first grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness, 11*(3), 339-374.

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

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

Kelly, D., Nord, C. W., Jenkins, F., Chan, J. Y., and Kastberg, D. (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. Washington, DC: National Bureau of Economic Research.

Ladnier-Hicks, J., McNeese, R. M., and Johnson, J. T. (2010). Third grade reading performance and teacher perceptions of the Scott Foresman Reading Street program in Title I schools in South Mobile County. *Journal of Curriculum and Instruction*, *4*(2), 51-70.

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., and Morris, P. (2018). *Strengthening children's math skills with enhanced instruction: The impacts of Making Pre-K Count and High 5s on kindergarten outcomes.* New York: MDRC.

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

National Assessment Governing Board. (2008). *Reading framework for the 2009 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Neidell, M., and Waldfogel, J. (2010). Cognitive and noncognitive peer effects in early education. *The Review of Economics and Statistics*, *92*(3), 562-576.

Phillips, D., Gormley, W., and Anderson, S. (2016). The effects of Tulsa's CAP Head Start program on middle-school academic outcomes and progress. *Developmental Psychology*, *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., and Jenkins, F. (2012). Highlights from TIMSS 2011: Mathematics and science achievement of US fourth-and eighth-grade students in an international context. NCES 2013-009. Washington, DC: National Center for Education Statistics.

Raudenbush, S. W., and Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475-499.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rittle-Johnson, B., and Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. *Oxford handbook of numerical cognition*, 1118-1134.

Rochester, S. E., Weiland, C., Unterman, R., McCormick, M., and Moffett, L. (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. *Early Childhood Research Quarterly*, *48*, 84-97.

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

Snow, C. E., and Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children*, *26*(2), 57-74.

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

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, *7*(4), 76-102.

Weiland, C. (2016). Launching preschool 2.0: A road map to high-quality public programs at scale. *Behavioral Science and Policy*, *2*(1), 37-46.

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

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

Weiland, C., and Yoshikawa, H. (2013). The impacts of an urban public prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills: Evidence from Boston. *Child Development*, *84*(6), 2112-2130.

Weiland, C., and Yoshikawa, H. (2014). Does peer socio-economic status predict children's gains in receptive vocabulary and executive function in prekindergarten? *Journal of Applied Developmental Psychology*, *35*, 422-432.

Weiss, M. J., Bloom, H. S., Savitz, N. V., Gupta, H., Vigil, A., and Cullinan, D. (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*, *10*(4), 843-876. Retrieved from https://doi.org/10.1080/19345747.2017.1300719.

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., and Herman, T. (2006). *An efficacy study on Scott Foresman's Reading Street program: Year one report*. Austin, TX: Magnolia Consulting.

Yoshikawa, H., Weiland, C., and Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, *26*(2), 21-35.

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

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

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York; Oakland, California; Washington, DC; and Los Angeles, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff members bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-prisoners, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.