Impacts of Five Expeditionary Learning Middle Schools on Academic Achievement

July 8, 2013

Ira Nichols-Barrer
Joshua Haimson

MATHEMATICA
Policy Research

Mathematica Reference Number:
40207.400

Submitted to:
Expeditionary Learning
247 West 35th Street, 8th Floor
New York, NY 10001
Project Officers: Tom Van Winkle, Mindy Spencer

Submitted by:
Mathematica Policy Research
955 Massachusetts Avenue
Suite 801
Cambridge, MA 02139
Telephone: (617) 491-7900
Facsimile: (617) 491-8044
Project Director: Ira Nichols-Barrer

Impacts of Five Expeditionary Learning Middle Schools on Academic Achievement

July 8, 2013

Ira Nichols-Barrer
Joshua Haimson

MATHEMATICA
Policy Research

ACKNOWLEDGMENTS

This report was made possible with the support of numerous individuals and organizations. First and foremost, we would like to acknowledge the New York City Department of Education and the District of Columbia Office of the State Superintendent of Education for generously making their data available to our research team and providing assistance and guidance to the study.

The study would not have been possible without contributions from several other individuals at Mathematica. In particular, Clare Wolfendale provided indispensible programming assistance during the study's data cleaning process, and Phil Gleason, Brian Gill, and Josh Furgeson all provided technical guidance on methodological issues related to the study's analyses. Autumn Parker led the formatting and production of the report

Finally, the study and this report benefited greatly from input at various stages from Mindy Spencer, Tom Van Winkle, and Scott Hartl of Expeditionary Learning. We thank all of the reviewers who commented on the report for their thoughtful responses and suggestions.

This page has been left blank for double-sided copying.

CONTENTS

This page has been left blank for double-sided copying.

TABLES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

Expeditionary Learning (EL) is a growing provider of curriculum and professional development services to teachers and school leaders. The EL model combines an interdisciplinary instructional approach with ongoing training and coaching for teachers and school leaders. The EL curriculum uses an experiential approach in which students conduct research projects to share with outside audiences. Learning expeditions—case-studies of academic topics—often bring together teachers from different subjects to coordinate shared projects; this curriculum includes several elements that are closely aligned with the Common Core standards for English-language arts and literacy. As of the 2010–2011 school year, EL's network included a total of 161 schools in 30 states.

This report presents findings from the first rigorous study of the impacts of EL schools. This research aims to use the best available quasi-experimental methods to estimate the impacts of five urban EL middle schools on students' reading and math test scores. Using the study's data on student characteristics, the report also provides additional descriptive information on the types of students who enroll in EL schools. The study's key results include the following findings.

Compared with local district schools, these EL middle schools enroll an elevated percentage of Hispanics and English-language learners. However, EL students are similar to the local student population with respect to their special-education status, eligibility for free or reduced-price meals, and prior achievement. To account for these differences in the impact analysis, we matched each EL student to a local comparison student with similar characteristics, and then compared the achievement outcomes of EL students with this matched comparison group.

The impact analysis indicates that the five EL middle schools have a positive and statistically significant impact on student achievement in reading and math. After students have been enrolled in EL for one year, we estimate that these EL schools have a positive and statistically significant impact of 0.06 standard deviations on reading test scores and an impact on math test scores (–0.02 standard deviations) that is statistically indistinguishable from zero. After two years, students experience significant and positive cumulative impacts in both subjects (0.11 standard deviations in reading and 0.09 standard deviations in math). Impact estimates remain positive and significant after three years, with effects of 0.16 in reading and 0.29 in math.

The magnitude of these impact estimates suggests that these EL schools are substantially increasing student achievement. Relative to a normal test distribution, these cumulative impacts are equivalent to moving a student from the 50th percentile to the 56th percentile in reading and to the 61st percentile in math after three years. Compared with national norms for middle-school learning growth, our results suggest that EL students experience impacts that are large enough to accumulate about an extra seven months of learning growth in reading and 10 months of extra learning growth in math after three years (Hill et al. 2008).

While we find positive and meaningful impacts at these five middle schools, further research is needed to more fully understand the effectiveness of the EL model. Future studies should include a larger sample of EL schools and examine whether EL schools impact other student outcomes beyond reading and math test scores. In addition, researchers should identify which components of EL's multifaceted approach are most strongly associated with achievement impacts. This promising line of research could provide important insights for school leaders, teachers, and policymakers.

This page has been left blank for double-sided copying.

I. INTRODUCTION

Expeditionary Learning (EL) is a rapidly growing provider of curriculum and professional development services to teachers and school leaders. Since 1993, when EL began implementing its approach in an initial set of demonstration schools, the model has spread to schools located throughout the United States. As of the 2010–2011 school year, EL's "whole school" reform network included a total of 161 schools in 30 states.

The EL model combines a highly detailed, interdisciplinary curriculum with training and coaching for teachers and school leaders.[1] The approach includes the following five dimensions:

1. Curriculum with learning "expeditions" that offer multidisciplinary, long-term explorations of issues or topics involving a combination of projects, fieldwork, and culminating performances. The EL curriculum includes several elements that are closely aligned with the Common Core standards for English-language arts and literacy.
2. Instructional methods that emphasize student interaction, critical thinking, and collaboration.
3. A focus on building a school culture that emphasizes quality work, student character, and citizenship.
4. Frequent student assessment against learning targets using achievement data.
5. Supports for focusing school leadership on student achievement, the use of assessment and other data, and shaping school structures to student needs.

To support implementation of this model, EL provides a combination of curriculum resources, professional development institutes in the summer and during the school year, and on-site classroom observation and coaching for teachers and school leaders.

Past descriptive studies of EL middle schools have found some potentially positive results, but the empirical methods used in these studies were not rigorous. Most recently, in 2011, researchers at the University of Massachusetts Donahue Institute completed a descriptive study of a single EL middle school in Rochester, New York, that compared the achievement of EL students with achievement at a set of eight district middle schools. The researchers found positive and significant potential effects in reading and statistically insignificant effects in math. However, the study did not conduct student-level matching or otherwise demonstrate that the comparison group of district students was similar to EL students, suggesting that the results could have been biased by differences between the two groups. Similarly, researchers at Mountain Measurement (2010) completed regression analyses comparing achievement growth at 24 EL schools with a group of comparison schools; the study found potentially positive impacts on reading and math test scores at mature EL schools (those that had been operating within the EL network for three or more years), but the analysis did not control for students' baseline achievement prior to receiving EL. In addition, in an older meta-analysis that included nine descriptive studies of student achievement at EL schools, Borman et al. (2001) found that these prior studies reported average potential EL effects of

[1] For a more detailed description of the EL model and the organization's whole-school reform principles and practices, see Expeditionary Learning (2011).

approximately 0.19 standard deviations (averaging across all subjects and outcomes). But this earlier research had serious methodological weaknesses. Several of the studies did not include a comparison group but instead compared achievement outcomes within EL schools before and after EL adoption. And among the studies that did include a comparison group, the research was not conducted using student-level data, meaning that the researchers could not account for differences between individual student characteristics or students' achievement prior to enrolling in EL schools.

This study aims to use the best available quasi-experimental methods to estimate the impacts of five EL middle schools on students' reading and math test scores, examining achievement outcomes up to three years after students enter EL. This report seeks to answer the following research question: Do EL services have an impact on student achievement outcomes? **Specifically, do students attending EL schools perform better in reading and math than they would have performed in other public schools?**

To answer this question, this study carefully matched each EL student to a comparison student with similar characteristics who attended a local district school, and then compared the achievement outcomes of EL students with this matched comparison group. Using the study's data on student characteristics, this report also provides additional descriptive information on the types of students who enroll in EL schools and explores whether the EL approach is more effective for particular subgroups of enrolled students in the sample. The findings from this study represent the most rigorous examination of EL impacts conducted to date.

Below, in Section II, we discuss the sample of EL schools included in the study and compare the characteristics of students attending these EL schools with the characteristics of other students in local district schools. Section III summarizes the study's empirical approach and presents our main findings regarding the estimated impacts of EL schools. Section IV concludes by outlining an agenda for future research related to the effectiveness of the EL curriculum and EL-related practices.

II. SAMPLE SELECTION AND DESCRIPTION

This study uses longitudinally linked, student-level data collected from two urban school districts: New York City and Washington, DC.[2] Within these districts, the study examined all EL schools that met our research selection criteria. First, we required the study schools to have been founded within EL's whole-school reform network: each of these schools was created with technical support from the EL organization, including interdisciplinary curriculum development as well as professional training and coaching services for teachers and school leaders.[3] Second, the sample was limited to schools that were founded in the 2010–2011 school year or earlier, to guarantee that at least two entering cohorts of EL students could be observed at each treatment site.[4] Lastly, the study included only EL schools that enroll new students in middle-school grades (grade 6 through grade 8), to ensure that we could observe multiple years of baseline (i.e., before EL enrollment) data for the students in the analysis. There were a total of five EL middle schools that met these selection criteria in the two study districts (for more details regarding the available data obtained for each EL school, see Appendix A).

Because the study sample is limited to five EL schools (out of the 62 middle schools[5] affiliated with EL during the study period), we used administrative records provided by EL to examine the characteristics of these study schools in relation to the national EL network (the records were current as of the 2010–2011 school year). As shown in Table II.1, the study schools have a similar average enrollment size (343 students) to the average in the EL network (357 students), and the average amount of time affiliated with EL in the study schools (5 years) is only slightly below the network's average (6 years). Similarly, the study sample includes a mix of charter schools (40 percent) and traditional public schools (60 percent) that is similar to the proportions found in EL's national network.

However, there are also several large differences between the study sample and the overall EL network. Perhaps most important, all of the study schools are located in urban areas, whereas only 53 percent of EL's national network is urban. Compared with the national network, the EL study schools are also more likely to enroll Hispanic students and less likely to enroll white students. The percentage of students eligible for free or reduced-price meals at study schools (63 percent) is also larger than the percentage in the EL network as a whole (48 percent), but the difference between these two measures is not statistically significant. In light of these differences between the study

[2] These two districts represent a convenience sample—Mathematica chose to make data requests in these two districts because both jurisdictions had shared data with the research team for similar projects in the recent past.

[3] Some charter schools and traditional public schools also join the EL network after operating independently for many years: in such cases, it would be more complex to empirically disentangle the impact of EL from the underlying effectiveness of the school before EL was introduced. In other words, because the five study schools were *founded* in partnership with EL, estimating school-level impacts on student achievement provides a direct means of testing the effectiveness of the EL model.

[4] Data provided by New York City included the 2011–2012 school year. However, the data provided by Washington, DC, ended in 2010–2011. Thus, in Washington, DC, the study could include only EL middle schools that were founded in the 2009–2010 school year or earlier. More details regarding the data provided by each district can be found in Appendix A.

[5] For the purpose of this analysis, we define middle schools to include all schools that serve students in grade 6 through grade 8. This includes schools that also serve additional grades (such as K-8 and K-12 schools).

schools and the overall EL network, it is not clear whether the study findings discussed below are fully representative of the results found in other EL middle schools located throughout the country.

Table II.1. Characteristics of All EL Middle Schools and EL Study Schools

	All EL Middle Schools	Study Schools
Total Enrollment (mean)	356.9	343.2
Number of Years Affiliated with EL (mean)	6.0	4.6
Located in Urban Area (percentage)	53.2	100.0*
Charter School (percentage)	43.5	40.0
Average Student Characteristics (Mean Percentage)		
Black	19.2	21.3
Hispanic	16.3	46.4*
White	53.8	24.7*
Free or Reduced-Price Meals	48.2	63.2
Special Education	14.3	17.4
English-Language Learner	9.2	11.2
Number of Schools	62	5

Source: EL Administrative Records

* Difference from the EL network is statistically significant at the .05 level, two-tailed test.

In addition to examining how the study schools compare with EL's national network, we also explored whether the students enrolling in study schools tend to differ from the students attending local district schools that are not affiliated with EL. Specifically, we used student-level records data to examine the baseline characteristics of the students who later attended one of the five EL schools in the study sample and compared the results with students who attended one of EL's feeder elementary schools or the district as a whole.[6] These analyses provide a way to test whether EL schools tend to attract different types of students than other district schools. The results are shown in Table II.2.

Key findings from this descriptive analysis include the following observations:

- **Compared with the local district population, EL students in the five sample schools are more likely to be Hispanic and more likely to be English-language learners.** Half of the EL students in the sample are Hispanic, whereas the non-EL students from feeder elementary schools are only one-third Hispanic. There is also a statistically significant difference between the proportion of EL students who are English-language learners (20 percent) and the proportion found in feeder schools (14 percent) or in the general district population (11 percent). In contrast, EL schools have a

[6] The full-district comparison group may include students from neighborhoods that differ from the areas directly served by EL. For this reason, we also analyze a narrower comparison group limited to the students who attended one of the subsets of district elementary schools (or "feeder" schools) attended by students who eventually enrolled in an EL middle school. Our analysis of student characteristics (for both the full-district comparison group and the feeder-school comparison group) only used administrative records from grade 5, before any of the study students enrolled in EL schools. Data on the comparison groups were limited to study cohorts that contained EL students.

substantially smaller proportion of African American students (22 percent), compared with both feeder schools (40 percent) and local districts (53 percent).

- **EL students in these schools have similar rates of FRPL eligibility and are equally likely to be in special education.** A substantial majority of EL students (71 percent) are eligible for reduced-price meals, which is very similar to the eligibility rate for feeder-school students (72 percent) and five percentage points below the overall district rate (76 percent). The proportion of EL students who received special education before enrolling in EL schools (19 percent) is slightly higher than the rate at feeder schools (17 percent) and similar to the proportion receiving special education in local districts (19 percent).

Table II.2. Characteristics of Students Who Attend EL Study Schools Compared with Students in Feeder Schools and All District Schools

	EL Students at Five Schools	Students at EL Feeder Schools	Students at All District Schools
Hispanic	0.50 N = 1,745	0.33** N = 77,623	0.27** N = 1,019,320
Black	0.22 N = 1,745	0.40** N = 77,623	0.53** N = 1,019,320
Female	0.50 N = 1,745	0.50 N = 77,631	0.49 N = 1,019,413
Free or Reduced-Price Lunch	0.71 N = 1,745	0.72 N = 77,627	0.76** N = 1,019,365
Special Education	0.19 N = 1,745	0.17* N = 77,628	0.19 N = 1,019,379
English-Language Learner	0.20 N = 1,745	0.14** N = 77,627	0.11** N = 1,019,371
Baseline Reading Score (mean z-score)	0.09 N = 1,686	0.09 N = 72,291	0.01** N = 975,290
Baseline Math Score (mean z-score)	0.10 N = 1,717	0.12 N = 74,725	0.01** N = 1,000,908

Note: Values are proportions unless otherwise indicated. Each table cell shows the number of students in the sample (all students with non-missing data for the relevant variable).

* Difference from EL students is statistically significant at the 0.05 level.

** Difference from EL students is statistically significant at the 0.01 level.

- **EL students in these schools also have similar baseline test scores to students from local schools.** As shown in Table II.2, at baseline (one year prior to entering an EL middle school), EL students had similar math and reading test scores to the scores of other students at feeder elementary schools: EL students scored 0.02 standard deviations lower in math and had equivalent scores in reading (neither difference is statistically significant). However, EL students had somewhat higher baseline test scores than the general district population: 0.08 standard deviations higher in reading and 0.09 standard deviations higher in math (both of these differences are statistically significant at the 0.05 level). In other words, there is no evidence that EL recruits higher-performing students

from feeder elementary schools, but this feeder-school population does tend to be somewhat higher-performing than the district average.

Together, these results provide little evidence to suggest that EL schools systematically recruit students who are more advantaged than the local student population. Rather, in some respects (such as the percentage of English-language learners), the EL students in this sample appear to be disadvantaged compared with local students from the same elementary schools, whereas in other respects (such as reduced-price-lunch eligibility, special education, or baseline test scores), differences between the two groups are negligible.

Separately, we also examined whether students at EL schools tend to repeat a grade more often during middle school. In addition to reflecting students' academic performance, grade repetition rates are often a direct consequence of school-level standards and policies and represent an important component of the educational time and resources devoted to each student over the course of middle school. To examine whether EL schools' grade repetition rates tend to differ from other local schools, we compared the average rate of repetition during middle school at EL with the average repetition rates across all local district schools. As Table II.3 shows, there are some small differences. While the grade 6 repetition rate at EL schools (3 percent) is similar to the repetition rate at district schools (2 percent), the grade 7 rate at EL schools (1 percent) is smaller than the rate at district schools (3 percent) by a statistically significant margin. As discussed in the following section, the study's empirical approach carefully accounts for any differences in grade repetition patterns among EL students and comparison students.

Table II.3. Grade Repetition Rates at EL Schools and District Schools

	EL	District
Grade 6	0.03	0.02
	N = 1,448	N = 941,806
Grade 7	0.01	0.03**
	N = 1,311	N = 861,590
Grade 8	0.01	0.02
	N = 973	N = 714,094

Note: Grade repetition represents the average proportion of students in each grade who will be retained in the same grade in the following year. Each table cell shows the number of students in the sample (grade repeaters plus non-repeaters).

* Difference from EL students is statistically significant at the 0.05 level.

** Difference from EL students is statistically significant at the 0.01 level.

III. EFFECTS OF EL MIDDLE SCHOOLS ON ACADEMIC ACHIEVEMENT

This evaluation estimates the average impact of the five EL middle schools on their students' academic achievement. Our approach seeks to measure the achievement growth of EL students relative to the outcomes these same students would have achieved if they had not enrolled in an EL school. To estimate these impacts, we match the EL students with similar students in their districts and compare the average outcomes of the two groups. This matching process was successful in that we identified a comparison group with very similar characteristics and similar baseline achievement in the prior two years before the EL students entered EL schools. After identifying this comparison group, we estimated impacts by running regressions that controlled for any remaining differences in students' demographic characteristics and baseline test scores.

A. Overview of the Estimation Strategy

The study's key outcomes for this impact analysis are student test scores in mathematics and reading. To facilitate making comparisons across districts with different tests, we standardized these test scores by subject, grade, and year using information from the entire sample of students in each district.[7]

The causal (internal) validity of the study's empirical approach depends on the ability of our methods to eliminate or minimize differences in key characteristics between students who enter EL in grade 6 or grade 7 and students in the comparison group who remain in non-EL public schools.[8] To achieve this, our approach used student-level data that included a rich set of student characteristics and multiple years of baseline (prior to EL entry) test scores. We used this information to identify a matched comparison group of students who are similar to EL students in terms of observed demographic characteristics and baseline test scores measured while they were in a non-EL elementary school. The study used "nearest neighbor" propensity score matching to identify this comparison group (for a complete description of the matching procedure, see Appendix B). As shown in Table III.1, there are no statistically significant differences between the baseline test scores of the study's treatment group and the matched comparison group. In addition, the magnitude of the baseline test score differences between these two groups is small (0.05 standard deviations in math and 0.04 standard deviations in reading), allowing the study to meet the What Works Clearinghouse evidence standards for a study of this kind (see Appendix B for additional details).

After we identified the matched comparison group, the second stage of our approach estimated impacts using ordinary least squares (OLS) regressions that control for any remaining baseline or pre-baseline differences between EL students and the matched comparison students. Specifically,

[7] Specifically, we used z-scores defined relative to the distribution of scores in each grade, year, subject, and district. For each student, we calculated the difference between the student's raw score and the district's mean score in that grade, year, and subject, and then divided the difference by the standard deviation of raw scores in the district in that grade, year, and subject. Thus, each impact estimate represents a change in z-scores—that is, a change in the number of standard deviations above or below the mean for the relevant cohort and district. For all cohorts, the distribution of z-scores has a mean of 0 and a standard deviation of 1 for both math and reading.

[8] Specifically, to produce unbiased impact estimates, the design must eliminate differences in student characteristics that could explain academic achievement outcomes and thus be confounded with the EL treatment effect.

the impact estimates adjust for any differences pertaining to demographic characteristics or students' prior two years of pre-EL math and reading test scores. A detailed description of the study's regression models also can be found in Appendix B. This combination of propensity score matching and OLS accounted for differences between the EL group and the comparison group in observed baseline characteristics and achievement scores. If there are no unmeasured differences between the two groups that are correlated with achievement outcomes, the study's analyses should produce unbiased estimates of the impacts of these EL schools. It should be noted, however, that this sample of EL students and matched comparison students could differ in unobserved ways at baseline (such as student motivation or parental attributes) that may also affect later test scores.

Table III.1. Balance Between the Study Sample of EL Students and Matched Comparison Students

Baseline Characteristic	EL	Matched Comparison	Difference	Number with Valid Data
Math Scores (mean z-score)	0.157	0.208	-0.051 (0.044)	3,016
Reading Scores (mean z-score)	0.100	0.139	-0.039 (0.042)	3,016
Female	0.499	0.508	-0.009 (0.026)	3,016
Black	0.197	0.204	-0.007 (0.023)	3,016
Hispanic	0.512	0.499	0.013 (0.025)	3,016
Special Education	0.189	0.181	0.008 (0.020)	3,016
Limited English Proficiency	0.213	0.198	0.015 (0.022)	3,016
Free or Reduced-Price Lunch	0.707	0.703	0.005 (0.022)	3,016

Note: Standard errors reported in parentheses. The total sample includes 1,508 EL students and 1,508 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference in reported values in the "EL" and "Matched Comparison" columns.

* Statistically different from zero at the 0.05 level, two-tailed test

** Statistically different from zero at the 0.01 level, two-tailed test

To calculate the average EL impact across the five schools in the sample, the analysis included all of the student cohorts that could be observed in each outcome year (the study estimated impacts at three different points in time: one year after students enter EL, two years after EL entry, and three years after EL entry). As a result, the number of cohorts in the sample declines in later outcome years. Also, at the time of the study one of the five EL schools had not operated long enough to observe Year 3 outcomes.[9] To obtain the average impact estimates, we calculated a separate impact estimate for each of the five EL schools; we then calculated the average EL effect, assigning an

[9] We also conducted a separate analysis using a restricted sample of student cohorts that could be observed longitudinally for three years in our data. The findings for this alternative sample did not differ substantially from the benchmark results shown here (see Appendix B for additional details).

equal weight to each EL school in the sample. For a detailed explanation of how the study calculated these average impact estimates, see Appendix B.[10]

We also addressed several other threats to the validity of the study's impact estimates: students moving from EL middle schools to other district schools (attrition from EL schools), attrition from study districts (that is, attrition from our data), and students who are retained in grade.

Attrition from EL Schools. Some students in the study sample depart EL schools before the end of 8th grade. This could potentially introduce a form of selection bias if the students who transfer before the end of middle school tend to perform worse than those who remain. In other words, an analysis that includes only persistently enrolled EL students could make the impact estimates look more positive than EL's true impacts. We addressed this problem by permanently assigning to the treatment group any student who enrolled at EL in grades 6 or 7, regardless of whether the student remained in an EL school or transferred elsewhere before the end of middle school. For example, a student who enrolled at EL in 6th grade but transferred out of EL the following year would remain in the study's treatment group throughout his or her middle-school years (including years when the student attended a non-EL school for grades 7 and 8). By holding EL responsible for the achievement of all the students who enroll, including those who transfer out before the end of middle school, this approach is likely to produce a conservative estimate of EL's full impact on students during the years they actually attend EL schools.

Analytic Sample Attrition. For a variety of reasons, some students may not have valid data in the year when a given outcome was measured. For example, some students may transfer to another district that did not provide data to the study, and other students may transfer to local private schools or drop out of school altogether. We categorize these cases when students disappear from the analytic sample as out-of-district transfers. If EL students transfer out of the district at a different rate from matched comparison students, it could undermine the validity of impact estimates. But in fact, our matched comparison group did not exit the analytical sample at an appreciably different rate than the study's sample of EL students: during the three follow-up years we examined, the difference in sample attrition rates for the two groups is below three percentage points (see Appendix B). A different type of analytic sample attrition might occur when students are missing one or more baseline or pre-baseline test scores. To address this, we imputed missing baseline data, ensuring that all students with at least one recorded baseline test score remain in the sample.[11] For a detailed discussion of our imputation methods, including results from an alternative set of impact estimates that do not use imputed baseline test scores, see Appendix B.

Grade Repetition. EL schools retain students in grade 7 at a slightly lower rate than do conventional public schools in their local districts (see Table II.3, discussed in the previous section). This produces a missing-data problem for the analysis of state test scores, as students who repeat a grade do not take the same tests as matched students from their original cohort. Because EL

[10] Appendix B also includes results from an alternative analysis that (instead of weighting each EL school equally) weighted the treatment schools by sample size to calculate the average impact estimates. This alternative weighting scheme did not change the study's main findings.

[11] Specifically, if a student is missing a baseline or pre-baseline test score (provided he or she has at least one baseline or pre-baseline score in the data), we use an imputation procedure to predict that student's missing baseline or pre-baseline scores and use those imputed score values during the matching process and final regression analysis.

students and comparison students are not retained at the same rate, our impact estimates could be biased if we simply excluded all of the retained students from the analysis (this approach would exclude a smaller proportion of EL students and a larger proportion of comparison students). To address this, in the impact analysis we used information on students' past performance to predict (impute) their outcome scores in the years after retention. For more details on this procedure, as well as a detailed discussion of alternative impact estimates we produced using a different approach to handle the scores of retained students, see Appendix B.

In sum, we have employed a variety of methods to address potential threats to validity, and we believe these methods are likely to yield reliable estimates of EL impacts. Indeed, previous studies have suggested that applying a combination of propensity score matching and OLS, as we did in this study, can succeed in replicating experimental impact estimates in certain contexts (Cook et al. 2008; Bifulco 2012; Fortson et al. 2012; Furgeson et al. 2012; Tuttle et al. 2013). Given these past results, we believe the study's matching-based methods represent a strong approach to estimating the impacts of EL schools.

B. Impacts of EL Schools on Test Scores

Below we summarize our main impact findings for all EL students as well as specific subgroups defined by various student demographic characteristics. Table III.2 shows the average estimated impacts of the study's five EL middle schools on reading and math test scores one to three years after students enroll in EL schools.

Table III.2. Mean Test Score Effects in Mathematics and Reading

Outcome	Year 1	Year 2	Year 3
Reading Impact	0.06* (0.03)	0.11** (0.04)	0.16** (0.06)
Number of EL Schools	5	5	4
Math Impact	-0.02 (0.03)	0.09** (0.04)	0.29** (0.07)
Number of EL Schools	5	5	4

Note: Regressions were performed separately for each EL middle school in the sample. Reported impacts are an average of equally weighted impact estimates from regressions of middle-school math and reading z-scores on indicator variables for the number of years after a student's enrollment in an EL middle school. After grade repetition, students were assigned the same z-score received in the last year prior to retention. The sample consists of students who enter EL in grades 6 or 7 matched by district and cohort to students who never enroll in EL; propensity scores were generated separately by EL school, using two years of baseline test scores and all available demographic characteristics. Regression controls include two years of baseline z-scores in math and reading (imputed if one baseline year was missing), as well as dummy variables for demographic characteristics, grade, and cohort. Regressions use robust standard errors (in parentheses) and are clustered on student identifiers.

* Statistically significant at the 0.05 level, two-tailed test

** Statistically significant at the 0.01 level, two-tailed test

The five EL middle schools have positive and statistically significant impacts on student achievement in reading and math.

Based on our impact estimates, the five EL schools have a positive impact on reading achievement after students are enrolled for one year and a positive impact on both reading and math after two years (Table III.2). After students were enrolled in EL for one year, we estimate that the

five EL schools have a positive and statistically significant impact of 0.06 standard deviations on reading test scores and an impact on math test scores (-0.02 standard deviations) that is statistically indistinguishable from zero. In the sample of students observed after two years, these EL schools have statistically significant impacts in both subjects, with impact estimates of 0.11 standard deviations in reading and 0.09 standard deviations in math. Impact estimates remain positive and statistically significant three years after students enter EL, with effects of 0.16 standard deviations in reading and 0.29 standard deviations in math. In short, we find that these schools have a pattern of positive and significant average impacts in both subjects. Below we discuss how to interpret the magnitudes of these impacts.

The impacts of these EL schools represent meaningful gains in student achievement.

In reading, the study estimated that these EL schools had a cumulative impact of 0.11 standard deviations after two years and 0.16 standard deviations after three years. Relative to a normal test distribution, these impacts are equivalent to moving a student from the 50th percentile to the 54th percentile after two years and to the 56th percentile after three years. Another way of interpreting these impact estimates is to compare the EL effect sizes to the national black-white achievement gaps in 8th grade (approximately 0.8 standard deviations in reading and 1.0 standard deviations in math) or to national norms regarding the amount of student learning growth that takes place during middle school (Hill et al. 2008). After enrolling in these EL schools, students experience reading impacts that are equal in magnitude to approximately 14 percent of the black-white achievement gap after two years and 20 percent of the black-white achievement gap after three years. Or in terms of learning growth, our results suggest that EL students experience reading impacts that are large enough to accumulate about an extra five months of learning growth after two years or an extra seven months of learning growth after three years.

In math, the EL schools have an average impact estimate of 0.09 standard deviations after two years and 0.29 standard deviations after three years. Relative to a normal test distribution, EL's cumulative math impacts are equivalent to moving a student from the 50th percentile to the 54th percentile after two years or to the 61st percentile after three years. These math-impacts estimates are equivalent to about 9 percent of the black-white achievement gap after two years and 29 percent of the black-white achievement gap after three years. Expressed in terms of learning growth, the impacts in math are roughly equivalent to three months of extra learning growth after two years or 10 months of extra learning growth after three years.

These results for the study's sample of EL schools—particularly in reading—are of a similar magnitude to some past findings on the impacts of high-performing charter schools. A lottery study of New York City charter schools estimated annual achievement impacts of 0.06 standard deviations in reading and 0.09 standard deviations in math (Hoxby et al. 2009). If these New York City charter schools accumulate such impacts annually over two years, the effects would amount to 0.12 standard deviations in reading and 0.18 standard deviations in math; the EL schools produce similar effect sizes in reading and smaller effects in math. Similarly, a national quasi-experimental study of KIPP charter schools found two-year impacts of 0.14 in reading and 0.27 in math (Tuttle et al. 2013); the EL schools produce impacts that are similar to KIPP impacts in reading but smaller than KIPP impacts in math. Evidence on the impacts of other charter-school management organizations (CMOs) suggests that these EL schools may also be outperforming the average CMO in reading. In a national quasi-experimental study of the impacts of 22 different CMOs, Furgeson et al. (2012) found that after two years, the average CMO had an impact of 0.03 in reading and 0.11 in math (neither effect was statistically significant). The average reading impacts of these EL schools are

larger than those of most other CMOs in that study; the math impacts of these EL schools resemble the math impacts among CMOs.

For several student subgroups of interest, the average EL impact is not appreciably different from the overall average impact among all EL students.

In addition to estimating impacts among all EL students in the sample, we also tested whether there are statistically significant differences in EL impacts for students with different characteristics. Specifically, we measured the difference between average EL impacts on reading and math achievement among members of a given subgroup as well as those outside that subgroup. We found that the average EL impacts are statistically similar to the impacts among each of the subgroups we tested. EL's math and reading impacts among males, African American students, Hispanic students, English-language learners, special-education students, and students eligible for free or reduced-price meals are not significantly different from EL's impacts on other types of students. However, it should be noted that the sample sizes in these subgroup analyses are limited, meaning that the analysis could not detect small differences between the group-level impacts. A detailed discussion of these subgroup results can be found in Appendix B.

IV. QUESTIONS FOR FUTURE RESEARCH

This report represents the first rigorous study of the impact of EL schools. Using careful quasi-experimental methods, we analyzed student achievement at a set of five EL middle schools and found a pattern of positive, statistically significant, and educationally meaningful impacts in both reading and math. Using the study's rich longitudinal data, we also examined the characteristics of students who enroll in EL schools and found little evidence that these EL schools attract a more advantaged student population than local district schools.

As the number of schools using the EL model continues to expand, it will be important to examine the impacts of the EL approach in greater depth and across a much larger sample of schools. Future research might usefully seek to answer several additional questions about the EL model and its effectiveness.

One set of questions pertains to the replicability of this study's results at other EL schools. Are schools throughout EL's national network improving student achievement in reading and math? While this study provides useful evidence on the effectiveness of five EL middle schools, it remains an open question whether other schools are producing similar results by applying the EL model. Future research should continue to examine a wider set of EL schools to test whether the approach is effective at scale and among various types of schools (particularly schools in suburban or rural settings).

A second set of research questions pertains to outcomes beyond reading and math test scores. The EL model includes an interdisciplinary curriculum that emphasizes student inquiry, critical thinking, and the integration of academics across multiple subject areas. For this reason, future research could examine whether EL schools are effective in raising student achievement in science and social studies, as well as examine the effects of EL on student behavior, attitudes, and other non-cognitive outcomes that are emphasized as part of the EL approach and mission.

Finally, researchers should identify which components of EL's multifaceted approach are most strongly associated with achievement impacts. EL schools receive a wide range of possible inputs, including the EL curriculum, professional development institutes, and on-site coaching and training. Given the broad array of services EL provides, policymakers and school leaders would find it valuable to understand which parts of the EL model are most effective. In particular, it is important to investigate whether there is an incremental impact of EL's intensive training and support services, compared with the impact of the EL curriculum alone. This research would provide useful lessons to the EL organization as it grows in scale and would help to inform school leaders as they increasingly seek to implement curriculum reforms that are aligned with Common Core standards. In addition, it is also likely that some schools implement the EL model with greater fidelity than others. By examining practices at a larger sample of EL schools, a future study could examine which components of EL implementation are more tightly linked with positive results. This promising line of research could provide important lessons regarding which aspects of the EL curriculum and support services would be most useful to other school leaders, teachers, and policymakers.

REFERENCES

Bifulco, Robert. "Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." *Journal of Policy Analysis and Management,* vol. 31, no. 3, 2012, pp. 729–751.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management,* vol. 27, no. 4, 2008, pp. 724–750.

Expeditionary Learning. "Expeditionary Learning Core Practices: A Vision for Improving Schools." New York, NY: Expeditionary Learning and Outward Bound, 2011.

Fortson, Kenneth, Natalya Verbitsky-Savitz, Emma Ernst, and Philip Gleason. "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, April 2012.

Furgeson, Joshua, Brian Gill, Joshua Haimson, Alexandra Killewald, Moira McCullough, Ira Nichols-Barrer, Bing-ru Teh, Natalya Verbitsky Savitz, Melissa Bowen, Allison Demeritt, Paul Hill, and Robin Lake. "Charter-School Management Organizations: Diverse Strategies and Diverse Student Impacts." Cambridge, MA: Mathematica Policy Research, January 2012.

Hill, Carolyn, Howard Bloom, Alison Rebeck Black, and Mark Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives,* vol. 2, no. 3, December 2008, pp. 172–177.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. "How New York City's Charter Schools Affect Student Achievement: August 2009 Report." Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009.

Tuttle, Christina Clark, Brian Gill, Phil Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alex Resch. "KIPP Middle Schools: Impacts on Achievement and Other Outcomes." Washington, DC: Mathematica Policy Research, February 2013.

U.S. Department of Education. *What Works Clearinghouse Procedures and Standards Handbook, Version 2*. Washington, DC: U.S. Department of Education, December 2008.

APPENDIX A: ADMINISTRATIVE DATA

In Appendix A, we describe the data used in this report in greater detail. Obtaining student-level longitudinal data was necessary to track individual EL and non-EL students in the baseline years prior to middle-school enrollment. All of the records obtained from Washington, DC, and New York City were de-identified; each student received a unique identifier code to permit longitudinal analyses. We requested variables from the districts' administrative data systems, including test scores in reading and mathematics, demographic characteristics, and schools attended and dates of enrollment. Within each district, Mathematica requested data for all school years, beginning with the year EL first opened a middle school and up to two years prior, to capture baseline data for the maximum number of cohorts.

Table A.1 summarizes the years of data obtained from each district, the demographic variables used in the analyses, and the number of EL middle-school cohorts with at least one year of baseline (pre-EL) test score data and at least one year of test score data after EL enrollment.

Table A.1. Administrative Data Used in Study Analyses

Jurisdiction	Years of Data Collected	Demographic Variables Analyzed	EL School (Grades Served)	Number of Cohorts with Impact Estimates
New York City Public Schools	02-03 to 11-12	Hispanic African American White English-language learner Special Education Free or reduced-price lunch	Washington Heights Expeditionary Learning School (6-8)	6 (06-07 to 11-12)
			Marsh Avenue Expeditionary Learning School (6-8)	3 (09-10 to 11-12)
			Metropolitan Expeditionary Learning School (6-8)	2 (10-11 to 11-12)
Washington, DC, Public Schools	00-01 to 10-11	Hispanic African American White English-language learner Special Education Free or reduced-price lunch	Capital City Public Charter Lower (K-8)	10 (01-02 to 10-11)
			Capital City Public Charter Upper (6-8)	3 (08-09 to 10-11)

Notes: A "cohort" is defined as a group of students who enrolled in that EL school for the first time in grade 6 or grade 7 at the beginning of the school year. In New York City, data on free or reduced-price meals used definitions of eligibility that vary from year to year; because the study approach included a consistent set of treatment and comparison cohorts for each EL school, these data were included in all analyses.

Once we obtained the administrative data, we implemented a data clarification protocol for each district. This process included confirming data variable definitions with district administrators, assessing data coverage gaps, and merging separate district data sets into a single longitudinally linked analytical data set. The analysis file was structured in "long" form, wherein a given student had a separate record for each year he or she appeared in the data.

APPENDIX B: EMPIRICAL METHODS AND SUPPLEMENTAL ANALYSES

A. Propensity Score Matching Procedures

As described in Section III of the report, our quasi-experimental approach identifies a matched comparison group of students who are similar to EL students and then applies an ordinary least squares (OLS) regression model to control for remaining differences. This appendix explains these procedures in greater detail; it also presents the results of alternative analyses that use different estimation procedures to assess whether the study's findings are robust to alternative approaches and assumptions.

The matching process was performed separately for each of the five EL middle schools in the sample. This matching process, described in detail below, consisted of three steps: (1) determining the covariates to be included in the matching model and estimating the matching model, (2) calculating propensity scores for sample members and selecting a matched comparison group based on these scores being close to those of EL students in the sample, and (3) testing the balance of baseline characteristics between our EL sample and matched comparison group.

Matching was conducted separately for each EL school in the sample. For the first step, we separated the students in each district-level data set into cohorts—grade-by-year groups for each EL middle-school entry grade (6th and 7th) in each year observed in the data. For each cohort of students at a given EL school, the pool of eligible comparison students was limited to those in the same district and grade as the EL students the year before they first enrolled in an EL middle school; comparison students were restricted to those never enrolled in EL at any time during middle school or elementary school. We then performed an iterative propensity score estimation procedure on a combined data set of all cohorts at a give EL school. The dependent variable in this propensity score model is an indicator of whether the student enrolled in an EL school in either grade 6 or grade 7.[12] Covariates in the model were selected using an iterative process that identifies the baseline demographic characteristics and test score variables, higher-order terms, and interaction terms that resulted in the best fit of the logistic model. (See Table B.1 for a full list of the potential covariates in each model.) At a minimum, we required the logistic model to include one year of baseline test scores in both math and reading. The other covariates were iteratively included and tested for whether they improved the fit of the logistic model. For this purpose only, we used a cut-off p-value of 0.20, instead of the traditional 0.05, to test for the significance of the covariates. If a potential covariate had a p-value of 0.20 or lower, it was retained in the matching model; it was dropped if its p-value exceeded 0.20.

Next, we calculated propensity scores for EL entry. For any given sample member, the propensity score was calculated by multiplying the model's estimated coefficients by the individual's values for the variables included in the propensity score model. We then performed nearest-neighbor matching (without replacement) of comparison group students to treatment group students, separately by cohort. In other words, for each EL student, we identified the non-EL

[12] We did not distinguish between students who enrolled for part of middle school or for the entire duration of middle school; before matching, all EL students in our data were grouped by the first recorded EL middle school they attended in our data.

district student whose propensity score was closest to that of the EL student. We then tested the balance of the EL group and the matched comparison group by conducting a test of the significance of differences between the two groups in their baseline test scores and other demographic variables (race/ethnicity, gender, special-education status, free or reduced-price-lunch status, and limited-English-proficiency status). For the matched comparison group sample associated with each EL school, we required the baseline test scores of treatment students and comparison students to be balanced in both math and reading; we also required there to be no more than one significant difference on any of the other demographic characteristics listed above. We consider a covariate to be balanced when the means of this covariate for the comparison group are not significantly different from the treatment group at the 5 percent level.[13]

If the first round of matching did not identify a comparison group meeting these criteria, we adjusted the propensity score estimation model for that EL school, re-estimated a new set of propensity scores, obtained a new matched comparison group, and tested for balance between the treatment group and the new matched comparison group. If balance was not achieved in the first round of matching for a given school, we adjusted the propensity score model by removing the variable or interaction term with the least statistical significance (that is, the variable or interaction term that was closest to our p-value cutoff of 0.20). These steps were iterated until we obtained a matched comparison group that achieved balance with the treatment group, according to our criteria.

Table B.1. List of Potential Covariates for Inclusion in Propensity Score Model

Math and reading baseline test scores from one year prior (always included)
Second- and third-order values of math and reading baseline test scores from one year prior
Observed and imputed (when missing) math and reading baseline test scores from two years prior
Observed (non-imputed) math and reading baseline test scores from two years prior
Set of math and reading imputation dummies indicating whether math and reading baseline test scores from one or two years prior are imputed (see Appendix E)
Dummy variables indicating whether student repeated a grade one or two years prior
Demographic variables (gender, race/ethnicity, special-education status, free or reduced-price lunch status, and limited-English-proficiency status, where available)
Interactions of baseline test scores from one year prior and all available demographic variables
Interactions of gender and race/ethnicity variables
Interactions of special-education status and race/ethnicity variables
Interactions of free or reduced-price lunch status and race/ethnicity variables
Interactions of English-language learner status and race/ethnicity variables

[13] The What Works Clearinghouse standards require that the baseline test scores of treatment and comparison groups differ by less than 0.25 standard deviations, provided baseline scores are controlled for in all estimation equations. As shown in Tables B.2, B.3, and B.4, in our sample the groups differ by less than 0.25 standard deviations in both subjects.

B. Baseline Equivalence of the Matched Sample

As described above, we identified a matched comparison group for each of the five EL schools in the analysis. However, the sample we used to estimate impacts on the key test score outcomes varied from the original sample of EL and comparison group students, depending on the outcome year. The matching process included all 6th and 7th grade student cohorts with at least one year of outcome data. The analytic sample size decreases in subsequent outcome years for two main reasons: first, more recent student cohorts had fewer years of available outcome data than earlier cohorts, so fewer were included. Second, within a given cohort, we observed sample attrition at the student level as students transfer out of the district or otherwise drop out of the data set. As a result, impact estimates beyond the first year after EL entry do not include all treatment and matched comparison students measured in Table B.2 (the sample sizes in later outcome years are shown in Table B.3 and Table B.4). To investigate whether the treatment and comparison groups maintained baseline equivalence, the following tables repeat the comparison of baseline scores and demographic characteristics for the portion of the initial sample included in each year's impact estimate. The table notes also describe the treatment and comparison sample sizes for each year, and demonstrate that the rate of analytic sample attrition in the treatment group did not differ substantially from the sample attrition rate in the matched comparison group. For example, in the second outcome year, there is a two percentage-point difference between the sample attrition rate of the treatment group (27.8 percent) and the comparison group (29.8 percent).

Table B.2. Balance Between EL Students and Matched Comparison Students in Year One

Baseline Characteristic	EL	Comparison	Difference	Number with Valid Data
Math Scores (mean z-score)	0.157	0.208	-0.051 (0.044)	3,016
Reading Scores (mean z-score)	0.100	0.139	-0.039 (0.042)	3,016
Female	0.499	0.508	-0.009 (0.026)	3,016
Black	0.197	0.204	-0.007 (0.023)	3,016
Hispanic	0.512	0.499	0.013 (0.025)	3,016
Special Education	0.189	0.181	0.008 (0.020)	3,016
English-Language Learner	0.213	0.198	0.015 (0.022)	3,016
Free or Reduced-Price Lunch	0.707	0.703	0.005 (0.022)	3,016

Note: Standard errors reported in parentheses. Total sample includes 1,508 EL students and 1,508 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "EL" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test
**Significantly different from zero at the 0.01 level, two-tailed test

Table B.3. Balance Between EL Students and Matched Comparison Students in Year Two

Baseline Characteristic	EL	Comparison	Difference	Number with Valid Data
Math Scores (mean z-score)	0.179	0.220	-0.041 (0.055)	2,148
Reading Scores (mean z-score)	0.141	0.160	-0.019 (0.054)	2,148
Female	0.501	0.520	-0.019 (0.032)	2,148
Black	0.207	0.214	-0.007 (0.029)	2,148
Hispanic	0.505	0.477	0.028 (0.031)	2,148
Special Education	0.191	0.187	0.003 (0.025)	2,148
English-Language Learner	0.205	0.199	0.006 (0.026)	2,148
Free or Reduced-Price Lunch	0.699	0.686	0.014 (0.028)	2,148

Note: Standard errors reported in parentheses. Total sample includes 1,089 EL students and 1,059 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "EL" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test
**Significantly different from zero at the 0.01 level, two-tailed test

Table B.4. Balance Between EL Students and Matched Comparison Students in Year Three

Baseline Characteristic	EL	Comparison	Difference	Number with Valid Data
Math Scores (mean z-score)	0.107	0.127	-0.020 (0.119)	1,116
Reading Scores (mean z-score)	0.030	0.033	-0.002 (0.097)	1,116
Female	0.545	0.559	-0.014 (0.065)	1,116
Black	0.215	0.232	-0.017 (0.064)	1,116
Hispanic	0.574	0.557	0.017 (0.066)	1,116
Special Education	0.175	0.224	-0.049 (0.054)	1,116
English-Language Learner	0.304	0.311	-0.006 (0.065)	1,116
Free or Reduced-Price Lunch	0.730	0.673	0.057 (0.056)	1,116

Note: Standard errors reported in parentheses. Total sample includes 587 EL students and 579 matched comparison students. Values are proportions unless otherwise indicated. Due to rounding, the value reported in the "Difference" column may differ slightly from the difference between the values reported in the "EL" and "Comparison" columns.

*Significantly different from zero at the 0.05 level, two-tailed test
**Significantly different from zero at the 0.01 level, two-tailed test

In addition, we also conducted a separate impact analysis that limited the sample to the subset of student cohorts observed longitudinally in our data for all three outcome years (this longitudinal

sample was present at only four of the five EL schools in the full study). As shown in Tables B.7 and B.8 (see model 5, at the conclusion of this appendix), the results of this longitudinal analysis using a smaller sample are nonetheless largely similar to the study's primary findings, which are based on the complete sample of cohorts and schools.

As shown in Tables B.2–B.4, students included in the matched samples for the three outcome years maintained baseline equivalence in prior math and reading scores. That is, in each of the three outcome years, the mean baseline scores of EL students are not significantly different from those of matched comparison students. We also tested for equivalence on demographic characteristics and did not find any large differences. The prevalence of specific demographic groups in the treatment and matched comparison outcome samples never differs by more than six percentage points. For all three outcome years (in other words, for each of the student samples used to estimate impacts in math and reading), there are no statistically significant differences between the observed demographic characteristics of EL students and the matched comparison group.

C. Impact Model and Covariates

As explained above, the first step in our matching-based impact estimation approach was to obtain a matched comparison group with characteristics that resemble the study's sample of EL students. To obtain impact estimates using this matched sample, we estimated an ordinary least squares (OLS) regression model that considered all math and reading test score data from grade 6 through grade 8 to measure students' outcome test scores. To make the analysis of state test scores comparable across districts, all raw test scores were converted to z-scores defined relative to the distribution of scores in each grade, year, subject, and jurisdiction. That is, in each of the two districts, we calculated the difference between each student's raw score and the mean score in that grade, year, and subject, and then divided the difference by the standard deviation of raw scores in the district in that grade, year, and subject. Thus, each z-score reflects the number of standard deviations above or below the mean for the relevant cohort and jurisdiction.[14]

In addition to baseline test scores, the model incorporated baseline (5th grade) demographic controls, including indicators for gender, race/ethnicity, free or reduced-price lunch status, special-education status, grade retention in a baseline year, and English-language-learner status; cohort (year by entry grade); outcome test grade level; and two years of baseline mathematics and reading test scores (4th and 5th grade for cohorts entering EL in grade 6; 5th and 6th grade for cohorts entering EL in grade 7). See Table B.5 for a full list of these covariates. The basic form of the model for each school is defined in equation 1:

$$(1) \quad y_{it} = \alpha + \beta X_i + \delta_1 T1_{it} + \delta_2 T2_{it} + \delta_3 T3_{it} + grade\ dummies\ +\ cohort\ dummies + \epsilon_{it}$$

where y_{it} is the outcome test score for student i in school year t; α is the intercept term; X_i is a vector of characteristics (demographic controls and two years of baseline test scores) of student i; and $T1_{it}$

[14] By definition, the distribution of student z-scores has a mean of 0 and standard deviation of 1 for each subject (math and reading) in each of the three outcome years examined in the impact analysis.

through $T3_{it}$ are binary variables for treatment status in up to three years,[15] indicating whether student *i* had first enrolled at EL one, two, or three years previously, as of school year *t*. For example, $T3_{it}$ would be equal to 1 for student *i* at time *t* if the student had first enrolled at EL at time (*t* - 3), regardless of whether the student was still enrolled at EL at time *t;* otherwise, *T3* would be equal to 0. ε_{it} is a random error term that reflects the influence of unobserved factors on the outcome; δ_1, δ_2, δ_3, and β are parameters or vectors of parameters to be estimated. As the estimated coefficient on the set of treatment indicators, δn represents the cumulative impact of *n* years of EL treatment. Robust standard errors were clustered at the student level since individual students could contribute up to four observations to the analysis sample.

We used the model to separately estimate the impact of each EL middle school in the sample. To calculate the average EL impact, the impact estimate for each EL school was given an equal weight. The standard error of the mean impact across all EL middle schools in the sample uses the pooled student-level variance of school-specific impact estimates for each outcome sample.

Table B.5. List of Covariates Included in OLS Model

Included Covariate
Math baseline test score from 1 year prior
Math baseline test score from 2 years prior
Reading baseline test score from 1 year prior
Reading baseline test score from 2 years prior
Gender indicator variable
Set of race/ethnicity indicator variables
Special-education status indicator variable
Free or reduced-price-lunch status indicator variable
English-language learner status indicator variable
Set of math and reading imputation dummies indicating whether math and reading baseline test scores from 1 and 2 years prior are imputed using method described in Appendix E, Section B
Dummy variables indicating whether student repeated grades in either of the two baseline years
Dummy variables for grades 5-8
Dummy variables for each student cohort in the sample

Note: Baseline test scores were imputed when missing. In some jurisdictions, data were not available on special-education status, free or reduced-price lunch status, or English-language learner status. For more details on the data provided by each jurisdiction, see Appendix A.

We also investigated whether the study's impact estimates were robust to an alternative weighting specification for each EL school. Specifically, we tested whether the average impact estimates were robust to an approach that weights each school-specific impact estimate by the number of students in the sample (this alternative approach gives the greatest weight to the schools that enroll the most students and were open for the longest period of time in our data). Results

[15] Due to a combination of data availability and the year when the EL school opened, at one EL school treatment students in the sample received no more than two years of EL treatment.

using these alternative weights are shown at the conclusion of this appendix in Tables B.7 and B.8 (see model 2). As shown in the tables, results that use these alternative weights are largely similar to average effect estimates that assign an equal weight to each school; under both approaches, impacts remain statistically significant and positive in reading for all outcome years. In math, using the alternative weights impacts were positive and statistically significant in all three outcome years (for comparison, when schools were weighted equally, the math impact estimates were positive and significant only in the second and third outcome years).

Finally, we tested whether the impact estimates may have been affected by details of the procedure we used to match comparison group students to EL students, which was based on nearest-neighbor matching without replacement. To do so, we estimated the impacts of the five EL schools using the entire district as a comparison group instead of identifying a matched comparison group. In other words, the comparison group is formed without propensity score matching, but the regression model in equation 1 is still used to control for baseline characteristics of EL students and comparison group students. Using a district-wide comparison group produces impact estimates (model 6 in Tables B.7 and B.8) that are very similar to the benchmark results—in both reading and math, the impact estimates have the same sign and statistical significance in all three outcome years, and the magnitude of each point estimate is similar to our benchmark results. Although these district-wide comparison group estimates are close to our matching results, there is a potential drawback to comparing EL students to all students in the relevant public school district. Under such an approach, the sample of comparison students may include individuals who are very different at baseline from the students who enroll in EL schools. OLS models adjust for these differences, but the adjustments depend on assumptions about the underlying relationship between each characteristic and the achievement results. Impact estimates that use a matched comparison group help to avoid relying on these assumptions, which is why our preferred matching-based impact estimates rely on propensity score matching. This ensures the treatment and comparison groups share similar demographic characteristics and prior achievement trajectories.

D. Imputation for Missing Baseline Data and Retained Students

This section explains in greater detail how our analysis handled two types of missing data: (1) students missing data on one of their test scores either one year before an EL entry grade or two years before a EL entry grade or (2) students who were retained in grade and therefore are missing a test score on the outcome test(s) given to the remaining cohort.

1. Imputation for Missing Baseline Data

Our benchmark analyses used data sets with imputed baseline test scores created by conducting single stochastic regression imputation for missing baseline test scores; imputation was completed separately by treatment status. This imputation process involved estimating the following model:

(2a) $$Yp\_math_{it} = \alpha + X_i\beta + \sum_r \varphi_r Yr\_math_{it} + \sum_{q=3}^{8} \gamma_q Yq\_reading_{it} + \varepsilon_{it}$$

(2b) $$Yp\_reading_{it} = \alpha + X_i\beta + \sum_r \varphi_r Yr\_reading_{it} + \sum_{q=3}^{8} \gamma_q Yq\_math_{it} + \varepsilon_{it}$$

where Yp_math_{it} is a single grade p math baseline test score for student i at time t; $Yp\_reading_{it}$ is a single grade p reading baseline test score for student i at time t; X_i is a vector of demographic characteristics (gender, race/ethnicity, special-education status, free or reduced-price lunch status, and English-language learner status, where available) of student i; Yr_math_{it} and $Yr\_reading_{it}$ are all available for grades 3–8, excluding grade p math and reading baseline or outcome test scores for

student *i* at time *t*; and Yq_math_{it} and $Yq\_reading_{it}$ are all available for grades 3–8 math and reading baseline or outcome test scores for student *i* at time *t*. Note that the treatment dummies are not part of the imputation model because imputation is performed separately for the treatment group and then the comparison group.

We first estimated equations (2a) and (2b) for baseline test scores one and two years prior to EL entry using those students in our sample who have non-missing scores on these tests. For students with missing values for a given test, we used that student's demographic characteristics and other non-missing test scores (in other words, values of the right-hand-side variables in equations 2a and 2b) and multiplied them by the estimated coefficients from the model. This gave us a predicted value of the missing test score for that student. We imputed missing baseline test scores only for students who had at least one non-missing baseline test score in either math or reading.

Finally, to obtain the imputed baseline test scores, we added a stochastic component to the predicted values Yp_math_{it} and $Yp\_reading_{it}$. For each student, the stochastic component is randomly selected from the set of all residuals estimated in equations (2a) and (2b) for the full sample. The stochastic component is included to ensure that the variance of the imputed baseline test scores is the same as that of the observed values.

To test whether our results are sensitive to this imputation strategy, we estimated our benchmark model using the subsample of students with complete baseline test score data—that is, we dropped students with missing baseline scores from the sample and compared the EL students for whom we did not impute scores to matched comparison students for whom we did not impute scores (see model 3, in Tables B.7 and B.8). The results for this smaller sample are very similar to our benchmark impact estimates: for all three outcome years, the sign and statistical significance of the EL impact in both subjects remains the same, and the magnitude of the impact estimates remains within 0.06 standard deviations of the benchmark estimates as well.

2. Imputation for Students Repeating a Grade

We also impute the math and reading state test scores of students who repeat a grade if they were retained in one of the study's three outcome years. For example, if a student in the treatment group entered EL in grade 6 and then repeated grade 6, he or she would still be in grade 6 (and would take the grade 6 state assessment) at the end of the second follow-up year; members of his or her cohort who remained on track would have taken the grade 7 state assessment in that year. Because the grade repeater's grade 6 assessment score would not be comparable to grade 7 scores, we treat this student's year 2 follow-up score as missing and impute its value. We use the following approach: for each grade repeater, in the year of repetition and subsequent years, we impute the student's z-score on the cohort-appropriate (rather than grade-appropriate) test by setting his or her score equal to the student's standardized score in the last year prior to grade repetition. In this example, we would use the standardized score of the grade repeater on the grade 6 assessment in the first follow-up year (the score from the first time the student took that assessment). In effect, this imputation procedure assumes students maintain the same percentile rank relative to their cohort in the year of grade retention and in all subsequent years. In other words, we assume that each retained student does neither better or worse in relative terms than before retention.

To test the sensitivity of our results to the method used for retained students, we also estimated EL impacts using an alternative approach to analyzing the test scores of retained students. In model 4 shown in Tables B.7 and B.8, we estimate the impacts of EL using the recorded test scores of

grade repeaters in all years, without any adjustments. In other words, within each student cohort this analysis compares the scores of retained students taking one test in a given year to the scores of non-retained students taking a different test (one grade level higher) in that year. Using the observed scores of retained students in all years does not change any of the study findings—EL's impact estimates retain the same sign and statistical significance in all outcome years, and the magnitude of the impact estimates changes by less than 0.02 standard deviations in both subjects.

E. EL impact estimates for student subgroups

In this section, we present the estimates derived to identify whether EL had differential impacts on particular subgroups of students. In general, our strategy to identify potential subgroup differences was to use interaction terms consisting of treatment indicators multiplied by subgroup variables. The coefficients on the interaction terms represent the marginal effect of EL for students in the specific subgroup above and beyond the average EL effect among other students. The statistical significance of the interaction term indicates whether the EL effect is different for the subgroup in question than for other EL students.

Table B.6 shows whether there are statistically significant differences in EL's impact on math and reading achievement for students with different characteristics. In other words, the results described in the tables show whether there is a significant difference between EL's average impact among members of the listed subgroup and the impact among those who are not members of the subgroup. A positive and significant interaction indicates that EL's average impact is higher for the listed subgroup relative to all other EL students. Each subgroup analysis included only EL schools in which more than 5 percent of its students were part of the subgroup of interest. Thus, as shown in these two tables, the sample of included EL schools varies depending on the subgroup being examined. To calculate the average of subgroup effect estimates at these schools, all of the included EL schools were weighted equally.

As shown in Table B.6, EL impacts do not differ in a majority of outcome years for students with any of the characteristics we tested (male students, Hispanic students, African American students, English-language learners, students receiving special education, and students eligible for reduced-price meals). However, the number of students in the sample with each of these characteristics was often small, and the number of schools that could be included in the subgroup analysis often differed depending on the characteristics being measured and the outcome year. As a result, these subgroup analyses had limited statistical power to detect small differences in effects.

Table B.6. Comparison of EL Effects on Subgroups to Effects on Other EL Students

Subgroup	Reading		Mathematics	
	Year 1	Year 2	Year 1	Year 2
Male	Larger [5]	Not Different [5]	Not Different [5]	Not Different [5]
Hispanic	Not Different [5]	Not Different [5]	Not Different [5]	Not Different [5]
African American	Larger [4]	Not Different [3]	Larger [4]	Not Different [3]
Special Education	Not Different [5]	Not Different [5]	Not Different [5]	Larger [5]
English-Language Learner	Not Different [4]	Not Different [3]	Not Different [4]	Not Different [3]
Free or Reduced-Price Lunch	Not Different [5]	Not Different [5]	Not Different [5]	Not Different [5]

Note: The number of EL schools in each analysis is shown in brackets. Table rows describe the difference in EL's average impact comparing members of the subgroup to those who are not members of the subgroup. A "larger" label indicates that the impact estimate is higher for the examined subgroup by a statistically significant margin ($p < 0.05$). A "smaller" label indicates that the estimate is lower by a statistically significant margin for the examined subgroup.

* Statistically significant at the 0.05 level, two-tailed test
** Statistically significant at the 0.01 level, two-tailed test

Table B.7. Comparison of Benchmark Impact Model and Alternative Models, Reading

Model	Year 1	Year 2	Year 3
1. Benchmark Model, Schools Weighted Equally	0.06* (0.03)	0.11** (0.04)	0.16** (0.06)
2. Benchmark Model, Schools Weighted by Sample Size	0.05* (0.02)	0.06* (0.03)	0.07* (0.03)
Alternative Approaches to Imputing Data			
3. Non-Imputed Baseline Data	0.05* (0.03)	0.11** (0.04)	0.20** (0.08)
4. Non-Imputed Grade Repeater Scores	0.06* (0.03)	0.11** (0.04)	0.15* (0.06)
Estimates Using a Consistent Sample of Cohorts (limited to four EL schools)			
5. Cohorts Observed for Three Outcome Years	0.04 (0.06)	0.17** (0.06)	0.17* (0.08)
Districtwide Comparison Group Without Matching			
6. Results with No Matching	0.05* (0.02)	0.13** (0.02)	0.12* (0.05)

Note: Each row shows EL impact estimates under different analytical approaches and assumptions, with standard errors in parentheses. Models 1 through 5 use the study's matched comparison group. In model 2, schools were weighted by sample size instead of being weighted equally; model 3 does not include imputed baseline test scores; model 4 uses the observed test scores of retained students; model 5 uses a longitudinal sample of students who were observed at four EL schools for all outcome years; model 6 includes all comparison students in local districts without matching.

* Statistically significant at the 0.05 level, two-tailed test
** Statistically significant at the 0.01 level, two-tailed test

Table B.8. Comparison of Benchmark Impact Model and Alternative Models, Mathematics

Model	Year 1	Year 2	Year 3
1. Benchmark Model, Schools Weighted Equally	-0.02 (0.03)	0.09** (0.04)	0.29** (0.07)
2. Benchmark Model, Schools Weighted by Sample Size	0.05** (0.02)	0.07** (0.02)	0.19** (0.03)
Alternative Approaches to Imputing Data			
3. Non-Imputed Baseline Data	-0.01 (0.03)	0.13** (0.04)	0.35** (0.08)
4. Non-Imputed Grade Repeater Scores	-0.02 (0.03)	0.09** (0.04)	0.30** (0.07)
Estimates Using a Consistent Sample of Cohorts (limited to four EL schools)			
5. Cohorts Observed for Three Outcome Years	0.05* (0.06)	0.17* (0.07)	0.32** (0.09)
Districtwide Comparison Group Without Matching			
6. Results with No Matching	-0.03 (0.02)	0.09** (0.03)	0.20** (0.05)

Note: Each row shows EL impact estimates under different analytical approaches and assumptions, with standard errors in parentheses. Models 1 through 5 use the study's matched comparison group. In model 2, schools were weighted by sample size instead of being weighted equally; model 3 does not include imputed baseline test scores; model 4 uses the observed test scores of retained students; model 5 uses a longitudinal sample of students who were observed at four EL schools for all outcome years; model 6 includes all comparison students in local districts without matching.

* Statistically significant at the 0.05 level, two-tailed test
** Statistically significant at the 0.01 level, two-tailed test

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC