Appendix A

Additional Information on Measures

Student Outcomes

Third-grade ELA and mathematics achievement. For third-grade reading and mathematics analyses, we used students' statewide mathematics and English Language Arts (ELA) standardized tests. 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 (see Weiland et al., 2019).

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

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

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

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 &

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."

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.

Predictors of Impact Variation

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. 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. Notably, in our focal years, Boston had relatively weak third grade performance, scoring in the bottom 11% of districts on the state third grade math test and the bottom 5% of districts for third grade reading (Massachusetts Department of Elementary and Secondary Education, 2014).

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). The state reported 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. These data were available for students in cohorts 2 through 4.

Proportion of low-income students. The Massachusetts State Department of Education releases data annually on the proportion of students from low-income families within its schools. During the 2014-15 school year, the state changed its definition of "low-income" slightly. For one of our school context variables – percentage low-income – the MA Department of Elementary and Secondary Education changed its definition in 2014-2015 (our last study) year. The prior definition counted as low-income any student who: 1) was eligible for free or reduced price lunch, 2) received Transitional Aid to Families benefits, and/or 3) was eligible for Supplemental Nutrition Assistance Program (SNAP). The updated measure added to this list students' foster care program status and Medicaid status (called MassHealth; MA Department of Elementary and Secondary Education, n.d.). To keep the definition consistent in all years, we used schools' previous low income score (2013-2014) as our measure of schools' 2014-2015 low-income status. 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 measures 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,[1] 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,

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

measures of student engagement and teacher effectiveness were highly correlated (Pearson correlation coefficient = 0.91, *p*-value = < 0.0001), so we 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.

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.

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

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

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

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

Internal Validity of the Analytic Sample

Table B.1

Balance on observables in the first-choice lottery sample

	Lottery winners	Control group	Estimated difference	P-value
Race/ethnicity (%)				
Hispanic	35.22	39.90	-4.68**	0.003
Black	25.00	23.35	1.65	0.271
White	26.73	24.34	2.39	0.144
Asian	10.13	7.28	2.85**	0.005
Other	2.92	4.14	-1.22	0.146
Male (%)	50.27	46.95	3.32	0.126
Eligible for free/reduced lunch (%)	57.66	56.68	0.98	0.604
Age	4.51	4.53	-1.97	0.117
Country of origin USA (%)	94.89	94.53	0.36	0.701
Home language (%)				
English	52.18	55.06	-2.88	0.133
Spanish	25.18	25.19	-0.01	0.994
Other	22.64	19.75	2.89	0.074
N children	1,101	2,081		

Note. There was a small amount of missing data on all baseline characteristics except age: 12 children (0.4%) were missing race/ethnicity and male information, 34 (1.1%) were missing male and free/reduced lunch information, 113 (4.2%) were missing country of origin information, and 5 (0.2%) were missing home language information. Means in the table were computed using non-missing data. Values for first choice lottery winners are the simple means for each requisite group. Values for the difference between lottery winners and control group members are obtained from a regression of a given baseline characteristic on a series of indicator variables that identify each lottery plus an indicator variable that equals 1 for lottery winners and 0 for lottery losers. The coefficient on lottery indicator equals the difference in the mean baseline characteristic between lottery winners and control group members, respectively. The value for control group members equals the corresponding value for lottery winners minus the estimated difference between lottery winners and control group members. A two-tailed t-test was applied to the estimated differences. An F-test was used to assess the statistical significance of the overall difference between lottery winners and control group members reflected by the full set of baseline characteristics in the table. The resulting F value is not statistically significant ($p = 0.2004$).
* P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

Table B.2

Percent of missing data on key outcomes

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

Note: ELA=English Language Arts.
* P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

Table B.3

Balance on baseline characteristics for students with test score data

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

Note: Nine students were missing free/reduced price lunch information; all other data was available for all students. Values for lottery winners are the simple means for each requisite group. Values for the difference between lottery winners and control group members are obtained from a regression of a given baseline characteristic on a series of indicator variables that identify each lottery plus an indicator variable that equals 1 for lottery winners and 0 for lottery losers. The coefficient on the lottery indicator equals the difference in the mean baseline characteristic between lottery winners and control group members. The value for control group members equals the corresponding value for lottery winners minus the estimated difference between lottery winners and control group members. A two-tailed t-test was applied to the estimated difference. An F-test was used to assess the statistical significance of the overall difference between lottery winners and control group members reflected by the full set of baseline characteristics in the table. The resulting F value was not statistically significant (p=.231).
* P-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

Appendix C

Pearson Correlation Coefficients

Table C.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-value < 0.05 for impact estimates. ** P-value < 0.01 for impact estimates.

Appendix D

Average Third-Grade Academic Proficiency Treatment Contrast

Table D.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.