AMERICAN INSTITUTES FOR RESEARCH®

Evidence of Deeper Learning Outcomes: Technical Appendix

Report #3 Findings From the Study of Deeper Learning: Opportunities and Outcomes

Prepared for:
The William and Flora Hewlett Foundation

Prepared by:
American Institutes for Research
1000 Thomas Jefferson St. NW
Washington, DC 20007
http://www.air.org

September 2014

Evidence of Deeper Learning Outcomes: Technical Appendix

Report #3 Findings From the Study of Deeper Learning: Opportunities and Outcomes

September 2014

Appendix Authors:
Kristina L. Zeiser, Jordan Rickles, James Taylor, and Michael S. Garet, American Institutes for Research

Principal Investigators:
Jennifer O'Day, American Institutes for Research
Michael S. Garet, American Institutes for Research

Study Team:

AMERICAN INSTITUTES FOR RESEARCH

Catherine Bitter
Megan Brown
Seth Brown
Marjorie Chinen
Mette Huberman
Jamie Shkolnik
Marian Eaton
Tara Zuber
Allison Waters

The Research Alliance for New York City Schools

Molly Alter
James Kemple

Contents

Acknowledgements: The authors would like to acknowledge the many people who helped make this study possible. We thank the thousands of students, teachers, principals, and network and district staff who agreed to provide responses to the study's many data collections. We extend our appreciation to the William and Flora Hewlett Foundation for the grant that made this study possible, and particularly to Kristi Kimball for her initiation of the project and Marc Chun and Barbara Chow for their consistent support of this study. Our thanks also go to Elizabeth Stuart at Johns Hopkins University and Luke Keele at AIR and Pennsylvania State University for technical advice, Kerstin Carlson LeFloch and Laura Salganik at AIR for their careful technical review, and Emma Ruckley for her thorough editing. We are also grateful to Jim Kemple and the staff at the Research Alliance for New York City Schools for their expert analysis in this collaboration. The statements, findings, and conclusions here are those of the authors and study leads and do not necessarily represent the viewpoint of these organizations or individuals.

I. Introduction

The *Study of Deeper Learning: Opportunities and Outcomes* is a proof-of-concept study focused on students who attended high schools with at least moderately well implemented network approaches targeting deeper learning (network schools) and schools that were not implementing network approaches targeting deeper learning but served similar populations of students (non-network schools). The study was conducted in pairs of network and non-network schools that serve similar disadvantaged student populations in several districts in California and New York City.

This appendix provides an extended description of the study's sampling procedures, data sources, and analytic methods. It begins by describing how network and non-network schools were selected and recruited to participate in the study. After presenting the characteristics of the participating schools, we describe the student samples, the levels of student attrition between Grade 9 entry and data collection, and the selection of student samples for primary data collection. After describing the instrumentation and administration of our three types of primary data collection—the student survey, the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA)-Based Test for Schools (PBTS), and teacher assignments—we provide information about the creation of weights and the statistical models used within the report. The appendix concludes with tables and figures that contain the findings discussed in the report.

II. Study Sample

A. Network School Recruitment and Comparison School Selection

In 2011–12, the Hewlett Foundation selected ten school networks to participate in what would become the "Deeper Learning Community of Practice." The purpose of this community of practice is to share strategies, tools, and lessons that both contribute to the work of the networks themselves and build the broader knowledge base about deeper learning. The main selection criterion for the networks were as follows:

- The networks needed to have experience in—and an explicit focus on—promoting a deep understanding of content and the kinds of competencies reflected in the Hewlett Foundation's identified dimensions of deeper learning.
- They needed to do this across whole schools serving diverse populations of students (rather than targeting only certain portions of the students or teachers in a school).

The Hewlett Foundation selected the Community of Practice networks prior to the start of the *Study of Deeper Learning: Opportunities and Outcomes*. The ten networks represented in this study have a well-established history of promoting deeper learning and all share an emphasis on providing educational opportunities for minority students and students from low-income families to prepare them for college and career. To address our primary research questions, we recruited a set of 20 network high schools from the ten networks. Criteria for network school selection are reported in Exhibit 2.1.

Given the small number of network schools in the sample, and given the criteria used to select the sample, the study's findings are limited in terms of their generalizability. For example, the ten networks include many schools that were excluded by the study's criteria (such as elementary and middle schools, very small schools, schools without substantial disadvantaged populations, and schools that opened very recently). Furthermore, because we included only moderate to high implementers of the network models, findings cannot be generalized to all schools trying to implement a deeper learning approach.

The network schools were drawn from ten different networks, and the treatment evaluated in this study is therefore heterogeneous. The networks' approaches vary, but as we discussed in Report 1 of the study (Huberman et al., 2014), the approaches in the sampled high schools typically included several common elements, including engagement in project-based learning involving collaboration and real-world experiences; use of authentic assessment (such as portfolios and exhibitions) to measure student achievement and progress; and development of personalized learning environments. The study was not designed to determine the relative effectiveness of the networks; rather, it was designed to assess whether schools can promote deeper learning across a variety of reasonably well-implemented approaches and a diversity of students.

Exhibit 2.1. Network and Non-Network School Eligibility Requirements

	Network School Criteria	Non-Network School Criteria
Regular high school (i.e., not a special education, vocational, or alternative high school)	✓	✓
Non-magnet school	✓	✓
Non-charter school		✓
Low grade is Grade 9		✓
Low grade is Grades K–9	✓	
High grade is Grade 12	✓	✓
25+% of students are eligible for free/reduced-price lunch	✓	✓
200+ students enrolled in Grades 9–12	✓	✓
Been in the network since the 2007–08 school year	✓	
Schoolwide implementation of the network approach	✓	
A moderate or high implementation rating from the network	✓	
Within the same district as a network school or a surrounding district		✓

Note: Some deeper learning networks begin focusing on deeper learning competencies before Grade 9. While these network schools included grades below Grade 9, we selected for our study students who did not attend a deeper learning network school until Grade 9. No non-network schools selected for the study had students below Grade 9.

To select non-network schools, we first identified schools with a population of incoming Grade 9 students similar to the incoming Grade 9 students at the network schools. We identified a set of eligible non-network schools located in the same school district as the network school (if the network school was operated by a school district), or within the surrounding school district of the network school (if the network school was operated by a charter school management organization). Schools were identified using the 2007–08, 2008–09, and 2009–10 Common Core of Data (CCD) and were deemed eligible if they met the criteria in Exhibit 2.1. Specifically, we used the 2007–08 data to determine if the school was in existence as of the 2007–08 school year, and we used averages from the 2008–09 and 2009–10 school years to determine the overall number of students and the percentage of students eligible for free or reduced-price lunch (FRPL). We expected the distribution of students across racial/ethnic categories to be relatively stable across years for most schools, so we relied on the 2009–10 data.[1]

Based on the CCD data, we identified up to five matches for each network school relying on Mahalanobis distances that were computed using four variables: the average percentage of students eligible for free or reduced-price lunch, the percentage of African American students, the percentage of Hispanic students, and the percentage of white students from the 2008–09 and 2009–10 CCD. To guard against matching dissimilar schools, we required comparison schools to be within one standard deviation of its paired network school on each of the four variables we used to calculate Mahalanobis distance. After receiving extant district data, we also compared the Grade 8 achievement of students in the network school and students in the selected comparison schools to determine priorities for school recruitment.

We encountered two challenges as we worked to secure the desired sample of schools. First, we found that some selected schools were reluctant to participate because of the data collection burden and their heavy workloads. Some candidate schools reported that they were overwhelmed by recent policy initiatives, standardized testing, preexisting research projects, staffing or facilities transitions, budgetary cuts due to the recession, and a range of other unique local factors. We employed a number of strategies to address this recruitment difficulty, including increasing incentives and honoraria for participation and involving the district leadership and/or research department in the recruitment process. Despite these efforts, some of the highest implementing network schools and some of the non-network schools that were our preferred choices (because they were the best matches based on demographic data from the CCD and achievement data from the districts) did not elect to participate in the study. Second, in some schools that agreed to participate, we encountered challenges in obtaining active parental consent for individual students' participation in the data collection activities in the districts for which it was required. While many schools were able to manage the active consent process with our assistance quite well, six schools were unable to collect sufficient numbers of signed consent forms to participate in the student-based data collections. As a result, analyses of student survey and PBTS data, which required parental consent, did not include all of the schools that were included in analyses of outcome data, which did not require parental consent (see Zeiser et al., 2014). As we discuss later in this appendix, we ran sensitivity analyses where possible to determine if these challenges affected study results.

[1] While we expected school characteristics to be reasonably stable from 2007–08 to 2009–10, schools that had recently opened might have experienced changes in enrollment during the first few years after opening. For example, if a school opened in 2007–08, and it first enrolled only Grade 9 students and added a grade each year, its highest grade would have been Grade 9 in 2007–08, Grade 10 in 2008–09, and Grade 11 in 2009–10. Similarly, the school's enrollment would have increased over the same period. As such, selection criteria were modified for recently opened schools. To ensure a sufficient sample size for schools that had recently opened, we removed schools with fewer than 200 students, *on average*, between the 2008–09 and 2009–10 school years (rather than within each school year), even if the school only had two and three cohorts of students in those years, respectively.

An overview of the matched pairs for which we were able to collect student survey data for this report's analysis is provided in Exhibit 2.2.

Exhibit 2.2. Description of School Pairs

		Enrollment	% Female	% African American	% Hispanic	% Asian	% FRPL
Pair 1 (CA)	Network (1N)	400	70	30	40	10	70
	Non-Network (1C)	2100	50	20	20	30	40
Pair 2 (CA)	Network (2N)	300	50	10	40	0	40
	Non-Network (2C)	1600	50	20	30	10	50
Pair 3 (CA)	Network (3N) [a]	400	50	20	50	10	60
	Non-Network (3C)	1800	50	40	20	20	50
Pair 4 (CA)	Network (4N)	300	50	0	90	10	50
	Non-Network (4C)	2300	50	0	90	10	70
Pair 5 (CA)	Network (5N)	400	50	0	100	0	40
	Non-Network (4C)	2300	50	0	90	10	70
Pair 6 (CA)	Network (6N)	600	50	10	10	10	30
	Non-Network (6C)	2600	50	10	30	0	20
Pair 7 (CA)	Network (7N1)	400	50	10	10	10	40
	Network (7N2)	400	50	10	10	10	40
	Non-Network (7C)	2500	50	10	30	10	50
Pair 8 (NY)	Network (8N)	500	60	10	20	10	40
	Non-Network (8C)	600	60	10	20	20	50
Pair 9 (NY)	Network (9N)	400	60	40	60	0	80
	Non-Network (9C)	400	40	40	50	0	70
	Non-Network (9Cb)	500	50	30	60	0	80
Pair 10 (NY)	Network (10N)	400	40	0	40	60	100
	Non-Network (10C1)	600	50	0	100	0	80
	Non-Network (10C2)	500	50	0	90	10	90
Pair 11 (NY)	Network (11N)	400	50	20	40	30	100
	Non-Network (10C1)	600	50	0	100	0	80
	Non-Network (10C2)	500	50	0	90	10	90
Pair 12 (CA)	Network (12N)	300	50	60	30	0	40
	Non-Network (3C)	1800	50	40	20	20	50
Pair 13 (NY)	Network (13N)	400	60	80	20	0	80
	Non-Network (13C)	400	60	70	20	0	80
Pair 14 (NY)	Network (14N)	400	50	80	20	0	100
	Non-Network (14C)	500	50	80	10	0	70
Pair 15 (NY)	Network (15N)	300	50	40	60	0	70
	Non-Network (9C)	400	40	40	50	0	70
Pair 16 (CA)	Network (16N)	300	60	0	80	10	70
Pair 17 (MN)	Network (17N)	200	40	80	0	0	100
Pair 18 (ME)	Network (18N)	300	50	20	10	0	0
Pair 19 (MA)	Network (19N)	700	50	20	40	0	60

(See notes on the following page.)

Note: School demographics from the 2010–11 Common Core of Data (CCD). To ensure school confidentiality, enrollment is rounded to the nearest 100 students and percentages are rounded to the nearest 10 percent.

Schools Included by Report:

Report 1. All network schools in this exhibit were included in qualitative analyses in Report 1 except school 13N, which was omitted due to incomplete qualitative data. All non-network schools were included in qualitative analyses in Report 1 except 13C (due to incomplete qualitative data) and 14C (which did not participate in qualitative data collection). All schools in Pair 1 to Pair 11 were included in the teacher survey sample.

Report 2. All schools from Pair 1 through Pair 11 were included in the student survey sample and were used in Report 2, with the exception of School 9Cb. School 9Cb was included in analyses of teacher assignments.

Report 3. All schools from Pair 1 through Pair 15 (excluding School 9Cb) were included in Report 3. School 9Cb was omitted because it did not participate in primary data collection. Schools in these pairs had student survey data, extant data, or both.

Details on Specific School Pairs:

Schools 4N and 5N are located in the same district, and we were able to recruit only a single non-network school in this district. The students in this non-network school were matched to students in both School 4N and School 5N.

Schools 7N1 AND 7N2 were associated with the same deeper learning network and resided on the same campus. Because the schools were small in size, we combined the students attending them and treated them as single network school in the analyses in reports 2 and 3, comparing it with 7C. For qualitative analyses and teacher survey analyses in Report 1, these two schools were counted as two separate network schools.

School 9Cb was originally selected as the non-network school for School 9N, but it did not reach the consent rate required to participate in the student survey and PBTS data collection, so School 9C was used instead. School 9Cb was included in the qualitative analyses and analyses of teacher assignments.

Due to small sample sizes, Schools 10C1 and 10C2 (non-network schools) were combined and treated as a single non-network school. Both non-network schools served populations that were similar to Schools 10N and 11N (network schools), which were associated with the same deeper learning network. The propensity scores for Pairs 10 and 11 were based on a combined sample that included both Schools 10N and 11N (network schools) and Schools 10C1 and 10C2 (non-network schools), because of the limited sample size within the individual network and non-network schools. Once the propensity scores had been computed, however, Pairs 10 and 11 were considered separate pairs for the purposes of the impact analysis and meta-analysis.

For the analysis of graduation, achievement test score, and postsecondary data, School 12N (a network school) was matched with School 3C (a non-network school), which was also used as the non-network school for School 3N (a network school).

For the analysis of graduation, achievement test score, and postsecondary data, School 15N (a network school) was matched with School 9C (a non-network school), which was also used as the non-network school for School 9N (a network school).

[a] Due to missing data in the 2010–11 CCD, demographic information for this school come from the 2011–12 CCD, and free or reduced-price lunch information for this school came from 2011–12 enrollment data from the California Department of Education, 2011–12.

B. Student Samples

The study is both retrospective and prospective. Using extant, student-level district data, we first identified cohorts of Grade 9 students entering selected high schools in prior academic years. We then prospectively followed these students to administer student surveys and assessments, observe high school graduation, and collect data on enrollment in postsecondary education. In each matched pair, the study focused on five student cohorts:

- Cohort 1: Students who entered Grade 9 in 2007–08
- Cohort 2: Students who entered Grade 9 in 2008–09
- Cohort 3: Students who entered Grade 9 in 2009–10
- Cohort 4: Students who entered Grade 9 in 2010–11
- Cohort 5: Students who entered Grade 9 in 2011–12

To account for preexisting differences between students attending network and non-network schools in our analyses, we restricted the sample to students who had data on Grade 8 characteristics, including middle school state standardized test scores, in the available district extant data (described below). This requirement restricted our student cohort samples to students who attended a district school in Grade 8, so our results may not generalize to students who attended a school in our sample in Grade 9 but attended a non-district middle school.

The analyses for this report (which is primarily based on student survey results) were based on students in Cohort 3 and Cohort 4. We chose to focus on these two cohorts for two reasons. First, we could not include Cohort 1 and Cohort 2 students because they had already graduated from high school by the time of our primary data collection in spring 2013, and thus they were not present to complete the study surveys and assessments. Second, we chose to focus data collection and analysis on students who had been exposed to the “treatment” for multiple years, which led us to exclude Cohort 5 students from this analysis. Therefore, student-level analyses in this report are based on students who entered Grade 9 in 2009–10 or 2010–11 and consented to participate in study data collection during spring 2013. (See Exhibit 2.3.) At that time, most students were in Grade 11 or 12.

For primary data collection, our goal was to collect data from a total of 260 students within each school pair (65 Grade 11 students and 65 Grade 12 students in the network and non-network schools). We selected student samples for primary data collection based on propensity score quintiles to ensure we were sampling similar groups of students in each pair of schools. The propensity score quintiles were defined based on the distribution of network students’ estimated propensity scores—the conditional probability of being assigned to the treatment condition (network school enrollment) given a set of observable covariates (Rosenbaum & Rubin, 1983). Propensity scores were estimated using students’ Grade 8 achievement scores (mathematics, language, and science if available), English language learner (ELL) status, gender, special education status, measures that captured students’ socioeconomic status, and race/ethnicity.

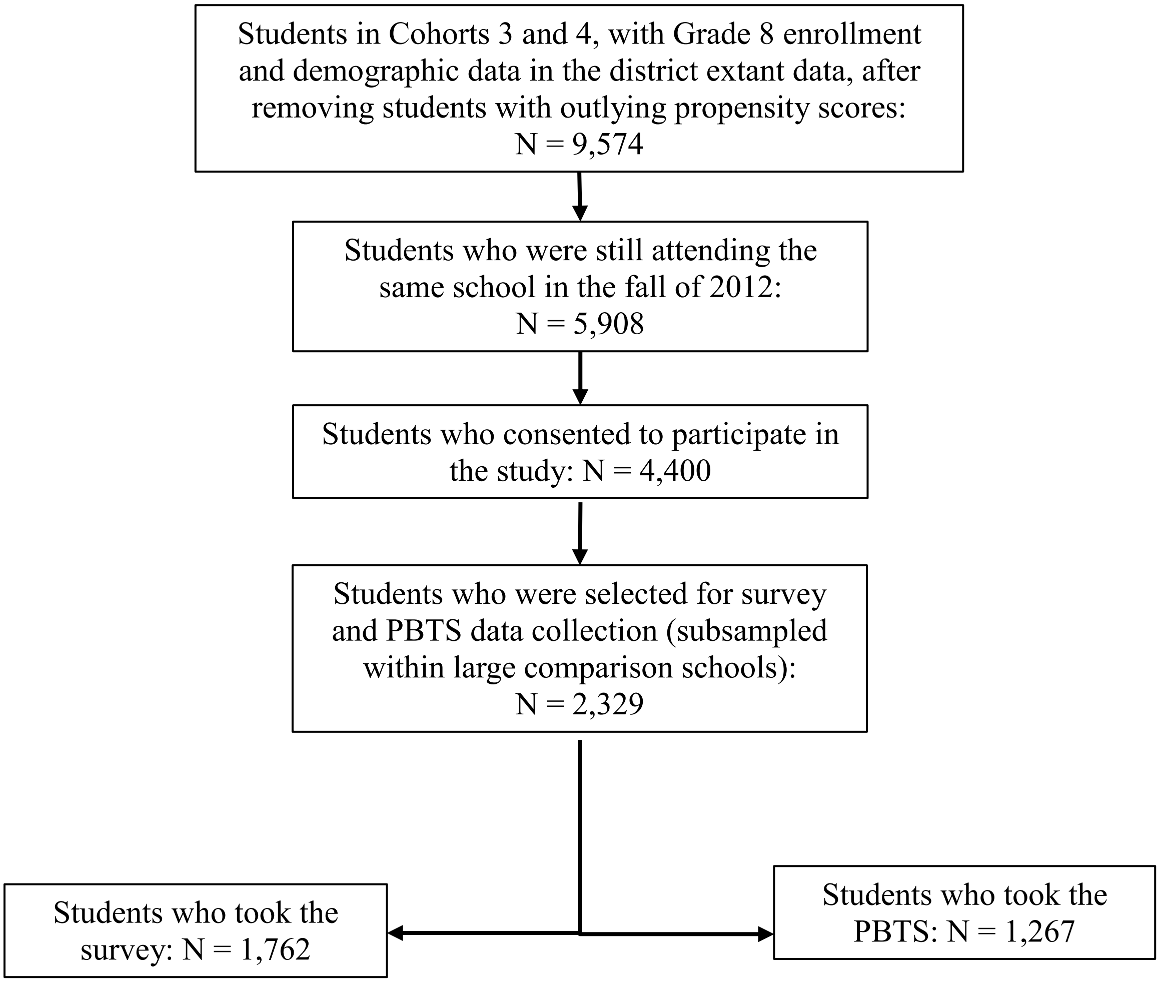
To ensure that the students we sampled in matched non-network and network schools had similar background characteristics, we removed students in non-network schools from the top propensity score stratum if they had unusually high propensity scores and from the lowest stratum if they had unusually low propensity scores. More specifically, we did not include non-network students whose estimated propensity scores fell outside the range of "common support," which is loosely defined as the range of propensity scores of students within the matched network school.[2]

Within each school pair, we sampled all consented students from network schools. However, because non-network schools tended to be larger in size, we subsampled consented students from these schools by randomly selecting students based on their propensity score quintile and the number of network students in the quintile. As a result, selected samples of network and non-network students had similar distributions of propensity scores within each matched pair of schools. Since the propensity scores reflect student background characteristics, the selected samples of network and non-network students also had similar characteristics. See Section IV.A for a more detailed discussion of the propensity score estimation process.

Analyses of high school graduation and postsecondary outcomes include all students with propensity scores that fall within the range of common support in Cohorts 1–3. Analyses of high school achievement test scores include all students within the range of common support in the first four cohorts. The only non-network students excluded from analyses of high school graduation and postsecondary outcomes were those whose propensity scores were not within the range of propensity scores among students who attended the matched network school. This resulted in excluding 93 non-network students from Cohorts 1, 2, and 3 (0.8 percent of all non-network students with propensity scores in these cohorts). For the analyses of high school achievement test scores, we also excluded students who did not remain in the district long enough to take the exam, and we incorporated attrition weights in these analyses. (See Section IV.A.)

[2] To ensure that we did not remove non-network students whose propensity scores were close in value to the propensity scores of network students, we created an allowable range of propensity scores that included the minimum and maximum propensity scores among network students. We determined the minimum allowable propensity score by subtracting 0.25 times the standard deviation of the propensity score distribution from the minimum propensity score for network students, and we determined the maximum allowable score by adding 0.25 times the standard deviation to the maximum propensity score for network students.

Exhibit 2.3. Number of Students From the Initial Grade 9 Sample to the Data Collection Sample (Cohorts 3 and 4)[3]



III. Data Sources and Measures

To address the primary research questions for this study, we collected outcome data from students and district administrative records. An overview of the data sources, including coverage across schools and students, is provided in Exhibit 3.1. Additional details about the data sources are available upon request. In addition to outcome data, student-level administrative records from the participating districts were collected for all students who entered Grade 9 in one of the five study cohorts in order to estimate propensity scores and include covariates in outcome models.

[3] As described in Exhibit 2.2, three non-network schools (School 4C, School 10C, and School 11C) were each included in two different school pairs so that they could be matched with two different network schools. Therefore, the counts presented in Exhibit 2.3 and the remaining exhibits include the non-network students within these schools twice. If we count only unique students, 1,575 unique students took the student survey; 1,146 unique students took the PBTS; and 1,108 unique students took both the student survey and the PBTS.

Exhibit 3.1. Outcome Data Sources and Sample Sizes

Data Source	Description	Sample	Number of Schools	Number of Participants	Analytic Sample Response Rate
OECD PISA-Based Test for Schools	Measures students' higher-order skills in reading, mathematics, and science	Students in Cohorts 3 and 4 with parental consent, who were subsampled for data collection, and who were in school during 2013 data collection	20 schools, 10 school pairs	1,267	61% overall 74% network students 54% non-network students
Achievement Test Score Data From District Data System	Measures students' performance on state-mandated achievement tests in English Language Arts (ELA) and mathematics	Students in Cohorts 1–4, all students in propensity score strata	ELA: 24 schools, 13 school pairs Mathematics: 23 schools, 12 school pairs	ELA: 14,343 Mathematics: 14,187	ELA: 82% overall, 82% network students, 82% non-network students Mathematics: 87% overall, 87% network students, 86% non-network students
Student Survey	Measures students' self-reported opportunities to engage in deeper learning, as well as interpersonal and intrapersonal outcomes (such as self-efficacy)	Students in Cohorts 3 and 4 with parental consent, who were subsampled for data collection, and who were in school during 2013 data collection	22 schools, 11 school pairs	1,762	76% overall 80% network students 73% non-network students
Graduation Data From District Data System	Measures students' graduation from a high school within the district within four years of entering Grade 9	Students in Cohorts 1–3, all students in propensity score strata	24 schools, 13 school pairs	13,831	N/A
Postsecondary Enrollment From the National Student Clearinghouse	Measures students' postsecondary enrollment and the types of postsecondary institutions attended	Students in Cohorts 1–3, all students in propensity score strata	22 schools, 11 school pairs	11,165	N/A

Note: High school graduation and postsecondary enrollment outcomes are measured among all students in Cohorts 1–3 who have propensity scores within the region of common support, and so the "N/A" in the last column indicates that these outcomes do not have attrition or non-response rates.

A. OECD PISA-Based Test for Schools (PBTS)

In accordance with the Hewlett Foundation's preference for using an off-the-shelf (rather than custom-made) assessment to compare student achievement in network and non-network schools, we considered three published assessments designed to measure outcomes aligned with deeper learning objectives: the College and Work Readiness Assessment (CWRA), the College Learning Assessment (CLA), and the OECD PISA-Based Test for Schools (PBTS). The CWRA and the CLA were eliminated from consideration because their assessment tasks are not designed to systematically measure core academic content knowledge. Further, the CLA was designed for college rather than high school students, and the CWRA was already used by some network schools and therefore would not allow for a fair comparison between students at network and non-network schools. We selected the PBTS because it includes a large number of test items focused on knowledge and application of core academic subjects at the high-school level, and because it would allow participating schools the opportunity to compare their performance to well-established international benchmarks.

Although the PBTS is designed to facilitate comparisons among 15-year-old students worldwide, we used it to compare the performance of students in Grades 11 and 12 (Cohorts 3 and 4), who were generally older than 15. The PBTS was administered to students whose parents consented to their participation. The sample was restricted to students who had been enrolled in their school since Grade 9, were enrolled as a student in Grade 11 or 12 during the winter/spring 2013 test administration, and had been sampled for primary data collection.[4]

Tests were administered by CTB/McGraw-Hill LLC (CTB) using test administrators trained in CTB testing procedures. In preparation for testing, the CTB testing coordinator worked with school staff to schedule the PBTS administration for dates on which the test takers were not expected to be taking other tests or to be unavailable for other reasons. In advance of the testing day, the CTB testing coordinator reviewed the list of sampled students with the school coordinator (the study's contact at the school) to identify students unavailable for testing either because they were no longer enrolled at the school or because school staff had determined that the extent of their special needs limited the utility of their participation in the test. The testing coordinator recorded the reasons for non-participation. If fewer than 70 percent of the targeted students participated in the initial test administration, one or two make-up sessions were scheduled.[5]

Testing sessions consisted of two 60-minute periods during which students responded to test items, with a five-minute rest break after the first hour. Test administration procedures—for example, the spacing and placement of test takers' seats; the distribution and labeling of test

[4] As a service to 15 of the participating schools, CTB also administered the PBTS to a sample of 15-year-old students to allow benchmarking of their performance relative to the performance of the PISA worldwide sample of 15-year-old students. While the study sample and the 15-year-old student sample were usually tested at the same time, results for the 15-year-old student sample are not discussed in this report. CTB analyzed the 15-year-old students' data and delivered reports directly to the schools.

[5] PBTS administration took place close to the end of the school year and some make-up testing sessions were therefore not well attended.

booklets; control of entry into and exit from the testing room; prohibitions on talking, using cell phones, and leaving the test area with any testing materials; and proctoring—were designed to maximize the security of test items and minimize interruptions and distractions during testing. All testing was conducted in English and no testing accommodations were offered.

The PBTS is designed to produce estimates of school-level performance, rather than the performance of individual students. The test follows an incomplete block design in which each student takes a test containing a fraction of the total PBTS item bank. The full PBTS item bank consists of 141 items in reading/English Language Arts (ELA), mathematics, and science.[6] The items are grouped into seven blocks. There are seven versions of the test, each containing three of these item blocks. Item blocks are spiraled through the seven forms. Each test form contains items in two or three subject areas.[7] Test administration procedures were designed to ensure that each test form was assigned randomly to roughly the same number of students stratified by grade and gender in each school.[8]

CTB estimated student scale scores and standard errors for each subject area. Scale scores were based on maximum likelihood estimates from a unidimensional item response model for each of the three subjects (reading, mathematics, and science). For analysis, these scores (originally in logits) were standardized based on the weighted comparison group mean and standard deviation in order to interpret results as effect sizes. If a student's test form did not include any items within a subject area (e.g., one test form included items in mathematics and science but did not contain any items in reading), the student was assigned a missing test score within that subject area and was excluded from analyses of that subject. Since test forms were distributed randomly, this type of missing data does not bias our results.

We excluded students identified as leaving the test administration early and completing less than 75 percent of the test items because we concluded that they did not fully participate in the test administration. A total of 52 students (4 percent of students who took the test) were removed from the sample for this reason. These students were classified as non-respondents and were included in the calculation of non-response weights.

---

[6] In each subject area, some items are multiple choice and others require a short or long constructed response.

[7] Of the seven test forms, four included items in each of the three subject areas (mathematics, science, and reading). One test form only contained items in mathematics and reading, one test form only contained items in mathematics and science, and one test form only contained items in science and reading. The number of items within each subject also varied across test forms.

[8] See the OECD PISA-Based Test for Schools website (http://www.oecd.org/pisa/aboutpisa/pisa-basedtestforschools.htm) for more information.

Exhibit 3.2. Descriptive Statistics for Maximum-Likelihood Estimates of Scores on the PBTS (Unweighted)

		Maximum Likelihood Estimates (Logits)				
	N	Mean	SD	Min	Max	ICC
Reading	1,079	-0.10	1.40	-5.32	5.20	0.20
Mathematics	1,082	-0.97	1.65	-5.98	4.97	0.22
Science	1,085	-0.34	1.19	-5.52	5.31	0.19

B. High School Achievement Test Scores

To measure the impact of attending a deeper learning network school on students' achievement test scores, we collected high school achievement test scores from New York City and the participating districts in California. In both New York City and California, students' achievement was measured using the score students received when they first took the mathematics and English Language Arts (ELA) tests within the first three years of high school. In California, all students took the test in either their second or third year of high school. In New York City, students may have also taken the test during Grade 9. Dummy variables were included in analysis models to account for the year (i.e., the first, second, or third year of high school) in which students took the test.

In California, we examined students' scores on the California High School Exit Exam (CAHSEE), which students take in Grade 10. The CAHSEE mathematics test largely measures students' knowledge of pre-algebra and the first year of algebra, while the ELA test measures students' content knowledge of ELA through Grade 10. In New York City, we examined students' test scores on the Integrated Algebra and Comprehensive English Regents tests. While Comprehensive English is the only ELA Regents test, students may take mathematics Regents tests in many subjects. We selected Integrated Algebra as the single mathematics Regents exam for analysis because (1) more students in our sample took this Regents exam than any other mathematics Regents exam; (2) Integrated Algebra is a lower level mathematics test, and therefore using scores on the Integrated Algebra test does not require us to exclude students who did not take higher level mathematics courses; and (3) approximately half of the mathematics CAHSEE exam assesses students' knowledge of algebra, meaning that the Integrated Algebra and mathematics CAHSEE exams are similar in terms of content. Achievement test scores in mathematics were not examined for one school pair in New York City because students who attended the network school in this pair were not required to take mathematics Regents tests.

Since not all students in our sample who attended schools in California persisted in the district until Grade 10, and not all students in New York City remained enrolled in the district for three consecutive years, attrition weights were applied to analyses of test score data. In addition, not all students who remained in the district had test scores in the district administrative records, and non-response weights were therefore applied to analyses of achievement test scores. These weights are described in detail in Section IV.A.

Achievement test scores in mathematics and ELA were standardized prior to data analysis so that results could be compared across states and interpreted as effect sizes. To standardize test scores in California, we used statewide means and standard deviations for the CAHSEE mathematics and ELA scores within the academic years that students in our sample took the tests.[9] In New York City, the year-specific means and standard deviations were calculated across all schools in New York City (rather than the state).

C. Student Survey

As part of the student survey development process, the survey was piloted in six network schools in spring 2012. To test the reliability of survey constructs and the survey administration processes, we subsampled 30 consented students from each of the high school grades to take the student survey. Items were added, dropped, or reworded based on findings from the pilot.

As part of the full study, student surveys were administered in spring 2013, when respondents were expected to be in Grades 11 and 12. At most schools, surveys were administered by members of the research team.[10] All schools were given the option of administering an online survey; paper surveys were administered in 18 schools and students took online surveys in four schools. The student survey included items (listed below) that measured opportunities to experience instruction focused on different dimensions of deeper learning and the competencies expected to result from exposure to deeper learning.

Each survey item had four response options. For example, the items that measured opportunities to learn had the following response options: none of my classes within the academic year (coded 0); one of my classes within the academic year (coded 1); two of my classes within the academic year (coded 2); and three or more of my classes within the academic year (coded 3). We estimated construct scores from the item-level responses with an ordered logit Rasch model (Yen, 1986), implemented with the WINSTEPS software package. The resulting Rasch scale scores are in the logit metric and have both negative and positive values. The value of zero is anchored to the average difficulty of the items included in the scale. In general, a student with a positive score tended to respond favorably (i.e., choosing the highest or second highest response option) on average, and a student with a negative score tended to respond negatively (i.e., choosing the lowest or second lowest response option) on average. The sample on which we calculated Rasch scores for each scale was restricted to students with missing data for no more than half of the items within the scale. Less than 5 percent of students within each school had missing data on each of the scales, with the exception of one non-network school, in which a technological glitch during survey administration caused all items from the first half of the survey to be deleted.[11] For the scales that were affected by this technological glitch, we excluded the school pair from the main analyses.

---

[9] http://www.cde.ca.gov/ta/tg/hs/resources.asp

[10] There was one school in which AIR staff were not present for survey administration due to scheduling issues. In addition, students in two schools who were not present for the first survey administration were asked to complete the online survey on their own time; AIR staff were not present for these makeup sessions.

[11] In one of the four schools in which the survey was administered online, a computer glitch deleted students' responses to the first half of the survey as soon as they advanced to the second half of the survey. While we

Exhibit 3.3 presents the overall mean, standard deviation, and intra-class correlation (ICC), by construct, for the Rasch scale score. The exhibit also reports the Rasch scores transformed into the original 0–3 response metric. The transformation to the 0–3 metric was based on the threshold parameters from the WINSTEPS output for each construct and the Rasch scale score for each individual. For example, the mean Rasch score of 0.78 for "academic engagement" is approximately equivalent to an average response of 2 (agree) on the survey response scale (from strongly disagree to strongly agree). Rasch scores above 0.78 imply stronger agreement with the academic engagement survey items and scores below 0.78 imply less agreement with the survey items. Throughout the report, we present differences between network and non-network students as standardized versions of the Rasch scale scores to facilitate interpretation of estimates in effect sizes. The scores were converted to *z*-scores based on the weighted comparison group mean and standard deviation.

We also report the ICC—which is the ratio of between-school variance to total variance for a given construct—in Exhibit 3.3. Higher values mean more variation between schools, and lower values mean that more of the variation was among students within each school. We expected constructs designed to be more "school-centric" (e.g., assessments aligned with deeper learning) to have higher ICCs than constructs designed to be more "student-centric" (e.g., perseverance).

Due to the large number of constructs measured in the student survey, we were concerned that our findings might be affected by the fact that we were making multiple comparisons with similar outcome measures. While some of the measures pertain to opportunities for different dimensions of deeper learning opportunities (e.g., opportunities for complex problem solving, opportunities for communication), other constructs are more similar in nature (e.g., academic engagement and motivation to learn). To ensure that using multiple measures of similar outcomes was not leading us to draw false conclusions about the impact of attending network schools, we performed qualifying tests. For these qualifying tests, we examined the impact of attending a network school on a composite measure based on multiple individual constructs. Results for the individual survey constructs were deemed significant only if the coefficient for both the qualifying test and the individual construct were statistically significant. Overall, three new composite measures were created to perform qualifying tests:

- Qualifying test for five measures of opportunity to learn (OTL) that did not fit perfectly within predefined domains of deeper learning: the average of the opportunities for assessments aligned with deeper learning, opportunities to receive feedback, opportunities for creative thinking, opportunities for interdisciplinary learning, and opportunities for real-world connections constructs
- Qualifying test for engagement/motivation: the average of the academic engagement and motivation to learn constructs
- Qualifying test for efficacy/locus of control: the average of the self-efficacy and locus of control constructs

corrected the computer issue and asked students to retake the student survey, only a small number of students retook the survey.

Exhibit 3.3. Descriptive Statistics for Rasch-Scaled Student Survey Constructs

		Rasch Logit Scale			Response Scale		
	N	**Mean**	**SD**	**ICC**	**Mean**	**SD**	**ICC**
Creative Thinking Skills (Degree of Truth Scale – Never to Always True)	1,672	1.77	2.30	0.01	2.01	0.72	0.01
Collaboration Skills (Degree of Truth Scale – Never to Always True)	1,676	2.19	2.08	0.05	2.23	0.71	0.05
Academic Engagement (Agreement Scale)	1,680	0.78	1.17	0.19	1.97	0.45	0.11
Motivation to Learn (Degree of Truth Scale – Never to Always True)	1,677	1.57	2.09	0.09	2.01	0.71	0.08
Self-Efficacy (Degree of Truth Scale – Never to Always True)	1,740	2.49	2.62	0.01	2.15	0.70	0.01
Locus of Control (Degree of Truth Scale – Never to Always True)	1,740	2.16	2.20	0.01	2.18	0.68	0.01
Perseverance (Degree of Truth Scale – Never to Always True)	1,673	2.59	2.61	0.01	2.18	0.74	0.01
Self-Management (Degree of Truth Scale – Never to Always True)	1,679	0.75	1.53	0.03	1.90	0.73	0.04

Note: A value of zero on the Rasch logit scale approximates the level at which students are equally likely to respond to items with a 1 (or below) or a 2 (or above) on the 0 to 3 scale. A positive mean value indicates that a larger percentage of students responded to items with values of 2 or 3, while a negative mean value indicates that a larger percentage of students responded to items with values of 0 or 1.

Detailed Description of Survey Constructs

Student Interpersonal and Intrapersonal Competency Outcomes

Academic Engagement

(Source: CCSR and Academic Engagement Scale – Behavioral Subscale)
Rasch reliability = .74; Cronbach's alpha = .77

Regarding your core academic classes (English, math, science, and social studies) this year, to what extent do you agree with the following statements?

	Strongly Disagree	Disagree	Agree	Strongly Agree
CCSR – Academic Engagement				
The topics we are studying are interesting and challenging.	☐	☐	☐	☐
I am usually bored by classes or activities.	☐	☐	☐	☐
I usually look forward to classes or activities.	☐	☐	☐	☐
Sometimes I get so interested in my work I don't want to stop.	☐	☐	☐	☐
I often count the minutes until class ends.	☐	☐	☐	☐
Academic Engagement Scale – Behavioral Subscale				
I always prepare for class.	☐	☐	☐	☐
I ask questions when I don't understand the lesson.	☐	☐	☐	☐
I actively participate in group activities.	☐	☐	☐	☐
I am usually distracted by my classmates.	☐	☐	☐	☐
I cut class when I'm bored.	☐	☐	☐	☐

Collaboration Skills

(Source: Original items, National Center for Research on Evaluation, Standards, & Student Testing [CRESST] – Personal Interaction Scale [Huang et al., 2010])
Rasch reliability = .83; Cronbach's alpha = .91

Now think about the group work you do for your classes. How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
When I work with a group, I tell the other members of my group when I think they are doing a good job. (*CRESST*)	☐	☐	☐	☐
When I work with a group, I make sure to be prepared and bring needed materials.	☐	☐	☐	☐
When I work with a group, I remember to do my part of a group project without being reminded.	☐	☐	☐	☐
When I work with a group, I finish my part of a group project on time.	☐	☐	☐	☐
When I work with a group, I help keep my group focused.	☐	☐	☐	☐

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
When I work with a group, I share my ideas with the group.	☐	☐	☐	☐
When I work with a group, I help my group figure out and fix any problems we face.	☐	☐	☐	☐
When I work with a group, I pay attention when my teammates talk.	☐	☐	☐	☐
When I work with a group, I consider everyone's ideas.	☐	☐	☐	☐
When I work with a group, I learn from other people in my group.	☐	☐	☐	☐

Creative Thinking

(Source: Original)
Rasch reliability: .77; Cronbach's Alpha: .84

How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
I am able to come up with new and different ideas.	☐	☐	☐	☐
I like to think of original solutions to problems.	☐	☐	☐	☐
I come up with new ways to do things.	☐	☐	☐	☐
I am an original thinker.	☐	☐	☐	☐
I have a better imagination than my friends.	☐	☐	☐	☐

Perseverance

(Source: Duckworth and Quinn's (2009) Perseverance of Effort scale, unless otherwise noted)
Rasch reliability = .79; Cronbach's alpha = .88

How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
I overcome setbacks to achieve important goals.	☐	☐	☐	☐
I am a hard worker.	☐	☐	☐	☐
I finish what I begin.	☐	☐	☐	☐
I achieve goals even if they take a long time.	☐	☐	☐	☐
I do a careful and thorough job. *(Original)*	☐	☐	☐	☐

Locus of Control

(Source: Levenson's (1981) Locus of Control construct).
Rasch reliability = .73; Cronbach's Alpha: .83

How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
I believe that whether or not I get to be a leader depends mostly on my ability.	☐	☐	☐	☐
When I make plans, I am almost certain to make them work.	☐	☐	☐	☐
I believe that I can pretty much determine what will happen in my life.	☐	☐	☐	☐
I believe that when I get what I want, it's usually because I worked hard for it.	☐	☐	☐	☐
I believe that my life is determined by my own actions.	☐	☐	☐	☐

Motivation to Learn

(Source: Pintrich and DeGroot's (1990) Motivated Strategies for Learning Questionnaire [MSLQ])
Rasch Reliability: .75; Cronbach's Alpha: .81

Think about the work you are doing in your classes this year. How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
It is important for me to learn what is being taught in my classes.	☐	☐	☐	☐
I think that what I am learning in my classes is useful for me to know.	☐	☐	☐	☐
I think what I am learning in my classes is interesting.	☐	☐	☐	☐
I prefer class work that is challenging so I can learn new things.	☐	☐	☐	☐
I try to learn from my mistakes in my schoolwork.	☐	☐	☐	☐

Self-Management

(Source: Student Culture of Excellence and Ethics Assessment Survey[12] *[CEEA] of High and Middle Schools, Xue and Sun's [2011] Self-Management Scale, College Student Experiences Questionnaire*[13] *[CSEQ])*
Rasch reliability = .81; Cronbach's alpha = .85

How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
I put off doing things that I don't like to do. *(CEEA)*	☐	☐	☐	☐
I set goals for doing better in school. *(CEEA)*	☐	☐	☐	☐
I make a to-do list every day. *(Xue and Sun)*	☐	☐	☐	☐
I make schedules to help myself finish tasks on time. *(Xue and Sun)*	☐	☐	☐	☐
I finish my tasks on time. *(Xue and Sun)*	☐	☐	☐	☐
I get all the help I can to help me reach my goals. *(Xue and Sun)*	☐	☐	☐	☐
I set long-term goals for myself. *(Xue and Sun)*	☐	☐	☐	☐
I can find the information I need to learn on my own. *(CSEQ)*	☐	☐	☐	☐
I feel good about my ability to learn whatever I want or need to know. *(CSEQ)*	☐	☐	☐	☐
I can learn effectively on my own. *(CSEQ)*	☐	☐	☐	☐
I feel like I am in charge of what I learn. *(CSEQ)*	☐	☐	☐	☐

[12] http://www.excellenceandethics.com/assess/CEEA_v4.5_matrix.pdf

[13] http://cseq.iub.edu/pdf/cseq_whole.pdf

Self-Efficacy

(Source: New General Self-Efficacy Scale by Chen, Gully, and Eden, 2001)
Rasch reliability = .84; Cronbach's alpha = .91

How often are the following statements true about you?

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
I believe I will be able to reach my goals.	☐	☐	☐	☐
I know I can complete difficult tasks.	☐	☐	☐	☐
I believe I can do whatever I decide to do.	☐	☐	☐	☐
I believe I will be able to overcome challenges.	☐	☐	☐	☐
I know I can do many different things well.	☐	☐	☐	☐
Compared to most other people, I can do most tasks very well.	☐	☐	☐	☐
Even when things are tough, I can perform quite well.	☐	☐	☐	☐

D. High School Graduation Data Received From School Districts

We obtained information on high school graduation status from the participating school districts in fall 2013. We defined high school graduation as graduation within the first four years after entry into Grade 9. As a result, all students in Cohorts 1 to 3 had sufficient time to graduate from high school on time. We did not collect data on high school graduation for students who left the district prior to graduation.

For our measure of on-time graduation within the same district, we classified all students as either graduates or non-graduates. Non-graduates included students who:

- Dropped out of school
- Left the school district prior to graduation
- Were still enrolled in the district at the time of data collection (fall 2013) and thus may have graduated in more than four years

Among students in Cohorts 1 to 3 (who entered Grade 9 between 2007–08 and 2009–10), approximately 62 percent graduated from a high school in the same district within four years.

Our definitions and rates are not comparable to traditional graduation rates based on aggregate data (such as the Averaged Freshman Graduation Rate) because those rates do not use longitudinal data to track student cohorts and transfers.[14] Additionally, our definition of

[14] For details on the NCES Averaged Freshman Graduation Rate, see: http://nces.ed.gov/pubs2011/dropout08/app_a3.asp

graduates and non-graduates does not directly correspond to official definitions used by California and New York City for graduation rate reporting at the school level. Both California and New York City use a four-year cohort graduation rate, where the numerator (graduates) consists of students who earn a high school diploma in their fourth year of high school from a given school, and the denominator (graduates + non-graduates) consists of first-time Grade 9 students in the cohort year, plus students who transferred into the cohort in the appropriate grade/year, minus students who transferred out of the school or met other discharge conditions (e.g., deceased or incarcerated).[15]

Like the traditional graduation rate definitions, our approach also uses a four-year cohort definition, but because our research design focuses on students who entered the network and non-network schools in Grade 9, we do not include students who transferred into the school in later grades. Our main graduation analysis also does not exclude students who transferred out of the school/district. Since attending a network school may affect rates of student transfer (i.e., exposure to instruction focused on deeper learning may discourage students from transferring to a different high school), excluding transfer students from analyses may bias the estimated impact of attending a deeper learning network school on high school graduation. As described in Section IV.B, we conducted sensitivity analyses examining alternative definitions of high school graduation that excluded transfer students. We also examined the impact of attending a deeper learning network school on graduation within five years of entering high school.

E. Postsecondary Outcome Data

We collected information on students' postsecondary enrollment outcomes from the National Student Clearinghouse. The National Student Clearinghouse (NSC) is a non-profit organization that collects student-level enrollment and degree completion information from postsecondary institutions in the United States. More than 3,600 institutions submit data to the NSC, accounting for approximately 98 percent of all students in postsecondary education.[16] Working closely with the districts participating in the study, we requested postsecondary data for students in Cohorts 1 to 3 within three districts, for a total of 11 pairs of schools. We requested data for all students who entered Grade 9 within our selected schools, including those who were not observed to graduate from high school within the district. As such, analyses include students who may have transferred to another district prior to graduating from high school and enrolling in college.

Postsecondary enrollment data were collected in fall 2013, when students in Cohort 1 were entering their third year of college (if they had progressed on time and enrolled in postsecondary education immediately after high school graduation) and students in Cohort 3 were entering their first year of college. Using the NSC data, the following postsecondary enrollment outcomes were measured:

[15] For details on the California graduation rate definition, see: http://dq.cde.ca.gov/dataquest/CohortRates/CohortOutcomeDefinitions2012_4_30.doc. For details on the New York City graduation rate definition, see: http://schools.nyc.gov/NR/rdonlyres/BD3585E6-B686-43F2-97F2-8F0EA3BF71FD/0/EducatorGuide_HS_11_25_2013.pdf

[16] http://www.studentclearinghouse.org/about/clearinghouse_facts.php

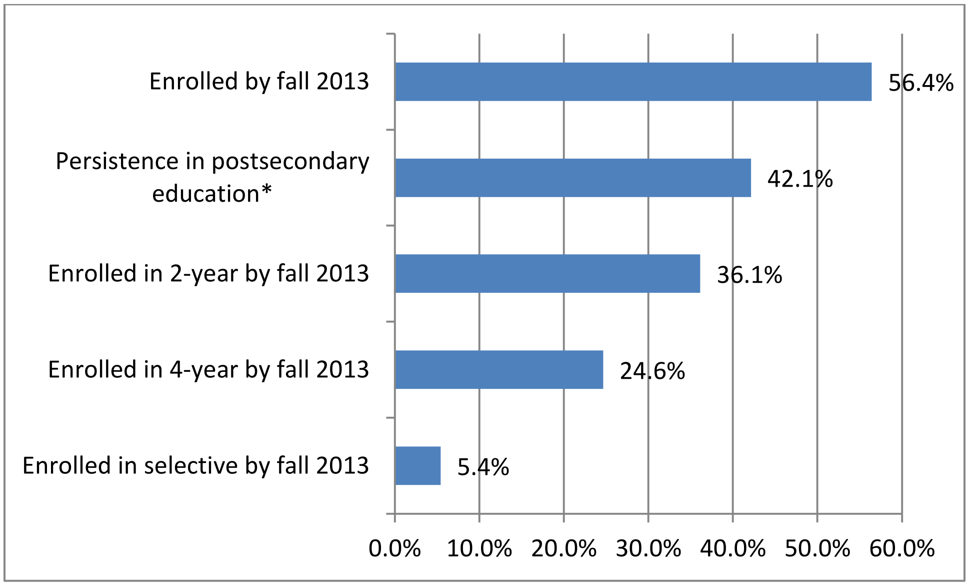
- Ever enrolled in a postsecondary institution by fall 2013
- Enrolled in a two-year institution by fall 2013
- Enrolled in a four-year institution by fall 2013
- Enrolled in a selective institution (defined below) by fall 2013
- Postsecondary persistence: enrolled in postsecondary education in the year following expected high school graduation and continued enrollment in the fall of the next year (only measured for Cohorts 1 and 2)

To define the selectivity of institutions, we drew on the definition used in the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS). IPEDS defines selective institutions as four-year institutions in which at least 80 percent of students are full-time students and the admissions requirements are defined as “more selective.”[17] Institutions are classified as “more selective” in the IPEDS data if the test scores of first-year students place the institution in the top 20 percent of institutions in the United States. One pair of schools was excluded from analyses of enrollment in selective institutions because no students attending the network school enrolled in a selective institution.

Postsecondary persistence was analyzed only for students in Cohorts 1 and 2 because students in Cohort 3 had not yet had the opportunity to enter the second year of college by fall 2013 (when the data were collected). Students were identified as persisting in postsecondary education if they enrolled in any postsecondary institution during the school year immediately following their expected high school graduation (August 2011–June 2012 for Cohort 1, August 2012–June 2013 for Cohort 2) *and also* enrolled in postsecondary education in the fall of the subsequent academic year (representing continuation in postsecondary education). All other students in Cohorts 1 and 2—including students who never enrolled in postsecondary education—were classified as not persisting in postsecondary education. Exhibit 3.4 provides the unweighted descriptive statistics for these five postsecondary enrollment outcomes for the students in our sample.

[17] The selectivity of postsecondary institutions is measured in the IPEDS data with the data field “CCUGPROF.” We classified institutions as selective if they had IPEDS codes of 12 (full-time, four-year, more selective, low transfer-in) or 13 (full-time, four-year, more selective, high transfer-in).

Exhibit 3.4. Descriptive Statistics for Postsecondary Outcomes (Unweighted)



* Persistence in postsecondary education is measured among students who entered Grade 9 in 2007–08 or 2008–09. All other postsecondary outcomes are measured among students who entered Grade 9 in 2007–08, 2008–09, or 2009–10.

Note: By fall 2013, it was possible that students had enrolled in multiple postsecondary institutions. As such, the rates of enrollment in two-year and four-year institutions are not mutually exclusive, and the sum of these rates do not equal the total percentage of students who ever enrolled in postsecondary education.

F. Student Background Data (Extant Data)

We obtained student-level administrative records from the participating districts containing data on student characteristics measured in Grade 8 and Grade 9. We used the record data to identify students to be included in our samples (i.e., first-time Grade 9 students) and to incorporate covariates in our analyses. Our study schools were located in multiple school districts, so consistent data was not available for all study schools. However, since school pairs were constructed within a district, we had the same set of student background characteristics for the two schools in any given pair.[18] Exhibit 3.5 lists the student background data we received from districts and details how many school pairs had each data element. As the exhibit indicates, we had two measures of student socioeconomic background: parents' education and students' free or reduced-price lunch status. For 13 of the 15 pairs, we had one of these proxies for socioeconomic status, and we received information for both indicators from the remaining two pairs. Only New York City provided data on Grade 8 attendance and students' age at Grade 9 entry.

A description of the students who attended network and non-network high schools in California and New York City for Cohorts 3 and 4—the cohorts that participated in primary data collection (i.e., the student survey and the PBTS)—is presented in Exhibit 3.6. Similarly, Exhibit 3.7 presents the descriptive statistics for Cohorts 1 to 3—the cohorts included in analyses of high

[18] One pair of schools contained a network and a non-network school in neighboring districts. The data elements available across the two districts were very similar.

school graduation and postsecondary enrollment outcomes.[19] The descriptive statistics represent all students within these cohorts, before adjustments were made for differences between students who attended network and non-network schools. As discussed in Section IV, student background characteristics were used in the estimation of weights and as covariates in analytic models.

Exhibit 3.5. Description of Student Background Data From Extant District Data

Measure	Description	Number of School Pairs With Available Data
Female	Dichotomous indicator of students' gender	15
Race/Ethnicity	Dichotomous indicators created for African American, Hispanic, white, Asian, and "other" races	15
Parents' Education	Categorical measure of parental education—specifically, the highest level of education obtained by either parent—using the following categories: some high school, high school diploma, some college, college degree, higher degree (above BA), and declined to report parents' education (varies slightly by district)	6
FRPL Status	Dichotomous indicator of whether student was eligible for the free or reduced-price lunch program, typically in Grade 8	9
English Language Learner (ELL)	Dichotomous indicator of whether the student was identified as an English language learner, typically in Grade 8	15
Individualized Education Plan (IEP)	Dichotomous indicator of whether the student had an Individualized Education Plan, typically in Grade 8	15
Prior Achievement in ELA	Standardized test score in English Language Arts (ELA) prior to entering high school, from Grade 8	13
Prior Achievement in Mathematics	Standardized test score in mathematics prior to entering high school, from Grade 8, including indicators for math test subject where relevant; standardized using the state mean and standard deviation for each year and grade level	13
Grade 8 Attendance Rate	Proportion of enrolled school days attended during Grade 8	7
Age	Age of student (in months) when first enrolled in Grade 9	7

[19] While descriptive statistics are not presented for the sample of students used for the analysis of high school achievement test scores, the only difference between this sample and the sample presented in Exhibit 3.7 is the inclusion of Cohort 4. As such, the characteristics presented in Exhibit 3.7 roughly approximate the sample used for the analysis of high school achievement test scores.

Exhibit 3.6. Descriptive Statistics for the Cohort 3 and Cohort 4 Study Sample in California and New York City, by Treatment Status

	California				New York City			
	Network		Non-Network		Network		Non-Network	
	N With Data	%	N With Data	%	N With Data	%	N With Data	%
Gender								
Female	1,061	52.6%	6,791	49.7%	556	57.6%	1164	50.8%
Race/Ethnicity								
White	1,061	17.3%	6,791	25.5%	556	26.1%	1164	8.8%
Black	1,061	13.9%	6,791	13.0%	556	20.0%	1164	9.8%
Hispanic	1,061	61.5%	6,791	46.5%	556	49.1%	1164	79.1%
Asian/Other	1,061	7.2%	6,791	15.0%	556	4.9%	1164	2.2%
Parents' Education								
Less than High School	865	29.0%	6,063	21.4%		N/A		N/A
High School Diploma	865	20.8%	6,063	22.8%		N/A		N/A
Some College	865	14.3%	6,063	19.9%		N/A		N/A
College Degree	865	18.3%	6,063	24.0%		N/A		N/A
Declined/Missing	865	17.6%	6,063	11.9%		N/A		N/A
Background								
FRPL Status	301	69.8%	1,916	53.0%	556	58.5%	1164	71.7%
ELL Status	1,061	23.2%	6,791	19.8%	556	36.2%	1164	55.4%
IEP Status	1,061	8.1%	6,791	9.3%	556	2.9%	1164	0.8%
Grade Level								
Grade 12	1,061	48.7%	6,791	50.1%	556	53.2%	1164	49.2%
	N With Data	**Mean**	**N With Data**	**Mean**	**N With Data**	**Mean**	**N With Data**	**Mean**
Prior Test Scores (Standardized)								
Grade 8 Math	1,061	-0.166	6,791	0.026	351[a]	0.117	354[a]	-0.116
Grade 8 ELA	1,061	-0.121	6,791	0.019	351[a]	0.012	354[a]	-0.012
Attendance Rate		N/A		N/A	351[a]	91.9%	354[a]	90.7
Age at Entry to Grade 9		N/A		N/A	556	14.9	1164	15.2
Total Sample	1,061		6,791					

Note: While the analyses of the effects of attending network schools were performed within pairs of schools, the descriptive statistics in this table are measured at the student level. As a result, larger schools implicitly received more weight.

[a] Prior achievement test score data and Grade 8 attendance rate data were unavailable for students in Pair 10 and Pair 11. The network and non-network schools within these pairs serve large populations of immigrant students and a substantial number of students within these pairs were not in the district or were not required to take the state assessment prior to Grade 9.

Exhibit 3.7. Descriptive Statistics for the Cohort 1, Cohort 2, and Cohort 3 Study Sample in California and New York City, by Treatment Status

	California				New York City			
	Network		Non-Network		Network		Non-Network	
	N With Data	%	N With Data	%	N With Data	%	N With Data	%
Gender								
Female	1,563	53.0%	11,147	47.7%	1,775	58.5%	2,528	49.7%
Race/Ethnicity								
White	1,563	14.4%	11,147	23.8%	1,775	12.0%	2,528	6.2%
Black	1,563	18.0%	11,147	17.2%	1,775	41.5%	2,528	26.0%
Hispanic	1,563	60.4%	11,147	44.4%	1,775	43.2%	2,528	65.9%
Asian/Other	1,563	7.2%	11,147	14.7%	1,775	3.3%	2,528	1.9%
Parents' Education								
Less than High School	1,298	25.2%	10,057	16.9%		N/A		N/A
High School Diploma	1,298	17.9%	10,057	20.2%		N/A		N/A
Some College	1,298	11.8%	10,057	16.8%		N/A		N/A
College Degree	1,298	14.7%	10,057	21.3%		N/A		N/A
Declined/Missing	1,298	30.4%	10,057	24.8%		N/A		N/A
Background								
FRPL Status	390	71.0%	2,792	50.6%	1,775	69.7%	2,528	74.8%
ELL Status	1,563	22.5%	11,147	18.2%	1,775	23.4%	2,528	37.0%
IEP Status	1,563	6.8%	11,147	9.2%	1,775	4.8%	2,528	2.1%
	N With Data	**Mean**	**N With Data**	**Mean**	**N With Data**	**Mean**	**N With Data**	**Mean**
Prior Test Scores (Standardized)								
Grade 8 Math	1,563	-0.163	11,147	0.023	1,399[a]	-0.796	1,318[a]	-0.816
Grade 8 ELA	1,563	-0.092	11,147	0.013	1,399[a]	-0.508	1,318[a]	-0.529
Attendance Rate		N/A		N/A	1,399[a]	88.5%	1,318[a]	89.1%
Age at Entry to Grade 9		N/A		N/A	1,775	14.9	2,528	15.1
Total Sample	1,563		11,147		1,775		2,528	

Note: While the analyses of the effects of attending network schools were performed within pairs of schools, the descriptive statistics in this table are measured at the student level. As a result, larger schools implicitly received more weight.

[a] Prior achievement test score data and Grade 8 attendance rate data were unavailable for students in Pair 10 and Pair 11. The network and non-network schools within these pairs serve large populations of immigrant students and a substantial number of students within these pairs were not in the district or were not required to take the state assessment prior to Grade 9.

IV. Analytic Methods

Because students were not randomly assigned to network and non-network schools, we cannot be sure that the students entering the two types of schools were equivalent on entry into Grade 9. We employed two strategies to take measured differences into account: weighting and covariate adjustment. We employed propensity score weighting to match the sample of students attending the non-network school in each pair as closely as possible to the sample of students attending the network school in the pair. We also used weights to reflect attrition between Grade 9 entry and the year of data collection, non-consent, subsampling of students in large non-network schools, and non-response. In addition, we used covariate adjustment to take any remaining differences between network and non-network students into account, and to improve the precision of the estimated effects.

These methods are described in the sections below.

A. Weighting

As described above, we applied weights to reflect two features of the study's design. First, we applied propensity score weights (Hirano et al., 2003) to account for measured pre-high school characteristics (including both demographic characteristics and Grade 8 achievement test scores) related to the decision to enroll in a deeper learning high school and likely related to student outcomes. Second, we applied attrition, sampling, and non-response inverse probability weights (IPW) to analyses so that results for the students from whom we collected data would be representative of the students who entered network and non-network schools in Grade 9. Inverse probability weights are commonly used to account for missing outcome data due to non-random attrition (Wooldridge, 2007; Ridgeway et al., 2013). Below, we discuss the four weights applied in the statistical analyses:

- The first weight accounts for differences in measured background characteristics associated with selection of a network or non-network school on entry to Grade 9.
- The second weight accounts for differences in the background characteristics of students who persisted in the same school between Grade 9 and the time of data collection *and* consented to participate in the study, and students who did not persist or consent.
- The third weight accounts for within-school variation in the probability of being selected for data collection among students who persisted and consented. This weight has a value of 1 for all network students (because we collected data from all consented network students), but this weight varies for non-network students because we sampled non-network students from propensity score strata in large schools.
- Finally, the fourth weight accounts for differences in the background characteristics of students who responded to the student survey among students sampled for data collection and those who failed to respond.

Analyses of survey data and scores on the PBTS were limited to students who were still attending the same school they entered in Grade 9 and who consented to participate in the study. In contrast, analyses of high school graduation and postsecondary data did not require student consent and did not require that students remain in the same school throughout high school. As a result (as Exhibit 4.1 demonstrates), different weights were applied to different analyses depending on the data source containing the outcome measure. While all analyses applied the first weight (to account for students' probability of enrolling in a network school), persistence and non-response weights were necessary only for specific outcomes. Specifically, because the timing and method of data collection differed across student outcomes, we calculated two persistence/consent weights—one for the survey/PBTS sample, and one for state achievement test scores. We also computed four non-response weights for the student survey, PBTS, and ELA and mathematics achievement test scores. Descriptive statistics for the weights used in the analyses of student survey data and PBTS scores are provided in Exhibit 4.2 and Exhibit 4.3.[20]

Exhibit 4.1. Definition of Weights for Different Outcome Analyses

	Weight 1: Student Selection Into Network Schools	Weight 2: Persistence and Consent	Weight 3: Sampling	Weight 4: Non-Response
Student Survey	*w*1	*w*2 (v1)	*w*3	*w*4 (v1)
PBTS	*w*1	*w*2 (v1)	*w*3	*w*4 (v2)
Achievement Test Scores From District	*w*1	*w*2 (v2)	N/A	*w*4 (v3) *w*4 (v4)
Graduation	*w*1	N/A	N/A	N/A
Postsecondary Data From NSC	*w*1	N/A	N/A	N/A

Notes:
*w*1 = student selection weight
*w*2 (v1) = student persistence/consent weight for survey and PBTS
*w*2 (v2) = student persistence weight for state achievement tests
*w*3 = student sampling weight
*w*4 (v1) = student response weight for survey outcomes
*w*4 (v2) = student response weight for PBTS
*w*4 (v3) = student response weight for state mathematics achievement test
*w*4 (v4) = student response weight for state ELA achievement test
N/A = no weight used

Weighting for Student Selection Into Network Schools

Students were not randomly assigned to network and non-network schools, so network and non-network students may not have had equivalent characteristics when entering high school. These preexisting student differences mean that any claims about a network school's effects on student experiences and outcomes could be biased if based on direct comparisons between network and non-network students. To account for these preexisting differences, we used inverse probability of treatment weighting (IPTW), which adjusts the comparison student sample to be more

[20] Descriptive information for the weights applied to analyses of high school graduation, postsecondary outcomes, and high school achievement test scores are available upon request. Due to issues of active parental consent and non-response associated with primary data collection, the weights documented in Exhibit 4.2 and Exhibit 4.3 are for the variables with the highest non-response rates.

representative of the network student sample based on measured student background characteristics. Assuming the measured student background characteristics accurately capture the important preexisting differences between network and non-network students, IPTW allows us to obtain valid estimates about what network students would have experienced if they had attended the non-network school.

IPTW is a propensity score-based method for selection bias adjustment (Hirano et al., 2003). A student's propensity score (p_i) is her or his predicted probability of attending a network school instead of a non-network school, given the measured student characteristics (X_i). To estimate propensity scores, we estimated separate logistic regression models for each school pair (j) and student cohort (k):

$$\ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_{0jk} + \beta_{1jk}X_{ijk},$$

where X_{ijk} represents the student characteristics listed in Exhibit 3.5 that were available for a given school pair.

The estimated propensity scores were then used to calculate IPTW weight for the non-network students, where a non-network student's weight equals the student's predicted odds of treatment assignment and a network student's weight equals one:

$$w1_{ijk} = T_{ijk} + \left(1 - T_{ijk}\right)\frac{p_{ijk}}{1-p_{ijk}},$$

where T_{ijk} equals 1 for students attending a network school and 0 for students attending a non-network school. With this weight, the comparison group was weighted to represent the network group to facilitate estimation of the average treatment effect on the treated (ATT). The IPTW weight used in this study had a value of 1 for all students attending a network school.

Weighting for Student Persistence and Consent

Our student survey analysis was designed to reflect the experiences of first-time Grade 9 students. However, we were not able to collect data on all entering Grade 9 students in the study cohorts because some students left the study schools prior to data collection (in their third or fourth year of high school) or because we were unable to obtain parental consent for data collection. On average, 62 percent of students in Cohorts 3 and 4 were still enrolled in the same school at the time of data collection, and 46 percent were both enrolled and provided consent to participate in the study (53 percent in network schools and 45 percent in non-network schools).

The sample of students who persisted and consented to data collection may of course differ in measured characteristics from the full sample of cohort students entering Grade 9. To account for this potential student attrition bias, we calculated an attrition weight based on the inverse probability that a student persisted in the same school from Grade 9 to the time of data collection and consented to participate in the study.[21]

We used generalized boosted regression (McCaffrey, Ridgeway & Morral, 2004) to estimate a student's probability of persisting and consenting for data collection. This method iteratively tries various combinations of student background covariates to predict the probability of persisting and consenting, searching for the combination that minimizes the differences in measured characteristics between students who persisted and those who did not, when the latter are weighted by the inverse probability of persisting and consenting. We used the *twang* package in the *R* statistical program to execute the generalized boosted regression. Following the recommendations set forth by the package authors (Ridgeway et al., 2013), we set the interaction depth to 4, shrinkage to 0.0005, and bagging to 0.50. A separate boosted regression was run for each school, with students in Cohorts 3 and 4 combined into one model for each school. Along with the student characteristics listed in Exhibit 3.5, a dichotomous indictor for cohort was included in the regression.

The estimated persistence and consent probabilities for student *i* in school j (pe_{ij}) were then used to calculate attrition weights:

$$w2_{ij} = \frac{1}{pe_{ij}}.$$

With this weight, eligible students were weighted to represent the cohorts entering Grade 9.

As shown in Exhibit 4.1, two different persistence/consent weights were calculated. First, to identify students to participate in the survey and PBTS data collections, we identified which students from Cohorts 3 and 4 were still attending the selected school in fall 2012 (i.e., when consent forms were distributed), and of those persisting students, which students consented to participate in the study. For these students, the same persistence/consent weight was applied to analyses of the student survey and PBTS data because students consented to participate in these two forms of data collection simultaneously.

A different persistence weight was applied to analyses of high school achievement test scores. While we did not need consent to obtain students' scores on the state assessments (because these data were provided by the participating school districts), we could not obtain achievement test scores for students who left the district prior to being eligible to take the exam. In California, students take the CAHSEE for the first time in the spring of their Grade 10 year. As a result, any student who left the district prior to entering Grade 10 was excluded, and persistence weights were calculated as the inverse probability of leaving the district prior to the Grade 10 year.[22] In

[21] Attrition weights and non-consent weights could not be calculated separately because some schools did not permit us to obtain identifying information for students who did not consent to participate in the study. It is for this reason that a single weight accounts for both attrition and non-consent.

[22] To apply consistent measurement rules across students in California and New York City, we also limited analyses to students in California who entered Grade 10 within the first three years after entering Grade 9.

New York City, students are not required to take Regents exams in a particular school year, but most students take the ELA and Integrated Algebra Regents exams within the first three years of high school. As a result, the persistence weights in New York City were computed as the inverse probability of leaving the district within the first three years of high school.

Weighting for Student Sampling

For student survey and PBTS data collection, we set a target survey sample size of 65 Grade 11 students (Cohort 4) and 65 Grade 12 students (Cohort 3) from each school. This target sample size was selected to provide sufficient power to detect effects of reasonable size, while minimizing burden and data collection costs. Because network schools were smaller in size, we administered the survey to all consented network students. In some network schools, fewer than 65 students within Grade 11 or Grade 12 consented to participate in the study. In order to collect data from a total of 260 students within each pair, we over-sampled non-network students within pairs in which the network school did not have 130 consented students in Grades 11 and 12. In small non-network schools (or non-network schools where only a small number of students consented to participate in the study), we also administered the survey to all consented students.

In large non-network schools with large numbers of consented students (such as large non-network schools with passive consent), we sampled a portion of consented students based on their propensity score strata (quintiles defined by the distribution of the matched network school). Once we observed the number of consented students from the matched network school within each stratum, we randomly sampled the same number of non-network students from each stratum. In addition, and in order to achieve the target of 260 completed student surveys for each matched pair, we randomly sampled the same number of additional non-network students from each stratum, so that the distribution of students across propensity score strata was preserved. Across the non-network schools, we sampled 39 percent of all consented students.

Since we subsampled students for the survey from propensity score strata, we calculated each student's probability of being sampled for the student survey (ps) based on the student's school, cohort, and propensity score stratum. In particular, in each school (j) where students were subsampled for survey data collection, we divided the number of students sampled (NS_{jkq}) within a specific cohort (k) and stratum (q) by the number of consented students (NC_{jkq}) within that cohort and stratum:

$$ps_{jkq} = \frac{NS_{jkq}}{NC_{jkq}}.$$

For students in schools where sampling was not necessary, including all network schools, ps_{ijkq} = 1. The sampling weight was applied only to analyses of student survey data and PBTS data. For the remaining data sources, we analyzed data for all students with propensity scores that fell within the range of common support.

Given the student's probability of sample selection, we calculated a sampling weight for each eligible student based on the inverse probability of sample selection:

$$w3_{jkq} = \frac{1}{ps_{jkq}}.$$

With this weight, students sampled for the student survey were weighted to represent the network school students who were eligible for survey administration (i.e., persisted in the school and consented to data collection). For students in a network school and students attending a small non-network school, the sampling weight was equal to 1. In the larger non-network schools, the sampling weight ranged from 1 to 15.07, with a mean of 2.38.

Weighting for Student Non-Response

To account for non-response in our analysis, we calculated a non-response weight based on the inverse probability that a student sampled for primary data collection participated in the student survey and the PBTS. Because the student survey and PBTS were not administered on the same day, there are separate non-response weights for analyses for these two data sources. In addition, non-response weights were calculated for analyses of state achievement test scores to account for missing test score data for students who were in the district long enough to have taken the test. Separate non-response weights were calculated for test scores in mathematics and test scores in ELA. The four non-response weights were calculated using methods similar to those described above for the calculation of persistence/consent weights. These weights can be interpreted as the inverse probability of having outcome data (from the student survey, PBTS, and high school achievement test scores in mathematics and ELA) among students who persisted in the school or district long enough to be eligible for the data collection (and were drawn into the sample, for the survey and PBTS).

The estimated response probabilities for student i in school j (pr_{ij}) were then used to calculate non-response weights for all students with outcome data:

$$w4_{ij} = \frac{1}{pr_{ij}}.$$

With this weight, students with non-missing outcome data were weighted to represent the target student sample.

Combined Analytic Weight

The four weights discussed above were combined into one weight that captured measured baseline differences between network and comparison students, as well as differences between student survey respondents and the target Grade 9 student cohorts. A convenient property of inverse-probability weighting is that different weights can be combined through multiplication (see, for example, Morgan and Todd, 2008). Therefore, each student's final analytic weight equals: $w1_{ijk} \times w2_{ij} \times w3_{jkq} \times w4_{ij}$. This weight represents the inverse of the combined probability of (1) being in a network school (p); (2) being eligible for data collection by persisting and consenting (pe); (3) being sampled for data collection among eligible students (ps | eligible); and (4) responding to the survey among sampled students (pr | sampled).

After calculating this analytic weight, we examined the distribution to identify outliers. We adjusted extreme weights to ensure that atypical cases did not disproportionately impact study results (Potter, 1988). Within each pair of schools, we calculated the mean of the analytic weight. We defined outlier weights as weights greater than three times the mean of the analytic weights within the school pair.[23] Overall, 39 students (2 percent of the sample) had outlying analytic weights, including 11 students at a network school and 28 students at a non-network school. To ensure that students with outlying weights did not disproportionately affect our analyses, we trimmed the analytic weights for these 39 students, setting the analytic weight equal to three times the pair-specific mean of the analytic weight. We incorporated this trimmed analytic weight into the analyses by using a survey design weight in our analytic models (discussed below). Exhibit 4.2 presents summary statistics for all of the individual weights as well as the final analytic weights (before and after trimming) for the survey sample. Exhibit 4.3 provides the same information for the weights applied to analyses of scores on the PBTS.

Exhibit 4.2. Descriptive Statistics for Individual and Combined Weights for Analyses of Student Survey Data, for Network and Non-Network Students

		N	Mean	S.D.	Min	Max
Weight 1: Weighting for School Selection	Network	687	1.00	0.00	1.00	1.00
	Non-Network	1075	0.35	0.42	0.00	4.05
Weight 2: Weighting for Attrition and Consent	Network	687	1.53	0.51	1.04	5.08
	Non-Network	1075	2.28	1.63	1.04	16.02
Weight 3: Weighting for Sampling	Network	687	1.00	0.00	1.00	1.00
	Non-Network	1075	2.38	2.55	1.00	15.07
Weight 4: Weighting for Non-Response	Network	687	1.20	0.21	1.02	3.44
	Non-Network	1075	1.24	0.26	1.00	3.81
Before Trimming: Analytic Weight	Network	687	1.83	0.68	1.12	6.68
	Non-Network	1075	1.21	1.30	0.01	14.76
After Trimming: Analytic Weight	Network	687	1.82	0.65	1.12	6.68[a]
	Non-Network	1075	1.16	1.06	0.01	7.13

[a] Trimming was performed within pairs. Respondents with outlying analytic weights were given a weight that was equal to three times the pair-specific mean analytic weight. Though the analytic weight was trimmed for 11 students attending network schools, the maximum weight did not change because there were zero network students with outlying analytic weights within the pair with the largest average analytic weight.

[23] In the absence of consensus about how to best trim outlying survey weights, we followed procedures that were used for the National Assessment of Educational Progress (NAEP) survey weights, as documented by the National Center for Education Statistics (https://nces.ed.gov/nationsreportcard/tdw/weighting/2002_2003/weighting_2003_base_schtrim.aspx).

Exhibit 4.3. Descriptive Statistics for Individual and Combined Weights for Analyses of PBTS Data, for Network and Non-Network Students

		N	Mean	S.D.	Min	Max
Weight 1: Weighting for School Selection	Network	570	1.00	0.00	1.00	1.00
	Non-Network	697	0.37	0.46	0.00	4.05
Weight 2: Weighting for Attrition and Consent	Network	570	1.51	0.51	1.04	5.03
	Non-Network	697	2.25	1.67	1.04	14.91
Weight 3: Weighting for Sampling	Network	570	1.00	0.00	1.00	1.00
	Non-Network	697	2.27	2.05	1	11.44
Weight 4: Weighting for Non-Response	Network	570	1.27	0.29	1.02	3.16
	Non-Network	697	1.56	0.56	1.03	6.29
Before Trimming: Analytic Weight	Network	570	1.90	0.73	1.11	5.83
	Non-Network	697	1.53	1.76	0.01	15.12
After Trimming: Analytic Weight	Network	570	1.90	0.71	1.11	5.83
	Non-Network	697	1.47	1.53	0.01	10.97

[a] Trimming was performed within pairs. Respondents with outlying analytic weights were given a weight that was equal to three times the pair-specific mean analytic weight. Though the analytic weight was trimmed for 11 students attending network schools, the maximum weight did not change because there were zero network students with outlying analytic weights within the pair with the largest average analytic weight.

To assess the quality of the final analytic weight, we examined (a) the degree to which network and non-network students had similar average student background characteristics after applying the final weight (to deal with potential selection bias due to the measured preexisting differences); and (b) the degree to which the intended population of students entering Grade 9 in network schools and the weighted analytic student sample had similar average student background characteristics (allowing us to generalize our results to the Grade 9 population).

A comparison of average student background characteristics before and after applying the final analytic weight for the survey sample is provided in Exhibit 4.4.[24] In this exhibit, the “unweighted” characteristics represent the characteristics of all students who attended network and non-network schools in Cohorts 3 and 4, while the “weighted” characteristics represent the characteristics of survey respondents after applying the final analysis weight. We also provide the average student background characteristics for the sample used for the analysis of high school graduation and postsecondary outcomes (Cohorts 1–3), before and after applying propensity score weights, in Exhibit 4.5. For each characteristic, we report the standardized mean difference (SMD), averaged across pairs. For a given pair and characteristic, the SMD is defined by the following equation:

[24] Analyses of covariate balance using the sample of students who took the PBTS were similar to the results in Exhibit 4.4. Results of these analyses are available upon request.

$$SMD = \frac{\bar{x}_n - \bar{x}_c}{sd_p},$$

where, $\bar{x}_n$ is the network school mean, $\bar{x}_c$ is the non-network school mean, and sd_p is the unweighted, pooled standard deviation for the original Grade 9 population. Across the student background characteristics, the SMD is below 0.25 standard deviations, which is a common threshold for baseline imbalance (What Works Clearinghouse, 2013). To account for imbalance that remains after weighting, we controlled for these covariates in the outcome models (discussed below).

Exhibit 4.4. Network and Non-Network Student Characteristics for Survey Sample Before and After Weighting: Cohorts 3 and 4[1]

	Unweighted Non-Network Mean	Unweighted Network Mean	SMD Before Applying Weights	Weighted Non-Network Mean	Weighted Network Mean	SMD After Applying Weights
Propensity Scores	0.22	0.31	0.84	0.31	0.31	0.06
Grade 8 Math Test Scores	0.05	-0.02	-0.08	0.10	0.03	-0.06
Grade 8 ELA Test Scores	0.00	-0.09	-0.10	-0.01	-0.03	-0.02
ELL Status	27.6%	31.5%	0.10	30.6%	30.7%	0.01
IEP Status	6.2%	6.2%	0.02	3.6%	6.0%	0.11
Gender (Female)	50.9%	52.7%	0.04	53.4%	54.4%	0.02
Race/Ethnicity (Black)	13.8%	15.2%	0.08	10.8%	15.1%	0.14
Race/Ethnicity (Hispanic)	54.6%	58.9%	0.05	64.2%	57.6%	-0.20
Race/Ethnicity (White)	20.7%	20.2%	0.03	17.7%	21.8%	0.17
Cohort (Grade 12)	50.0%	50.3%	0.01	52.6%	43.8%	-0.18

[1] Unweighted means and SMDs were calculated separately for each pair, using the population of incoming Grade 9 students in network and non-network schools, while weighted means and SMDs were calculated separately for each pair, using the weighted sample of survey respondents. The results shown in the exhibit are based on an equally weighted average across the pair-specific means and SMDs. Pair-specific results are available upon request.

Exhibit 4.5. Network and Non-Network Student Characteristics for Graduation and Postsecondary Analysis Sample Before and After Weighting: Cohorts 1–3[1]

	Unweighted Non-Network Mean	Unweighted Network Mean	SMD Before Applying Weights	Weighted Non-Network Mean	Weighted Network Mean	SMD After Applying Weights
Propensity Scores	0.263	0.357	0.85	0.354	0.357	0.02
Grade 8 Math Test Scores	-0.227	-0.336	-0.11	-0.344	-0.336	0.01
Grade 8 ELA Test Scores	-0.191	-0.273	-0.08	-0.235	-0.273	-0.04
ELL Status	21.0%	24.8%	0.10	24.4%	24.8%	0.02
IEP Status	5.9%	5.9%	0.01	5.1%	5.9%	0.04
Gender (Female)	49.3%	54.1%	0.10	53.4%	54.1%	0.01
Race/Ethnicity (Black)	26.7%	28.8%	0.08	27.7%	28.8%	0.07
Race/Ethnicity (Hispanic)	47.5%	51.6%	0.09	53.2%	51.6%	-0.06
Race/Ethnicity (White)	16.4%	14.3%	-0.05	13.7%	14.3%	0.00

[1] Unweighted means and SMDs were calculated separately for each pair, using the population of incoming Grade 9 students in network and non-network schools, while weighted means and SMDs were calculated separately for each pair using the weighted sample. The results shown in the exhibit are based on an equally weighted average across the pair-specific means and SMDs. Pair-specific results are available upon request.

B. Statistical Models

Within-Pair Effect Estimation: Doubly Robust Regression Model

To estimate the effects of enrolling in a deeper learning network school instead of a non-network school, we first conducted pair-by-pair analyses.[25] The analysis method is considered doubly robust (Funk et al., 2011) because it accounts for observed differences in network and non-network students in two ways: (1) through propensity score weighting, and (2) through regression-based covariate adjustment. To apply both the propensity score weight and the regression-based covariate adjustment, we used the following weighted ordinary least squares regression model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{3j}X_{ij} + e_{ij},$$

where Y_{ij} is a given opportunity to learn (OTL) measure for student *i* in school pair *j*; T_{ij} is a dichotomous indicator for whether the student enrolled in the network school (T_{ij}=1) or the non-network school (T_{ij}=0) in the fall of Grade 9; and X_{ij} is a vector of available student background characteristics listed in Exhibit 3.5, as well as a dichotomous indicator for whether the student was in Cohort 3 or Cohort 4. We applied the combined analytic weight, so the estimated effect is representative of students who enrolled in a network school in the fall of Grade 9.

The main parameter of interest is β_{1j}, which is the effect of enrolling in a network school instead of the matched non-network school in a given school pair. Since we standardized measures prior to analysis, estimates of β_{1j} can be interpreted as the estimated effect size for network school enrollment in pair j.

Analyses of PBTS data used a variation on this model that allowed us to measure the error associated with PBTS score. Accounting for measurement error was particularly important for analyses of PBTS data because different students received different forms of the PBTS and each student responded to only a subset of items within each subject area. For analyses of PBTS scores, we used a two-level, variance-known, hierarchical linear model (Raudenbush and Bryk, 2002). The first level of analysis accounted for the error associated with PBTS scores, while the second level mirrored the equation above for estimating within-pair effects.

Averaging Pair-Specific Effect Estimates: Meta-Analysis

The main results presented in the report are estimates of the effect of attending a network school, averaged across the pairs for which we have data. We view the results as pertaining only to the particular schools included in our sample and not to a wider population. Thus, we used a fixed-

[25] We conducted separate pair-specific analyses (instead of combining data into one analysis) for two main reasons. First, data access limitations precluded combining student data from California and New York City. Second, because the available student background characteristics differed across districts, pooling the data would have required restricting the data to a subset of the characteristics. By conducting separate analyses for each pair, we were able to maximize the number of student background characteristics we could include in the analyses.

effects meta-analysis approach (Hedges and Vevea, 1998) to calculate the average effect across the school pairs:

$$\overline{ES} = \frac{\sum_{j=1}^{11} w_j \hat{\beta}_j}{\sum_{j=1}^{11} w_j},$$

where $\widehat{\beta_j}$ is the estimated network effect for pair j, and w_j is the inverse of the variance of pair j's estimate (i.e., one divided by the standard error squared).[26] This calculation is the precision-weighted mean effect size of the pair-specific effect estimates, where estimates with more precision (less error variance) receive more weight in the average.

A power analysis conducted prior to data collection indicated that with this design we should be able to detect effect sizes as small as 0.07 to 0.09 standard deviations for survey and PBTS outcomes, depending on the specific assumptions. (See Exhibit 4.6.) After completing the analyses, the realized minimum detectable effect size (MDES) for the interpersonal and intrapersonal competency outcome measures was between 0.13 and 0.16 depending on the measure. The realized MDES for PBTS scores was between 0.09 and 0.11 (depending on the subject area) while the realized MDES for state achievement test scores was 0.04.

Exhibit 4.6. Minimum Detectable Effect Size (MDES) for Survey and PBTS Outcomes Using a Fixed Effects Model, Based on Different Assumptions About the Percentage of Variance Explained by Blocking (School Pairs and Propensity Strata) and By Covariates

	% Variance Explained by Blocking		
% Variance Explained by Covariates (R^2)	10%	15%	20%
50%	0.092	0.089	0.087
60%	0.082	0.080	0.078
70%	0.071	0.069	0.067

Note: The MDES is based on a two-tailed significance test, alpha=.05, with power=80 percent. The sample includes eight pairs of schools, two cohorts per school, five strata per cohort, and 8.4 treatment students and 13.8 non-network students per stratum per cohort on average.

Subgroup Analysis

For each effect estimate, we also examined whether the effect of network school enrollment differed across student subgroups. We examined the following subgroups:

- Gender: male versus female

[26] Meta-analyses may be conducted using either a fixed or random effects approach (Hedges and Vevea, 1998). Random-effects meta-analysis would assume that the schools in the study were drawn from a larger population, and the goal of these models would be to estimate the effect of attending a network school for the population.

- Cohort: Grade 11 (Cohort 4) versus Grade 12 (Cohort 3)[27]
- Free/Reduced-Price Lunch Status: eligible versus not eligible[28]
- Prior English Language Arts Achievement: high achieving versus low achieving

We used two different approaches to create subgroups based on prior English Language Arts achievement. First, we compared students' Grade 8 ELA test scores to the state average[29] test score for the year in which the test was taken. Students were classified as low achieving if their test score fell below the state mean and high achieving if their score fell above the state mean. Second, we calculated the average Grade 8 ELA test score within each school pair and classified students as low achieving if their test score fell below the pair-specific mean and high achieving if their score fell above the mean. The first definition compares students' test scores to a statewide benchmark, while the second measure directly compares students' test scores to the test scores of their peers within the same school pair.

To test whether effects differed significantly across subgroups, we estimated a model similar to the model described above, adding the interaction of network enrollment and the dichotomous subgroup indicator:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}X_{ij} + \beta_{3j}S_{ij} + \beta_{4j}(S_{ij} \times T_{ij}) + e_{ij},$$

where *S* is the dichotomous subgroup indicator. In this model, the primary parameter of interest is β_{4j}, which captures the network effect difference for the subgroup. Each subgroup analysis was performed independently, and so only one interaction term was added to the model at a time. We used the same meta-analytic approach described above to calculate average subgroup effects across the pairs.

Sensitivity Analyses

As a sensitivity analysis for analyses of interpersonal and intrapersonal competency outcomes, we removed pairs in which (1) the response rate for the survey fell below 70 percent; and/or (2) the difference in response rates between the matched network and non-network school in the pair was larger than 10 percentage points. This resulted in the removal of five matched pairs of schools from survey analyses. Analyses using the remaining six pairs showed that the differences between network and non-network students for four outcomes—collaboration skills, academic engagement, motivation to learn, and self-efficacy—were still significant and were larger in magnitude. In addition, we ran sensitivity analyses using a fixed effects model with equally weighted (rather than inverse-variance weighted) pairs of schools and a random effects model.

---

[27] We did not perform subgroup analyses by cohort for high school achievement test scores, high school graduation, and postsecondary enrollment outcomes.

[28] Information regarding students' eligibility to receive free or reduced-price lunch (FRPL) was not available in all districts. Due to the small number of schools that participated in the PBTS for which we had this information, we did not perform subgroup analyses for PBTS scores by FRPL eligibility status.

[29] In New York City, test scores were compared to the New York City average ELA test score for the appropriate year.

The findings in these sensitivity analyses closely resembled the main findings, though the effects on collaboration skills and motivation to learn that were significant in the fixed effects model were not significant when a random effects meta-analysis was performed. Detailed results are available upon request.

We performed several sensitivity tests for the analyses of effects on the PBTS. First, we performed the meta-analysis excluding one pair of schools in which school staff requested that several of the non-network students leave the testing session early.[30] In another sensitivity analysis, we removed four pairs of schools that had an overall response rate below 60 percent. The results of these sensitivity analyses were consistent with the main effects presented in the report. An additional supplemental analysis removed six pairs of schools in which the overall response rate was below 65 percent. After removing these schools, the effect of attending a network school on PBTS reading scores was not significant, but the effects for math and science remained significant.

For high school graduation, high school achievement test scores, and postsecondary enrollment outcomes, we conducted a set of sensitivity analyses to examine the effects of using different comparison schools. In five cases, the best match for a network school was not willing to participate in the primary survey and PBTS data collection at the time of recruitment.[31] As a result, we chose an alternative matched non-network school. To examine the sensitivity of the results to using these alternative matches, we compared the results from (1) the main analysis (in which the network school was matched with the non-network school that participated in the primary data collection but was not the best-matched non-network school); and (2) an alternative analysis (in which the network school was matched with the best-matched non-network school). Overall, results using the best-matched non-network school were consistent with results that used the non-network school that participated in primary data collection. Results of these alternative analyses are available upon request.

For on-time high school graduation, three additional sensitivity analyses were performed. For the report, we defined graduates as students who graduated within four years from a school within the same district that students entered in Grade 9. A non-negligible percentage of students may have transferred out of the district prior to graduation, and so we conducted two additional sensitivity analyses, which excluded:

- Students who were identified as transferring out of the district in the district’s administrative records[32]

[30] In the main analysis, we handled this issue using attrition weights.

[31] While a total of five pairs of schools included a network school and a non-network school that was not the best-matched comparison school, only four of these pairs provided graduation and achievement score data, and four of these pairs provided postsecondary data.

[32] The quality of information surrounding the reason students withdrew from the district varied across school districts. In some districts, transfer students were easily identifiable; in other districts, students disappeared from enrollment data without a clear reason for withdrawal. For this sensitivity analysis, we removed two pairs of schools for which the district did not provide information about why students withdrew from the district.

- All students who were not enrolled in the district in the fall of the fourth year of high school

For these sensitivity analyses, attrition weights were calculated to account for the inverse probability that students transferred out of the district (and were enrolled in the fall of the fourth year of high school). The weights were calculated using the procedures described in Section IV.A for the other attrition/consent weights. Finally, in a third sensitivity analysis, we looked at the effects of attending a network school on graduation within five years of high school entry.[33] All three of these sensitivity analyses showed that attending a deeper learning network school had a positive impact on the odds of graduating from high school, and that the effects were larger than the effects in the main analyses.

Finally, we performed sensitivity analyses for students' postsecondary enrollment outcomes including different numbers of covariates in logistic regression models. In logistic regression models, coefficients may be unstable when quasi-complete separation occurs (Allison, 2008). Quasi-complete separation may occur with binary outcomes when one or more binary covariates in the statistical model predict the outcome perfectly (or nearly perfectly). While the effects of attending a deeper learning network school on enrollment in a four-year institution and enrollment in a selective institution did not lose significance across different model specifications, the magnitude of these effects did vary across models, particularly for the outcome of "enrolled in a selective institution" (which ranged from 1.448 to 1.649 across model specifications). In the report, we present findings from a conservative model that included sufficient background characteristics to ensure an appropriate comparison between network and non-network students, while not including a large number of covariates that may lead to problems of separation in logistic regression models.[34]

[33] This sensitivity analysis only included students in Cohorts 1 and 2 because students in Cohort 3 had not been enrolled in high school for five years at the time of data analysis.

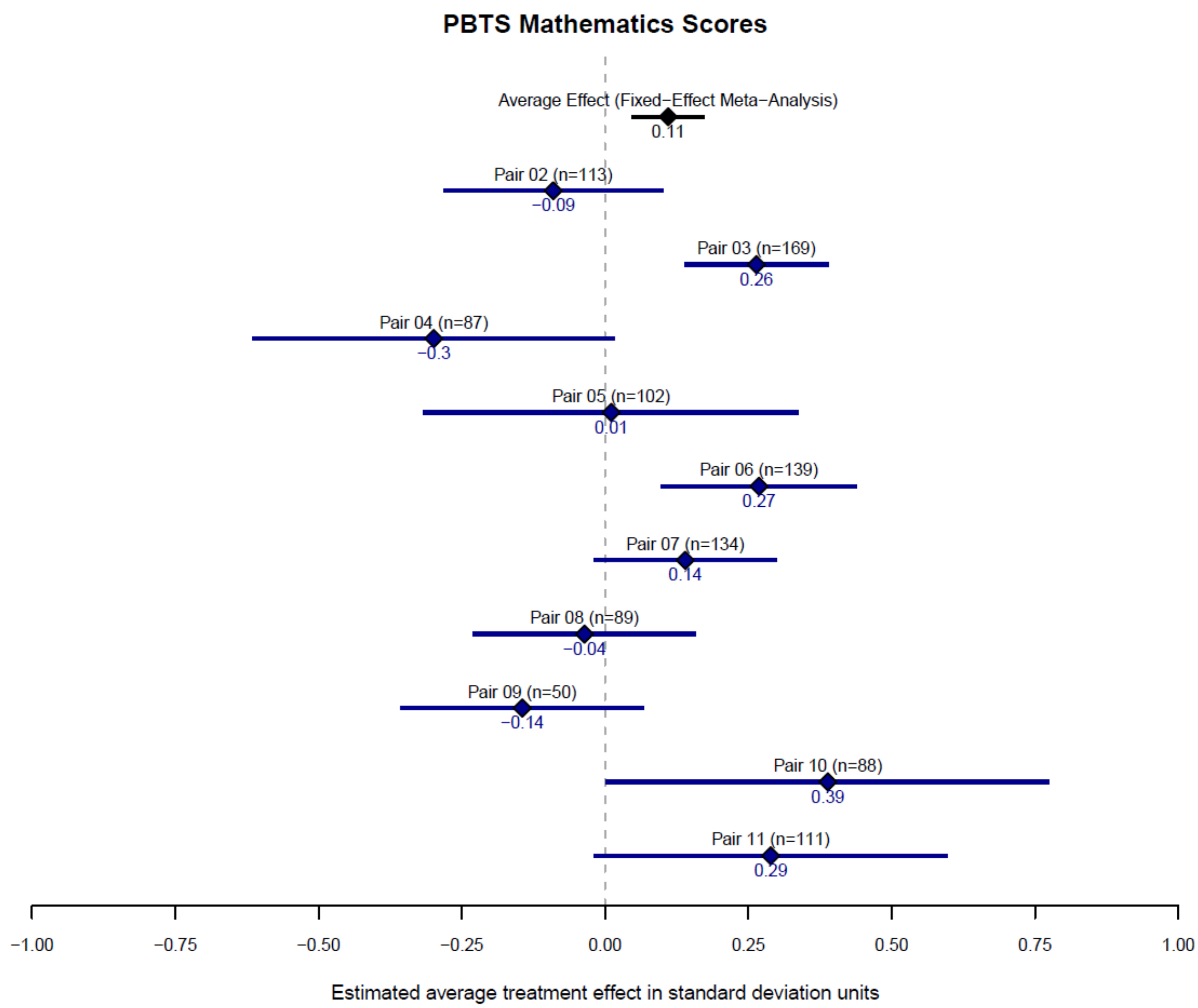
[34] The covariates included in the main statistical models include two indicators for cohort (with the oldest cohort as the reference category); two indicators for race/ethnicity (white and Asian, with "nonwhite" as the reference category); English language learner status; eligibility to receive free or reduced-price lunch (where available); an indicator for students with an individualized education plan (IEP); an indicator for students whose parents had less than a high school education (where available); Grade 8 mathematics test scores; and Grade 8 ELA test scores.

V. Detailed Results

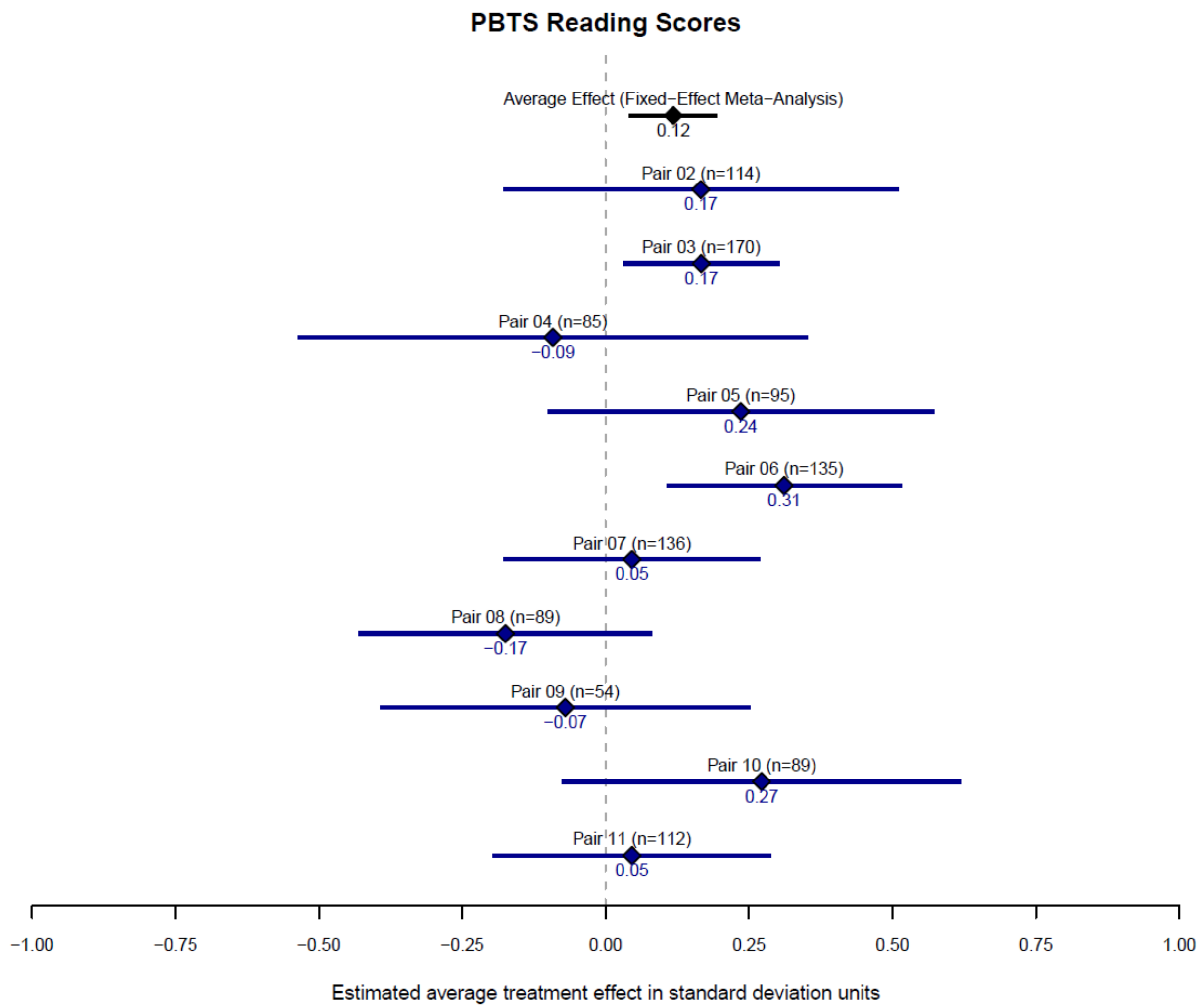
In this section, we provide supplemental figures and tables presenting more detailed information for the results described in the report.

A. PBTS Achievement Effect Estimates: Pair-Specific Results

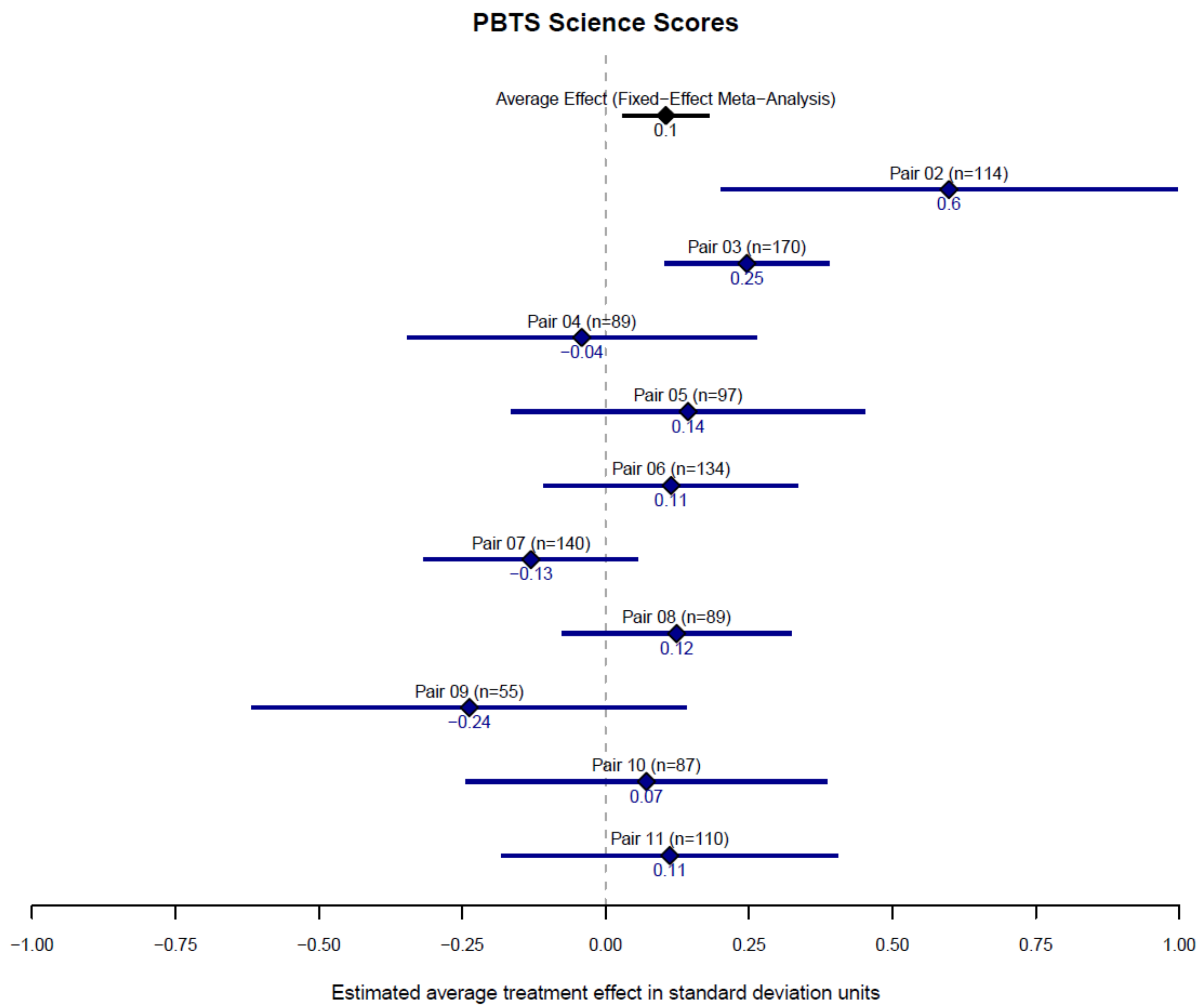
This section provides forest plots that display the meta-analytic average estimate and the pair-specific estimates of the differences in PBTS achievement scores for students attending network and non-network schools. Estimates for PBTS scores in mathematics, reading, and science are presented in separate forest plots.



Note: A significant value of i-squared for this outcome ($p < 0.001$) indicates that the effect of attending a network school significantly differs across school pairs.



Note: A non-significant value of i-squared for this outcome (p = 0.154) indicates that the effect of attending a network school does not significantly differ across school pairs.



Note: A significant value of i-squared for this outcome (p = 0.018) indicates that the effect of attending a network school significantly differs across school pairs.

B. High School Achievement Test Score Effect Estimates: Pair-Specific Results

Exhibit 5.1 presents the pair-level findings for high school achievement test scores in mathematics and ELA.

Exhibit 5.1: Results for High School Mathematics and ELA Achievement Test Scores in Cohorts 1–4, Overall Average and by Pair (Effect Sizes)

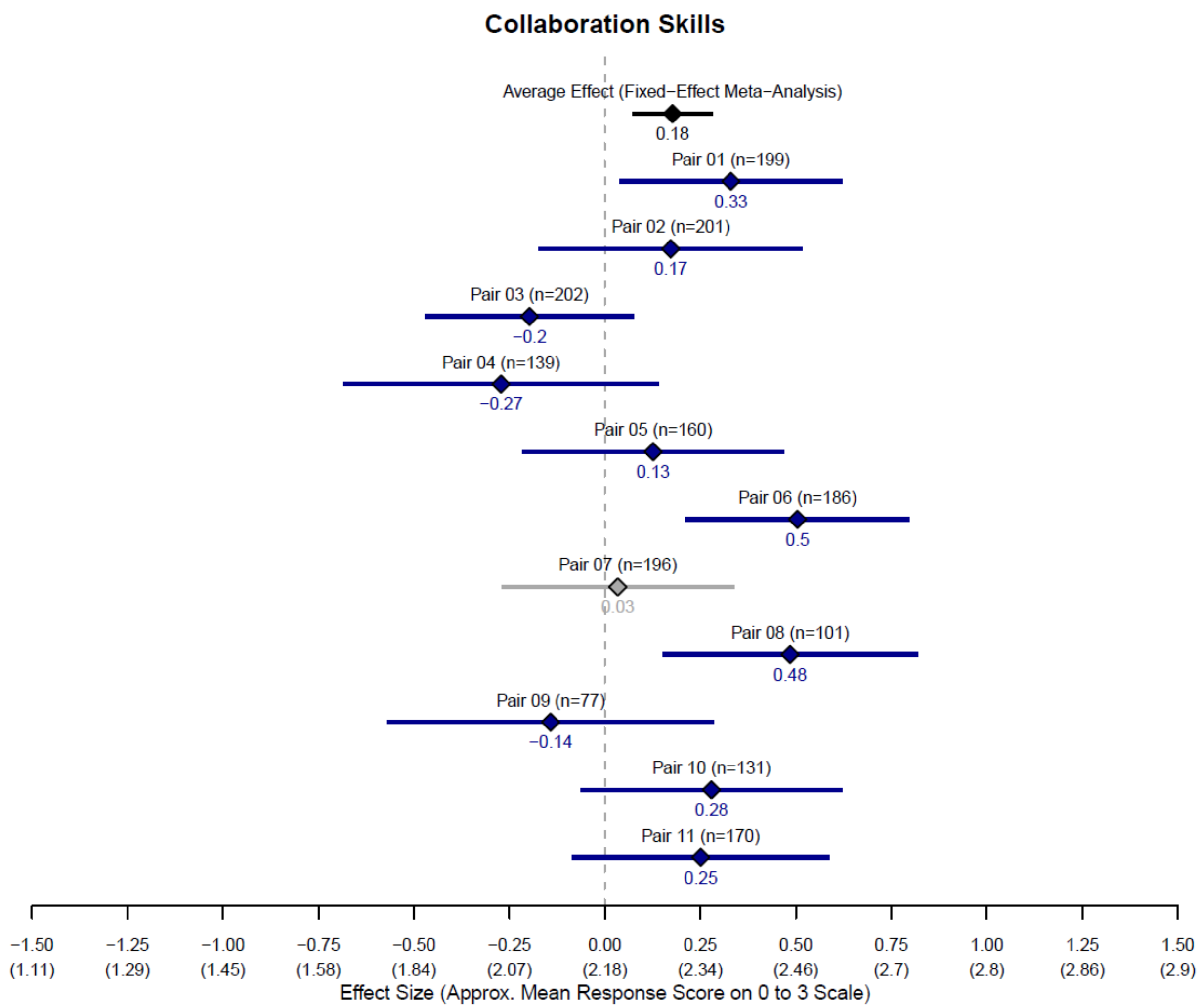
Pair	Mathematics	ELA
Meta-Analytic Results (Average)	**0.097**	**0.052**
Pair 3*	0.270 (<0.001)	0.293 (<0.001)
Pair 4	-0.117 (0.001)	ns - (0.959)
Pair 5	0.174 (<0.001)	ns + (0.906)
Pair 6	ns - (0.159)	ns + (0.173)
Pair 7	0.069 (0.038)	ns + (0.701)
Pair 8	0.163 (0.002)	0.136 (0.013)
Pair 9	ns + (0.816)	-0.248 (0.004)
Pair 10	ns - (0.384)	ns - (0.376)
Pair 11	N/A	0.481 (<0.001)
Pair 12*	ns + (0.067)	ns + (0.217)
Pair 13	0.497 (<0.001)	0.320 (<0.001)
Pair 14	ns + (0.482)	ns + (0.113)
Pair 15	ns - (0.750)	-0.302 (<0.001)

Note: Numbers in parentheses are p-values; ns+ indicates that effects are positive and non-significant; ns- indicates that effects are negative and non-significant. A significant value of i-squared for these outcomes ($p < 0.001$ for both mathematics and ELA) indicates that the effect of attending a network school significantly differs across school pairs.

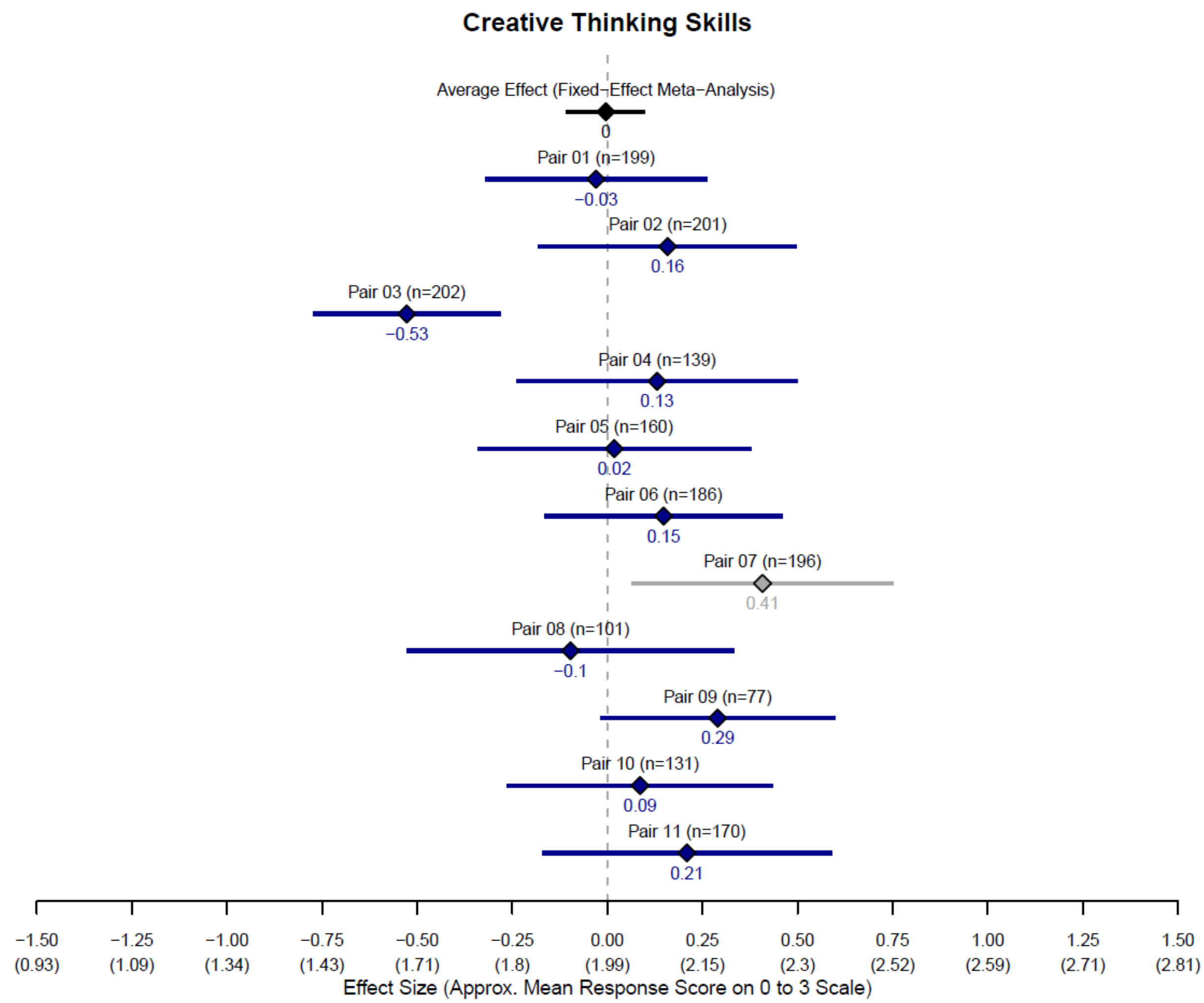
* The data supplied for these network schools contained poor-quality high school achievement test score data for the 2008–09 academic year, leading to a large amount of missing data. As a result, analyses for these pairs exclude Cohort 1. Other cohorts were not affected by this data quality issue because they were not eligible to take the CAHSEE exam (i.e., they had not yet reached Grade 10) by the 2008–09 academic year.

C. Interpersonal and Intrapersonal Competency Outcome Effect Estimates: Pair-Specific Results

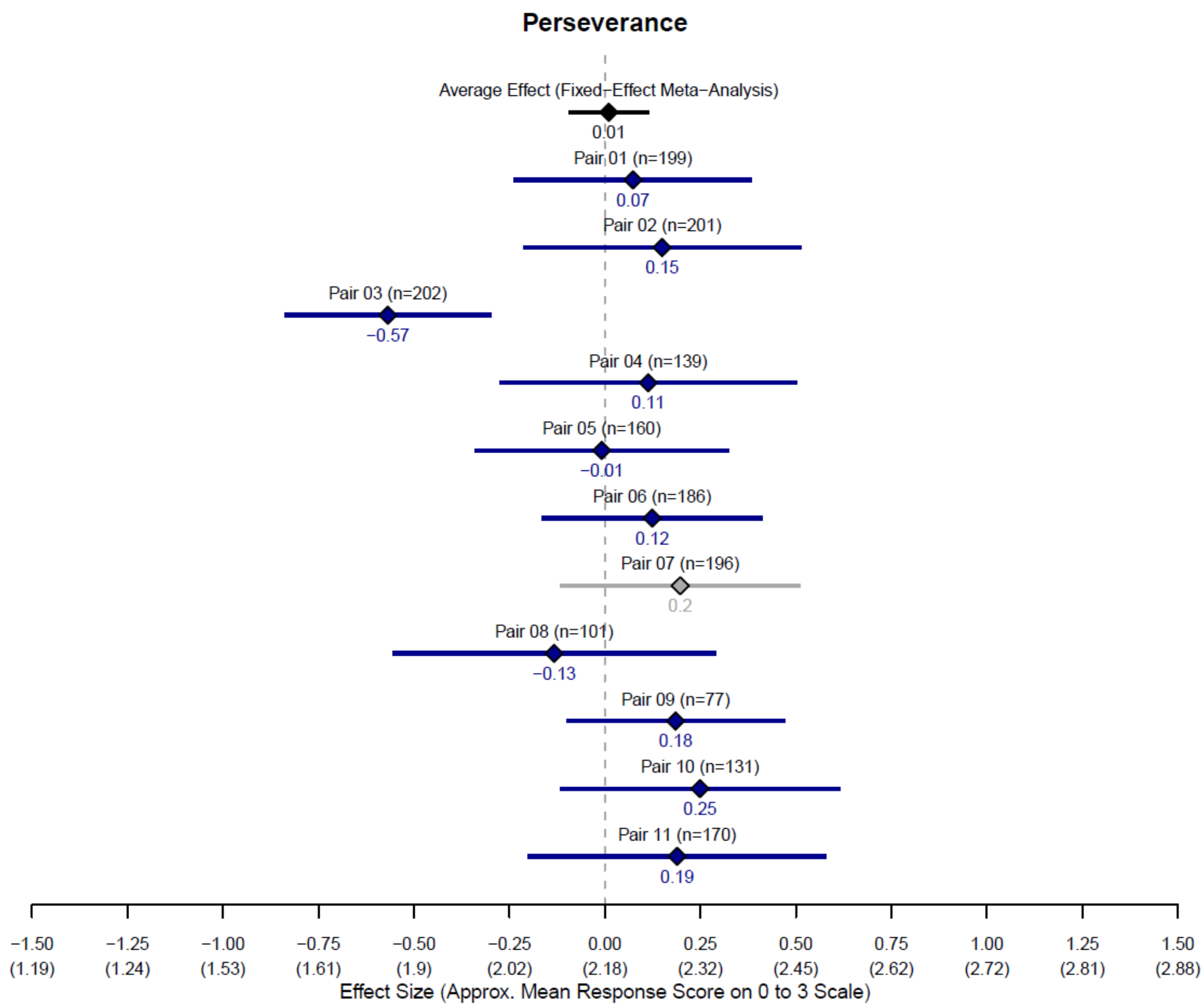
This section provides forest plots that display the meta-analytic average estimate and the pair-specific estimates of the differences in interpersonal and intrapersonal competency measures for students attending network and non-network schools. Estimates for each interpersonal and intrapersonal competency are presented in separate forest plots.



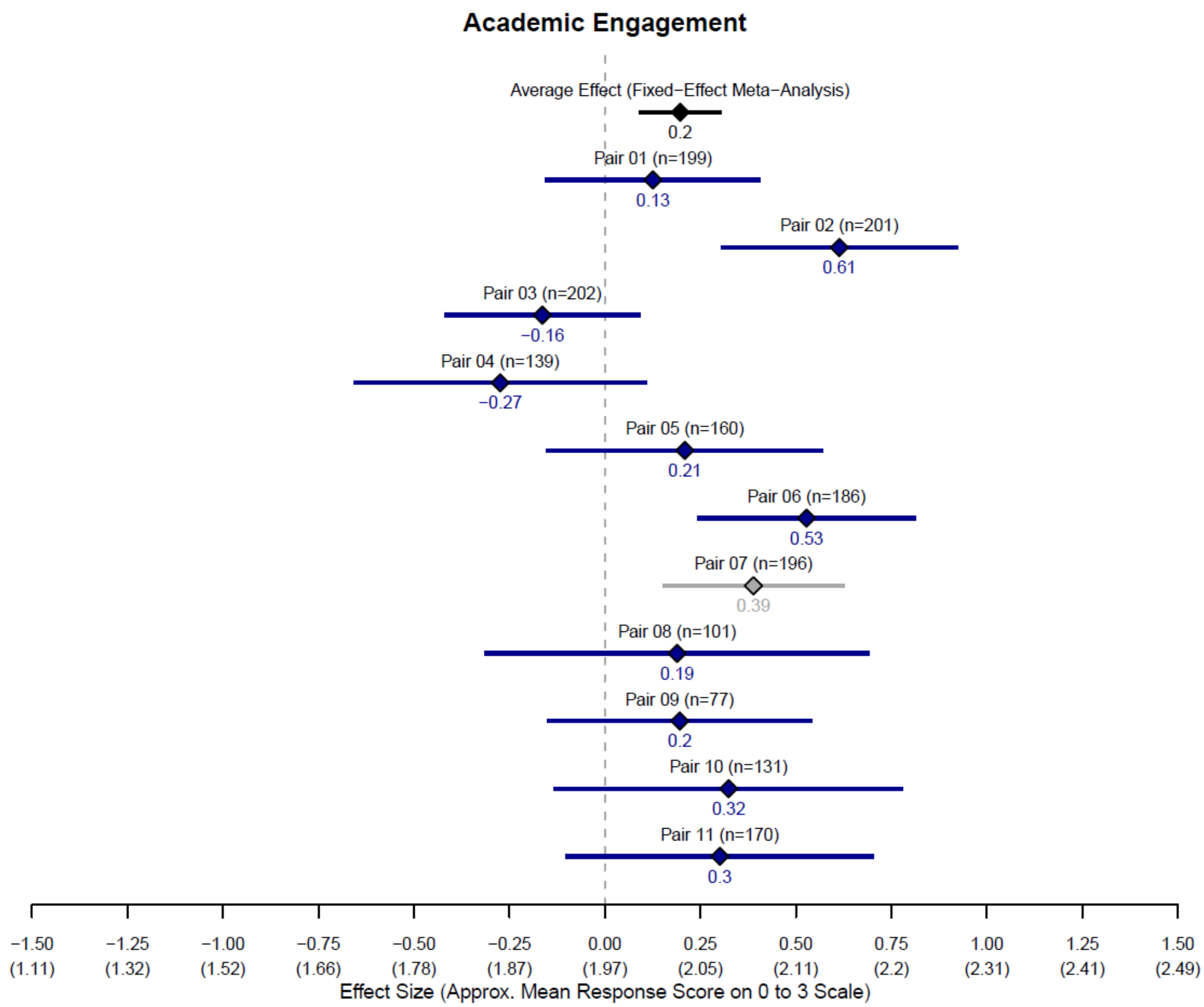
Note: A significant value of i-squared for this outcome (p = 0.005) indicates that the effect of attending a network school significantly differs across school pairs.



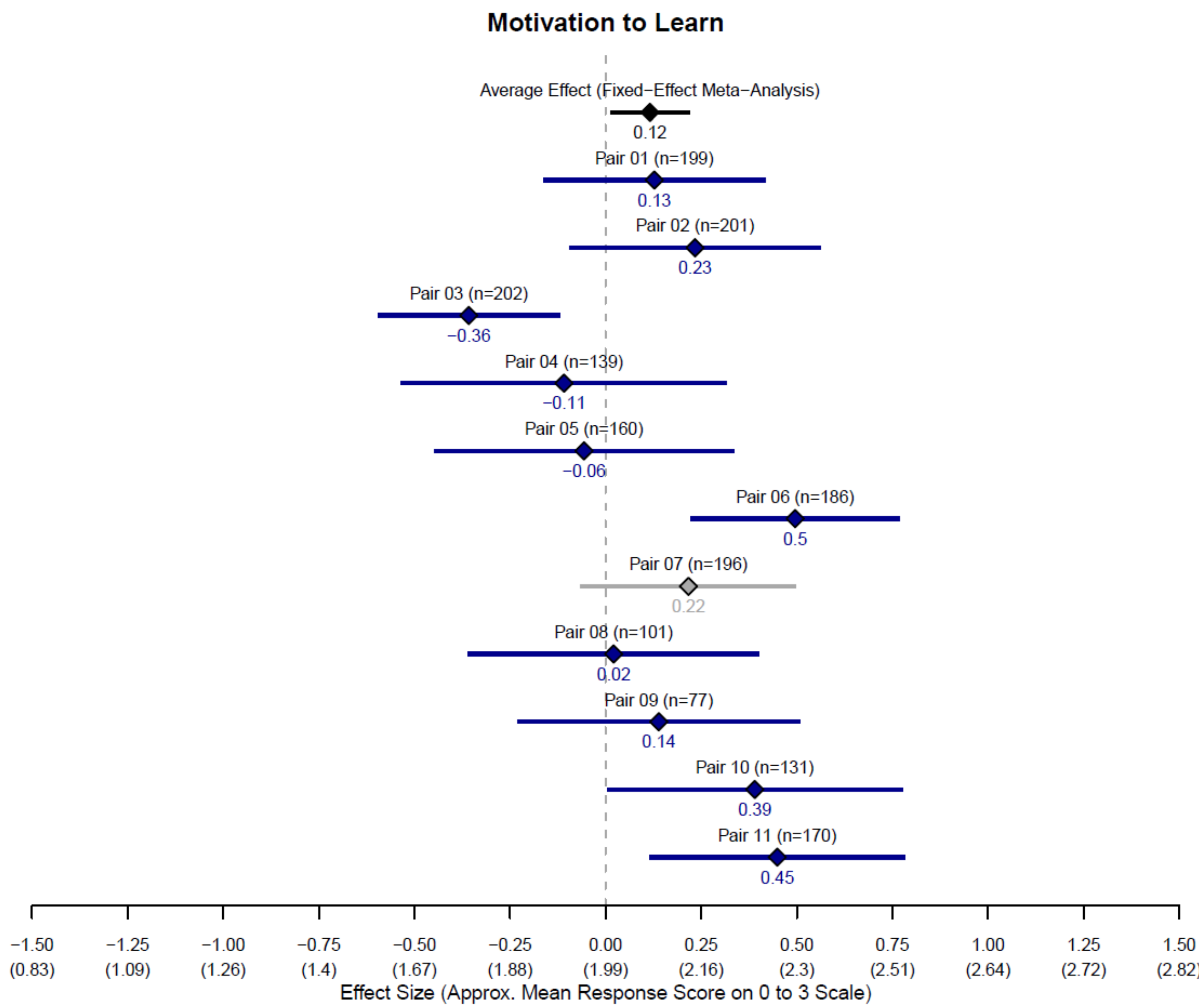
Note: A significant value of i-squared for this outcome (p = 0.003) indicates that the effect of attending a network school significantly differs across school pairs.



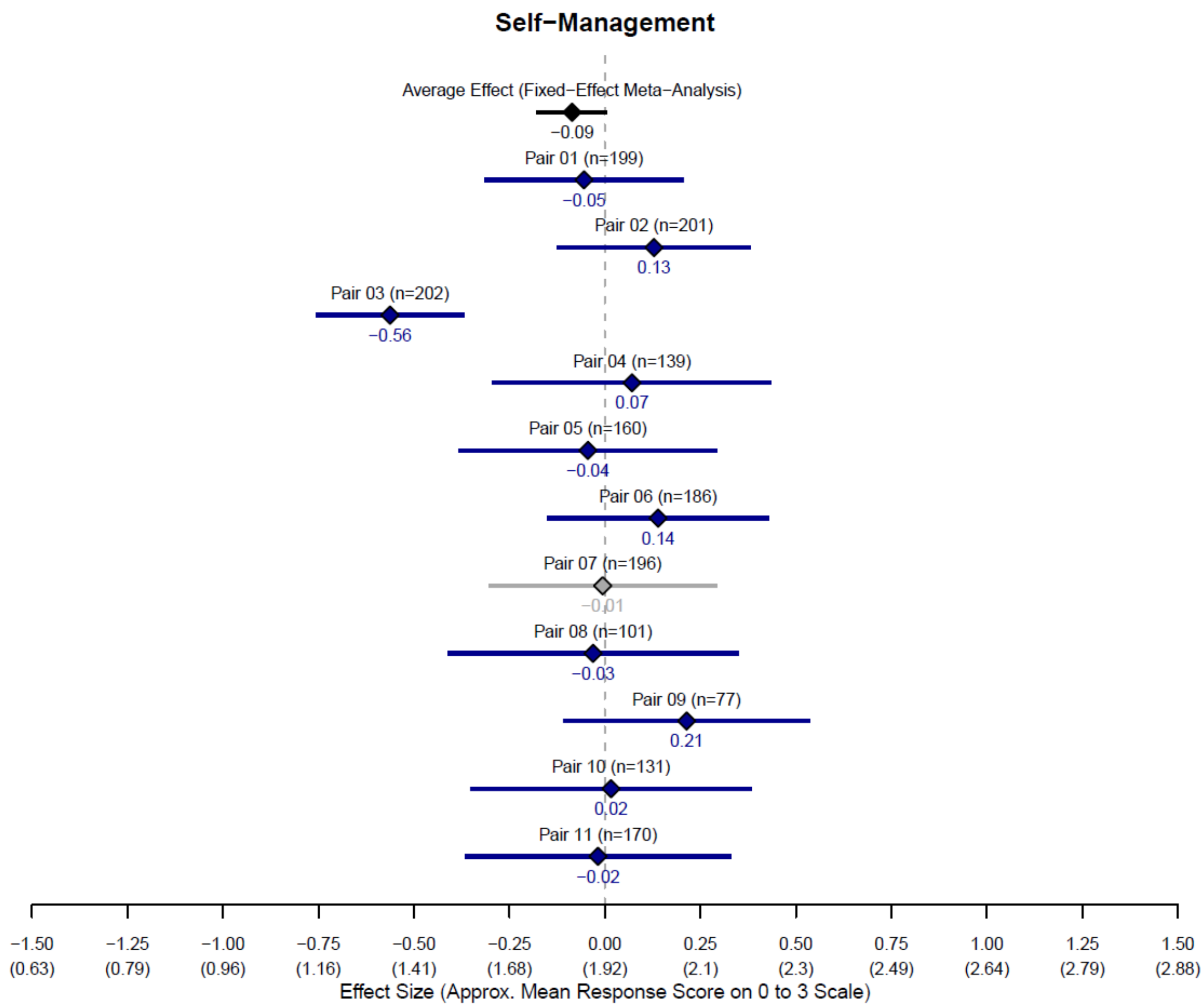
Note: A significant value of i-squared for this outcome (p = 0.005) indicates that the effect of attending a network school significantly differs across school pairs.



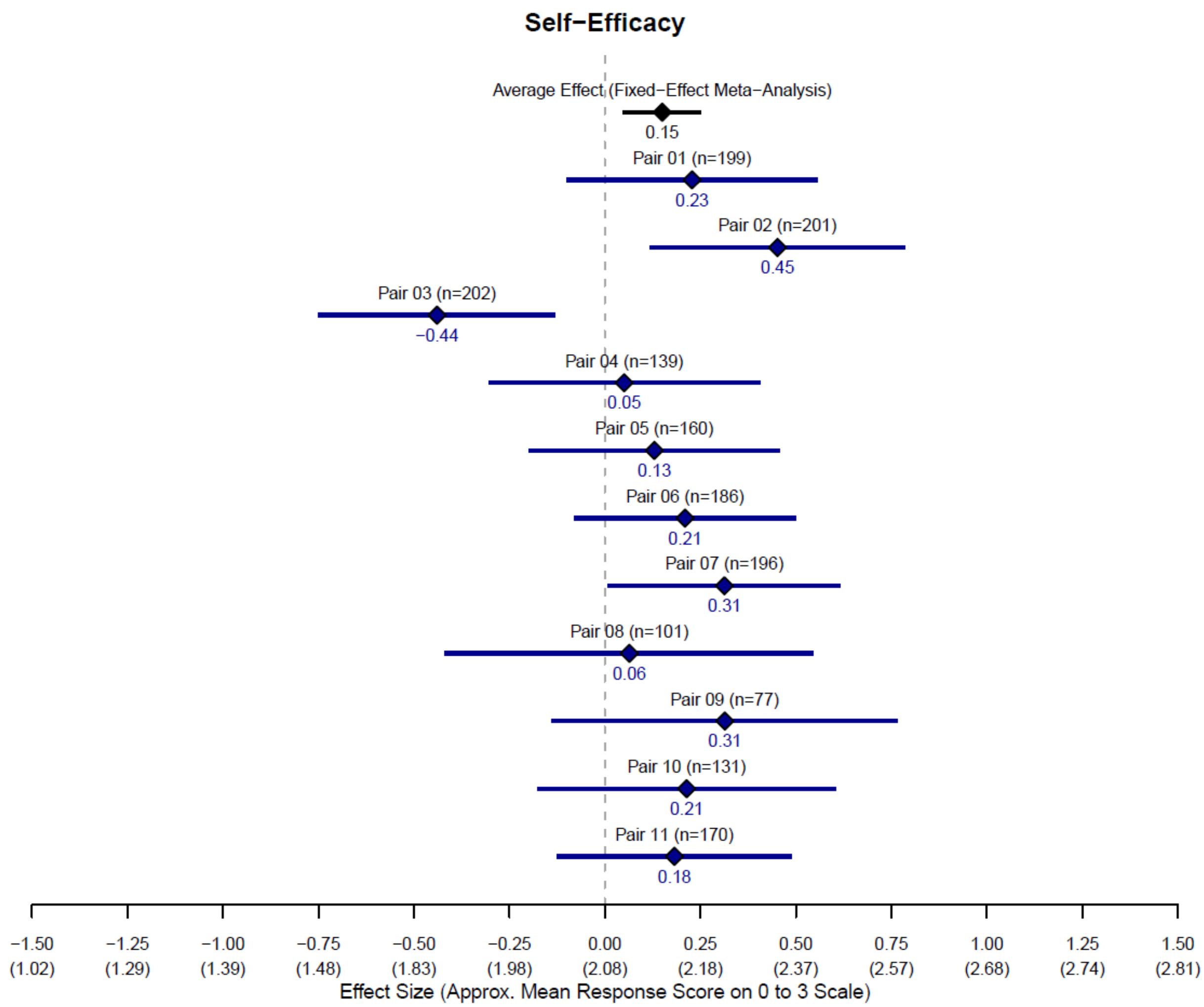
Note: Because we examined the effect of attending a network school on multiple interpersonal and intrapersonal competencies, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on two intrapersonal competencies that are likely to be highly correlated: academic engagement and motivation to learn. The qualifying test involved examining the impact of network school attendance on a composite of these two intrapersonal competency measures. The impact on the composite measure was significant for Pairs 2, 3, 6, and 10. A significant value of i-squared for this outcome ($p = 0.002$) indicates that the effect of attending a network school significantly differs across school pairs.



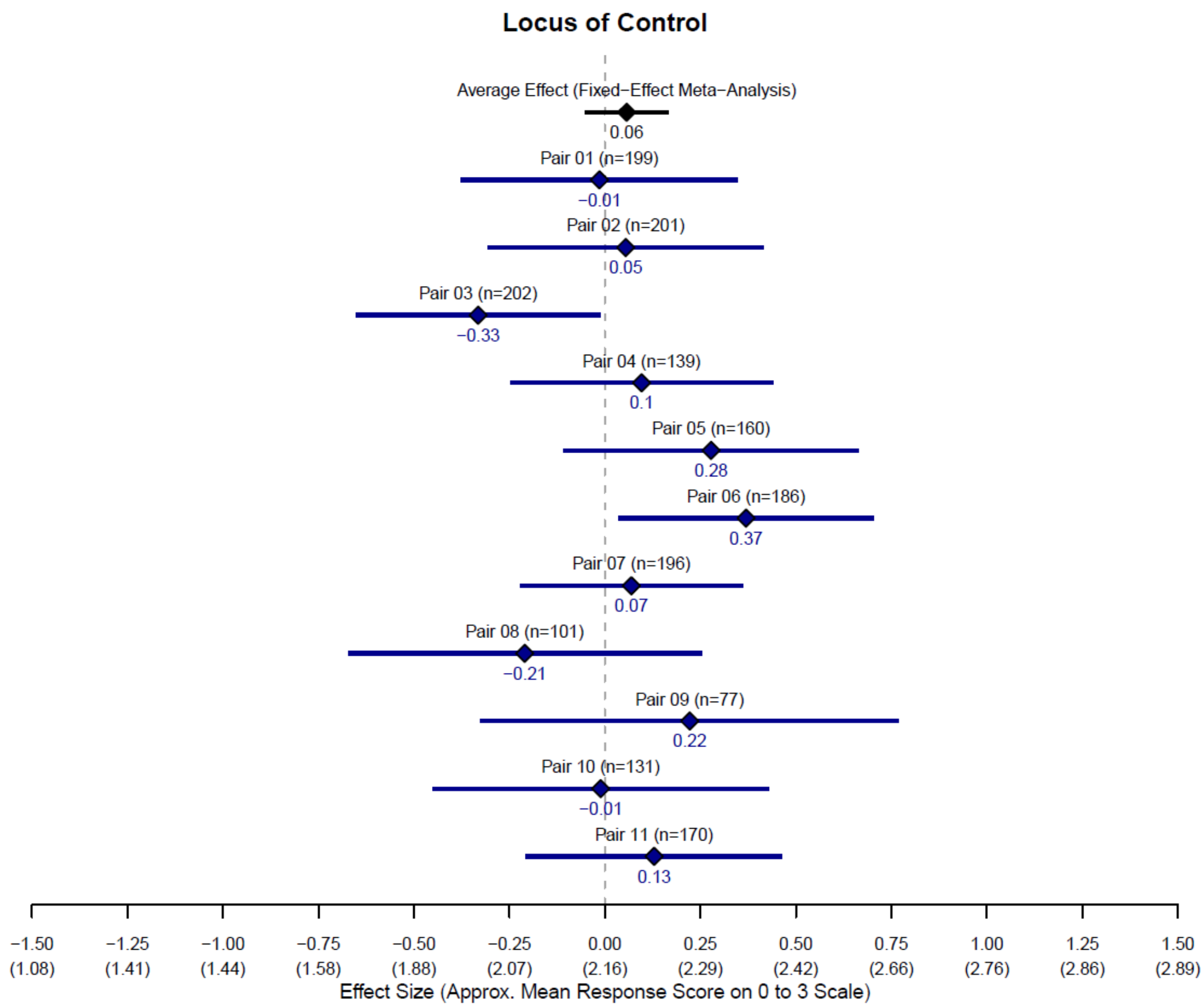
Note: Because we examined the effect of attending a network school on multiple interpersonal and intrapersonal competencies, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on two intrapersonal competencies that are likely to be highly correlated: academic engagement and motivation to learn. The qualifying test involved examining the impact of network school attendance on a composite of these two intrapersonal competency measures. The impact on the composite measure was significant for Pairs 2, 3, 6, and 10. Thus, the result for motivation to learn for Pair 11 (shown above) could be due to chance. A significant value of i-squared for this outcome ($p < 0.001$) indicates that the effect of attending a network school significantly differs across school pairs.



Note: A significant value of i-squared for this outcome ($p < 0.001$) indicates that the effect of attending a network school significantly differs across school pairs.



Note: Because we examined the effect of attending a network school on multiple interpersonal and intrapersonal competencies, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on two intrapersonal competencies that are likely to be highly correlated: self-efficacy and locus of control. The qualifying test involved examining the impact of network school attendance on a composite of these two intrapersonal competency measures. The impact on the composite measure was significant for Pair 3. Thus, the results for self-efficacy for Pair 2 and Pair 7 (shown above) could be due to chance. A significant value of i-squared for this outcome (p = 0.033) indicates that the effect of attending a network school significantly differs across school pairs.



Note: Because we examined the effect of attending a network school on multiple interpersonal and intrapersonal competencies, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on two intrapersonal competencies that are likely to be highly correlated: self-efficacy and locus of control. The qualifying test involved examining the impact of network school attendance on a composite of these two intrapersonal competency measures. The impact on the composite measure was significant for Pair 3. Thus, the result for locus of control for Pair 6 (shown above) could be due to chance. A non-significant value of i-squared for this outcome (p = 0.257) indicates that the effect of attending a network school does not significantly differ across school pairs.

D. Effect Estimates for Graduation in the District Within Four Years: Pair-Specific Results

Exhibit 5.2 presents the pair-level findings for graduation in the same school district within four years of entering high school.

Exhibit 5.2. Results for High School Graduation Among All Students in Cohorts 1–3, Overall Average and By Pair: Odds Ratios

Pair	Graduation	
Meta-Analytic Results (Average)	**1.467**	
Pair 3	ns +	(0.917)
Pair 4	3.505	(<0.001)
Pair 5	1.978	(<0.001)
Pair 6	ns -	(0.830)
Pair 7	ns -	(0.452)
Pair 8	1.894	(0.037)
Pair 9	ns +	(0.059)
Pair 10	ns -	(0.193)
Pair 11	3.214	(<0.001)
Pair 12	ns -	(0.673)
Pair 13	2.160	(<0.001)
Pair 14	ns +	(0.789)
Pair 15	ns +	(0.088)

Note: Numbers in parentheses are p-values; ns+ indicates that effects are positive and non-significant; ns- indicates that effects are negative and non-significant. A significant value of i-squared for this outcome ($p < 0.001$) indicates that the effect of attending a network school significantly differs across school pairs.

E. Postsecondary Outcomes Effect Estimates: Pair-Specific Results

Exhibit 5.3 presents the pair-level findings for postsecondary enrollment outcomes.

Exhibit 5.3. Results for Postsecondary Outcomes Among All Students in Cohorts 1–3, Overall Average and by Pair: Odds Ratios

Network School	Enrolled in Postsecondary Education by Fall 2013	Enrolled in a Two-Year Institution by Fall 2013	Enrolled in a Four-Year Institution by Fall 2013	Enrolled in a Selective Institution by Fall 2013	Postsecondary Persistence
Meta-Analytic Results (Average)	**ns +**	**ns -**	**1.265**	**1.632**	**ns -**
Pair 1	ns + (0.392)	1.503 (0.006)	0.629 (0.031)	ns + (0.231)	ns - (0.654)
Pair 2	ns - (0.798)	ns - (0.391)	ns + (0.136)	4.764 (0.012)	ns - (0.657)
Pair 6	ns + (0.229)	ns + (0.689)	1.723 (0.037)	3.258 (0.007)	ns - (0.559)
Pair 7	ns - (0.058)	0.649 (0.010)	ns + (0.159)	ns + (0.310)	ns - (0.319)
Pair 8	ns - (0.829)	0.082 (<0.001)	4.306 (<0.001)	ns - (0.226)	ns + (0.337)
Pair 9	ns + (0.339)	ns + (0.311)	ns - (0.977)	ns + (0.560)	ns + (0.834)
Pair 10	ns - (0.090)	ns - (0.074)	ns - (0.464)	N/A	0.440 (0.011)
Pair 11	ns + (0.087)	ns - (0.703)	1.686 (0.025)	12.948 (0.002)	ns + (0.289)
Pair 13	1.603 (0.016)	1.825 (0.009)	ns + (0.267)	ns - (0.870)	ns + (0.115)
Pair 14	ns - (0.587)	ns - (0.442)	ns + (0.668)	ns - (0.668)	ns + (0.851)
Pair 15	ns - (0.709)	ns + (0.642)	ns - (0.286)	ns + (0.408)	ns - (0.452)

Note: Numbers in parentheses are p-values; ns+ indicates that effects are positive and non-significant; ns- indicates that effects are negative and non-significant. A significant value of i-squared for these outcomes (Enrolled in Postsecondary Education: $p = 0.050$; Enrolled in a Two-Year Institution: $p < 0.001$; Enrolled in a Four-Year Institution: $p < 0.001$; Enrolled in a Selective Institution: $p = 0.017$) indicates that the effect of attending a network school significantly differs across school pairs. A non-significant value of i-squared for these outcomes (Postsecondary Persistence: $p = 0.210$) indicates that the effect of attending a network school does not significantly differ across school pairs.

F. Effect Estimates: Subgroup Results

As discussed in Section IV.B, we conducted analyses of the different effects of attending network schools for the following subgroups: male versus female (gender); Grade 11 students versus Grade 12 students at the time of survey and PBTS data collection (cohort); students eligible for free or reduced-price lunch versus students ineligible for free or reduced-price lunch (FRPL status); and students with low Grade 8 achievement versus students with high Grade 8 achievement (prior achievement). The results are shown in Exhibit 5.4. (Exhibit 5.4 presents the significant coefficients associated with the interaction term between subgroup membership and the indicator for attending a deeper learning network school.) For those coefficients that achieved statistical significance, Exhibit 5.5, Exhibit 5.6, and Exhibit 5.7 provide more detailed information regarding the differential impacts of attending a deeper learning network school by gender and prior achievement.

As demonstrated in Exhibit 5.4, the impact of attending a deeper learning network school on various student outcomes differed by gender and prior achievement. There were few differences when outcomes were analyzed by cohort or free or reduced-price lunch status.

Exhibit 5.4. Meta-Analytic Results for Subgroup Analyses for Student Outcomes: Interaction Terms Reported as Effect Sizes and Odds Ratios

Network School	Gender: Female Versus Male	Cohort: Grade 12 Versus Grade 11	FRPL: Yes Versus No	Achievement: Low Versus High (Absolute Reference)	Achievement: Low Versus High (Pair Reference)
Subgroup results for interpersonal and intrapersonal competency outcomes, PBTS scores, and high school achievement test scores (in effect sizes)					
Creative Thinking Skills	ns+	ns-	ns-	ns-	ns-
Collaboration Skills	0.307	ns+	ns-	ns-	ns+
Perseverance	0.333	ns+	ns-	-0.321	-0.227
Self-Management	0.244	ns+	ns-	-0.212	-0.219
QT: Engagement and Motivation	0.278	0.215	ns-	-0.370	-0.251
Academic Engagement	0.331	0.281	ns+	-0.395	-0.249
Motivation to Learn	0.287	ns-	ns-	-0.374	-0.279
QT: Self-Efficacy and Locus of Control	0.286	ns+	-0.283	ns-	ns-
Self-Efficacy	0.251	ns+	-0.300	ns-	ns-
Locus of Control	0.251	ns+	ns-	ns+	ns+
PBTS: Math	ns-	ns+	N/A	ns-	ns-

Network School	**Gender: Female Versus Male**	**Cohort: Grade 12 Versus Grade 11**	**FRPL: Yes Versus No**	**Achievement: Low Versus High (Absolute Reference)**	**Achievement: Low Versus High (Pair Reference)**
PBTS: ELA	ns-	ns+	N/A	ns+	ns+
PBTS: Science	ns-	ns-	N/A-	ns-	ns-
High School Achievement Test Scores in Math	ns+	N/A	ns+	ns-	ns-
High School Achievement Test Scores in ELA	ns+	N/A	ns+	ns+	ns+
Subgroup results for graduation and postsecondary outcomes (in odds ratios)					
High School Graduation in the Same District Within Four Years	ns-	N/A	ns-	ns-	ns+
Enrolled in Postsecondary Education by Fall 2013	ns-	N/A	ns-	1.357	1.399
Enrolled in a Two-Year Institution	ns+	N/A	ns-	ns+	ns+
Enrolled in a Four-Year Institution	0.710	N/A	ns-	1.763	1.659
Persistence in Postsecondary Education	ns-	N/A	ns-	ns+	ns+

Note: QT refers to the composite measure that was evaluated for the qualifying tests described in Section III.C; ns + denotes a non-significant positive interactive effect; ns - denotes a non-significant negative interactive effect.

Gender. The subgroup analyses revealed that attending a deeper learning network school had significant positive effects on collaboration skills, perseverance, academic engagement, motivation to learn, self-efficacy, and locus of control among female students, but it did not significantly affect these outcomes among male students. (See Exhibit 5.5.) In addition, while attending a deeper learning network school did not have an effect on self-management among female students, it had a negative effect among male students. These findings indicate that attending a deeper learning network school may have had a more positive effect on interpersonal and intrapersonal outcomes among female students than among male students. However, these analyses also show that while attending a deeper learning network school did not have an impact on enrollment in four-year institutions among female students, it had a positive effect among males.

Exhibit 5.5. Interpretation of Significant Subgroup Results: Estimated Effects of Attending a Deeper Learning Network School for Male and Female Students[35]

	Treatment Effect Among Males (B_1)	**Treatment Effect Among Females (B_1+B_2)**	**Difference in Treatment Effect (B_1-[B_1+B_2]=B_2)**
Subgroup results for interpersonal and intrapersonal competency outcomes (in effect sizes)			
Collaboration	-0.026	0.283*	0.307*
Perseverance	-0.147	0.146*	0.333*
Self-Management	-0.171*	-0.011	0.244*
QT: Engagement and Motivation	0.024	0.227*	0.278*
Academic Engagement	0.048	0.305*	0.331*
Motivation to Learn	0.017	0.244*	0.287*
QT: Self-Efficacy and Locus of Control	-0.036	0.213*	0.286*
Self-Efficacy	0.034	0.236*	0.251*
Locus of Control	-0.044	0.162*	0.251*
Subgroup results for postsecondary enrollment outcomes (in odds ratios)			
Enrolled in a Four-Year Institution by Fall 2014	1.605*	1.120	0.710*

* Denotes a significant effect of attending a deeper learning network school in columns one and two and a significant difference in effects in column 3.

Prior Achievement. In order to examine the differential effects of attending a deeper learning network school among students who entered high school with high or low prior achievement, we identified levels of prior achievement using two methods: (1) using district (or state) average test scores; and (2) using within-pair average test scores. (See Exhibit 5.6 and Exhibit 5.7.)

[35] The effect of attending a deeper learning network school among students in the advantaged group (subgroup = 0) is the coefficient for the indicator for attending a network school within the model that includes the interaction term. The effect among students in the disadvantaged group (subgroup = 1) is equal to the sum of the coefficient for the indicator for attending a network school and the coefficient for the interaction term. We summed these two coefficients within each matched pair of schools and then meta-analyzed this new treatment effect coefficient. The standard error of this coefficient—which was calculated within each matched pair of schools—takes into account the standard error of the main effect of attending a network school, the standard error of the interaction term, and the covariance of these two coefficients in the model.

The two methods produced similar results. Attending a network school had a positive, significant effect on academic engagement and motivation to learn among students with high prior achievement, but it had no effect among students with low prior achievement. Moreover, while attending a deeper learning network school did not affect self-management among high-achieving students, it had a significant negative effect among students with low prior achievement. In contrast, attending a deeper learning network school increased the odds of enrolling in postsecondary education, and more specifically enrolling in a four-year institution, among low-achieving students, but it did not have an effect on these outcomes among high-achieving students.

Exhibit 5.6. Interpretation of Significant Subgroup Results: Estimated Effects of Attending a Deeper Learning Network School for Students With Above-Average and Below-Average Incoming ELA Achievement (Defined Based on California State/New York City Average)

	Treatment Effect Among High-Achieving students (B_1)	**Treatment Effect Among Low-Achieving Students (B_1+B_2)**	**Difference in Treatment Effect (B_1-[B_1+B_2]=B_2)**
Subgroup results for interpersonal and intrapersonal competency outcomes (in effect sizes)			
Perseverance	0.113	-0.126	-0.321*
Self-Management	-0.012	-0.164*	-0.212*
QT: Academic Engagement and Motivation to Learn	0.320*	-0.067	-0.370*
Academic Engagement	0.417*	-0.016	-0.395*
Motivation to Learn	0.232*	-0.098	-0.374*
Subgroup results for postsecondary enrollment outcomes (in odds ratios)			
Enrolled in Postsecondary by Fall 2013	0.858	1.226*	1.357*
Enrolled in a Four-Year Institution by Fall 2013	1.089	1.595*	1.763*

* Denotes a significant effect of attending a deeper learning network school in columns one and two and a significant difference in effects in column three.
Results were not estimated for Pairs 10 and 11 due to missing data on students' incoming test scores.

Exhibit 5.7. Interpretation of Significant Subgroup Results: Estimated Effects of Attending a Deeper Learning Network School for Students With Above-Average and Below-Average Incoming ELA Achievement (Defined Based on Averages Within Pair and Cohort)

	Treatment Effect Among High-Achieving Students (B_1)	**Treatment Effect Among Low-Achieving Students (B_1+B_2)**	**Difference in Treatment Effect (B_1-[B_1+B_2]=B_2)**
Subgroup results for interpersonal and intrapersonal competency outcomes (in effect sizes)			
Perseverance	0.128	-0.111	-0.227*
Self-Management	0.040	-0.171*	-0.219*
QT: Academic Engagement and Motivation to Learn	0.259*	-0.014	-0.251*
Academic Engagement	0.345*	0.040	-0.249*
Motivation to Learn	0.199*	-0.051	-0.279*
Subgroup results for postsecondary enrollment outcomes (in odds ratios)			
Enrolled in Postsecondary by Fall 2013	0.885	1.252*	1.399*
Enrolled in a Four-Year Institution by Fall 2013	1.035	1.824*	1.659*

* Denotes a significant effect of attending a deeper learning network school in columns one and two and a significant difference in effects in column three.
Results were not estimated for Pairs 10 and 11 due to missing data on students' incoming test scores.

References

Allison, P. D. (2008, March). Convergence failures in logistic regression. In *SAS Global Forum* (Vol. 360, pp. 1-11).

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, *4*(1), 62–83.

Duckworth, A. L, & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, *91*, 166–174. Retrieved from: http://www.sas.upenn.edu/~duckwort/images/Duckworth%20and%20Quinn.pdf

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Huang, D., Leon, S., Hodson, C., La Torre, D., Obregon, N., & Rivera, G. (2010). *Exploring the effect of afterschool participation on students' collaboration skills, oral communication skills, and self-efficacy*. (CRESST Report 777). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Huberman, M., Bitter, C., Anthony, J., and O'Day, J. (2014). *The shape of deeper learning: Strategies, structures, and cultures in deeper learning network high schools.*

Levenson, H. (1981). Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 15–63). New York: Academic Press.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, *38*, 231–281.

Pintrich, R. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33–40.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 453–458).

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

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., Griffin, B. A., & Burgette, L. (2013). Twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 1.311.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

What Works Clearinghouse. (2013). *Procedures and standards handbook* (version 3.0). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*(2), 1281–1301.

Xue, G., & Sun, X. (2011). Construction and validation of self-management scale for undergraduate students. *Creative Education*, *2*(2), 142–147.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

Zeiser, K.L., Taylor, J. Rickles, J., Garet, M. S., and Segeritz, M. (2014). *Evidence of deeper learning outcomes*.