AMERICAN INSTITUTES FOR RESEARCH®

Providing Opportunities for Deeper Learning: Technical Appendix

Report #2 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

Providing Opportunities for Deeper Learning: Technical Appendix

Report #2 Findings From the Study of Deeper Learning: Opportunities and Outcomes

September 2014

Appendix Authors:

Jordan Rickles, Catherine Bitter, James Taylor, Kristina L. Zeiser, American Institutes for Research

Principal Investigators:

Jennifer O'Day, American Institutes for Research

Michael S. Garet, American Institutes for Research

Study Team:

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Megan Brown
Connie Chandra
Marian Eaton
Alison Hauser
Mette Huberman
Jamie Shkolnik
Allison Waters
Tara Zuber

The Research Alliance for New York City Schools

Michael Segeritz
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, 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 to Marc Chun and Barbara Chow for their consistent support throughout the research process. Our thanks also go to Kerstin Carlson Le Floch and Laura Salganik at AIR for their careful technical review and to 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.

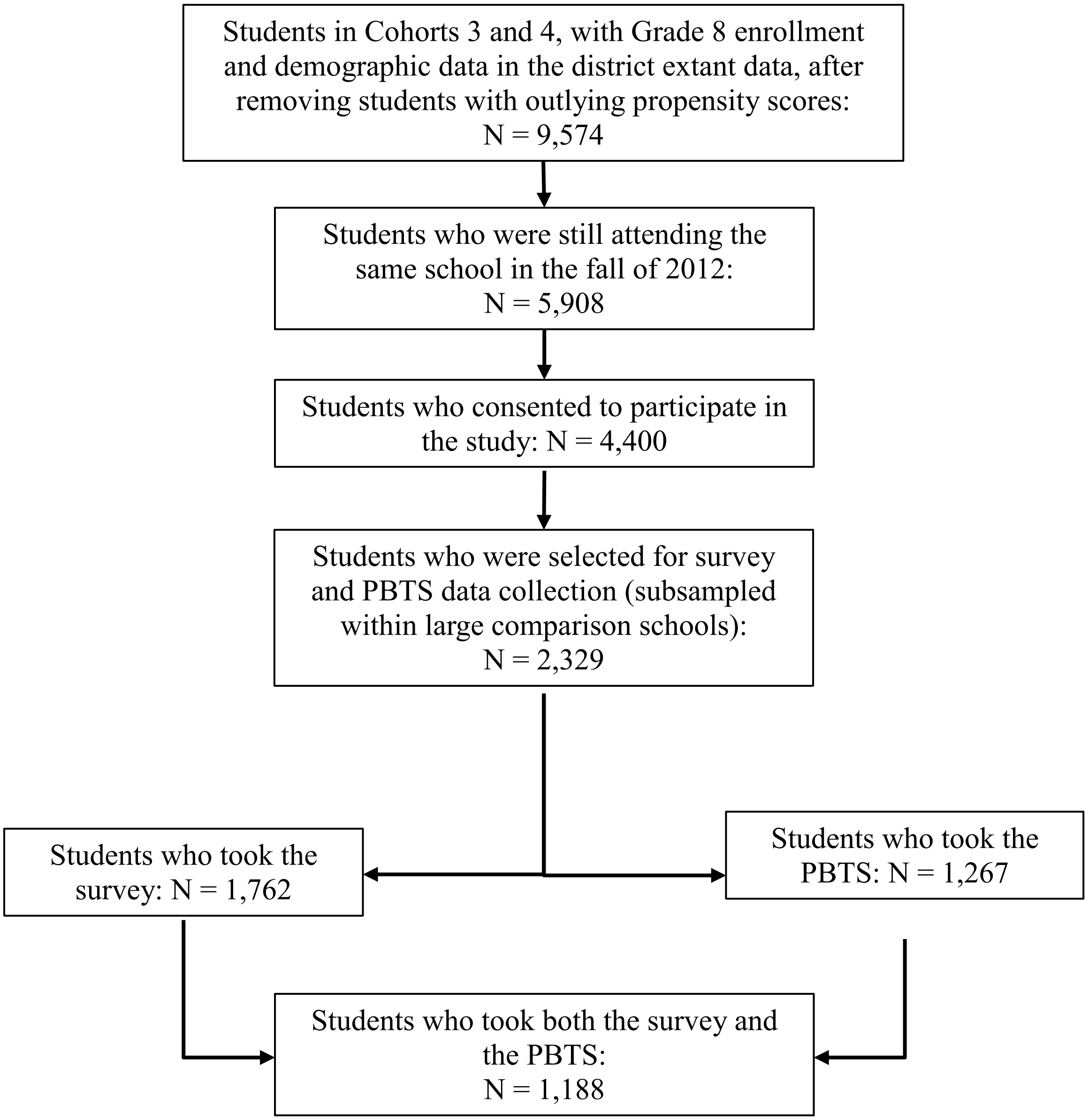
For primary data collection, our goal was to collect data from a total of 260 students within each school pair (65 Grade 11 students and 65 Grade 12 students in the network and non-network schools). We selected student samples for primary data collection based on propensity score quintiles to ensure 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.

[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]



[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.

III. Data Sources and Measures

To address the primary research questions for this study, we collected outcome data from students and teachers. 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.

Exhibit 3.1. Outcome Data Sources and Sample Sizes

Data Source	Description	Sample	Number of Schools	Number of Participants	Analytic Sample Response Rate
Student Survey	Measures students' self-reported opportunities to engage in deeper learning, as well as interpersonal and intrapersonal outcomes (such as self-efficacy and academic engagement)	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
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
Teacher Assignments	Measures deeper learning opportunities present in teacher assignments	Students in Cohorts 3 and 5 (stratified sample of 40 students per school)	Mathematics: 17 schools, 8 school pairs English Language Arts (ELA): 22 schools, 11 school pairs	Mathematics: 359 ELA: 488	Mathematics: 66%[a] overall, 73% network students, 60% non-network students ELA: 63% overall, 71% network students, 55% non-network students

[a] Three pairs of schools were removed from the analysis of mathematics teacher assignments because zero teachers handed in assignments from one non-network school that was matched with two network schools (removing three schools from analyses), and zero teachers handed in assignments in a network school from a third pair (removing two more schools from analyses). These response rates were calculated after removing the three pairs of schools.

A. 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.[4] 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.[5] For the scales that were affected by this technological glitch, we excluded the school pair from the main analyses.

Exhibit 3.2 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

[4] 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.

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

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 3.2. 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

Exhibit 3.2. Descriptive Statistics for Rasch-Scaled Student Survey Constructs

		Rasch Logit Scale			Response Scale		
	N	Mean	SD	ICC	Mean	SD	ICC
Opportunity for Complex Problem Solving (# of Classes – 0 to 3+)	1,681	0.15	1.15	0.10	1.66	0.90	0.09
Opportunities for Creative Thinking (# of Classes – 0 to 3+)	1,745	0.91	2.31	0.13	1.80	0.91	0.12
Opportunities to Communicate (combined) (# of Classes – 0 to 3+)	1,677	0.54	1.48	0.14	1.86	0.97	0.17
Opportunities to Collaborate (# of Classes – 0 to 3+)	1,756	0.95	1.40	0.18	2.14	0.84	0.17
Opportunities to Learn How to Learn (# of Classes – 0 to 3+)	1,680	1.81	1.79	0.12	2.33	0.79	0.08
Opportunities for Feedback to Students (# of Classes – 0 to 3+)	1,751	0.96	1.79	0.15	1.97	0.85	0.13
Opportunities for Assessments Aligned With Deeper Learning (# of Classes – 0 to 3+)	1,746	0.73	1.53	0.17	1.97	0.89	0.15
Opportunities for Interdisciplinary Learning (Frequency Scale – From Never to All the Time)	1,737	-0.40	2.35	0.16	1.30	0.85	0.14
Opportunities for Real-World Connections (# of classes – 0 to 3+)	1,746	0.57	1.77	0.11	1.80	0.85	0.11
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

	Rasch Logit Scale				Response Scale		
	N	**Mean**	**SD**	**ICC**	**Mean**	**SD**	**ICC**
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

Opportunities for Deeper Learning

Opportunities for Complex Problem Solving

(Source: National Survey of Student Engagement[6] *[NSSE])*
Rasch Reliability: .0.90; Cronbach's Alpha: .93

In how many of your English, math, science, and social studies classes this year do you do the following?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I analyze an idea, experience, theory, or story by examining its various parts.	□	□	□	□
I combine many ideas and pieces of information into something new and more complex.	□	□	□	□
I judge the value and reliability of an idea.	□	□	□	□
I use ideas or concepts from one class to help solve a problem in another classroom.	□	□	□	□

Opportunities for Complex Problem Solving in English Language Arts

(Source: Consortium on Chicago School Research[7] *[CCSR])*
Rasch Reliability: .83; Cronbach's Alpha: .89

Think about your English classes you've taken this year. In these classes, how often do you do the following?

	Never	Some Of The Time	Most Of The Time	All Of The Time
I discuss my point of view about something I've read.	□	□	□	□
I discuss connections between what we are reading in class and real-life people or situations.	□	□	□	□
I discuss how culture, time, or place affects an author's writing.	□	□	□	□
I explain how writers use tools like symbolism and metaphor to communicate meaning.	□	□	□	□
I improve a piece of writing as a class or with partners.	□	□	□	□
I debate the meaning of what we are reading in class.	□	□	□	□

[6] http://nsse.iub.edu/html/engagement_indicators.cfm
[7] http://ccsr.uchicago.edu/surveys/documentation?cat=4&content_id=25

Opportunities for Complex Problem Solving in Mathematics

(Source: CCSR)
Rasch Reliability: .71; Cronbach's Alpha: .76

Now just think about your math classes this year. In these classes, how often do you do the following?

	Never	Some of the Time	Most of the Time	All of the Time
I write a few sentences to explain how I solved a math problem.	□	□	□	□
I write a math problem for other students to solve.	□	□	□	□
I discuss possible solutions to problems with other students.	□	□	□	□
I use math to solve real-world problems.	□	□	□	□
I solve a problem with multiple steps that take more than 20 minutes.	□	□	□	□

Opportunities for Complex Problem Solving in Science

(Source: Original)
Rasch Reliability: .86; Cronbach's Alpha: .91

Now just think about your science classes you've taken this year. In these classes, how often do you do the following?

	Never	Some of the Time	Most of the Time	All of the Time
I form hypotheses by asking questions and defining problems.	□	□	□	□
I create physical models representing scientific ideas.	□	□	□	□
I plan and carry out experiments.	□	□	□	□
I interpret data and explain what the results mean.	□	□	□	□
I use equations to help me analyze data or solve a problem.	□	□	□	□
I use data to support a hypothesis or argument.	□	□	□	□
I am required to judge the value and quality of information.	□	□	□	□

Opportunities for Creative Thinking

(Source: Original)
Rasch reliability: .79; Cronbach's Alpha: .88

Still think about your English, math, science, and social studies classes this school year. For how many of these classes is each statement true?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I am encouraged to come up with new and different ideas.	☐	☐	☐	☐
I need to think of original solutions to problems.	☐	☐	☐	☐
I am asked to come up with new ways to do things.	☐	☐	☐	☐
I am challenged to create new ideas.	☐	☐	☐	☐
I have to use my imagination.	☐	☐	☐	☐

Opportunities to Communicate

(Source: Original, based on the Common Core State Standards)
Rasch Reliability: .83; Cronbach's Alpha: .90

How many of your teachers (in your core academic subjects) this year ask you to do the following?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I write for different purposes (for example, to explain or to persuade).	☐	☐	☐	☐
I write for different audiences.				
I write and revise a piece of writing through multiple drafts.	☐	☐	☐	☐
I use technology and the Internet to write and get feedback on our writing (for example, on a message board or blog).	☐	☐	☐	☐
I write what I want in a journal, diary, or blog at least once a week.	☐	☐	☐	☐
I lead a group or class discussion.	☐	☐	☐	☐
I share my opinions in a class discussion.	☐	☐	☐	☐
I give presentations with visual aids, such as pictures, videos, charts, or graphs.	☐	☐	☐	☐
I give presentations.	☐	☐	☐	☐
I give presentations for different types of people, such as other students, parents, or people outside of school.	☐	☐	☐	☐

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I discuss how well other students present their ideas in presentations.	☐	☐	☐	☐
I use information from different types of sources, such as videos, pictures, graphs, charts, and presentations.	☐	☐	☐	☐

Opportunities to Collaborate

(Source: Various, listed below)
Rasch Reliability: .69; Cronbach's Alpha: .93

In how many of your core academic classes this year do you do each of the following?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I work in groups of two to six students. *(NYC student survey[8])*	☐	☐	☐	☐
I work with other students on projects during class. *(NSSE)*	☐	☐	☐	☐
I work on assignments with my classmates outside of class. *(NSSE)*	☐	☐	☐	☐
To do well in class I need to work with others. *(Original)*	☐	☐	☐	☐

Opportunities to Learn How to Learn

(Source: Original, Measures of Effective Teaching[9] [MET] project)
Rasch Reliability: .52; Cronbach's Alpha: .78

Think about the teachers of your English, math, science, and social studies classes this year. For how many of these classes is each statement true?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
My teacher gives us activities to do, other than just listening to him or her. *(MET project)*	☐	☐	☐	☐
My teacher lets me test or try out my ideas to see if they work. *(MET project)*	☐	☐	☐	☐
My teacher helps me learn to use different sources of information.	☐	☐	☐	☐
My teacher asks me to think about how I learn best.	☐	☐	☐	☐

[8] http://schools.nyc.gov/Accountability/tools/survey/2011surveysamples
[9] http://www.metproject.org/

Opportunities to Receive Feedback

(Source: Original, unless otherwise noted)
Rasch Reliability: .75; Cronbach's Alpha: .84

Think about your core academic classes this year ***and the feedback you receive about your work in those classes***. For how many classes is each statement below true?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
My teacher gives me feedback on most of my work.	☐	☐	☐	☐
My teacher gives me specific suggestions about how I can improve my work. *(CCSR)*	☐	☐	☐	☐
I learn a lot from my teacher's feedback on my work.	☐	☐	☐	☐
I get useful feedback on my school work from other students.	☐	☐	☐	☐
I sometimes receive feedback on my work from someone other than the teacher or other students, such as my parents.	☐	☐	☐	☐
My teacher often asks me to revise my work after I get feedback.	☐	☐	☐	☐

Opportunities for Assessments Aligned With Deeper Learning

(*Source: All original items*)
Rasch reliability=.77; Cronbach's alpha = .86

Still thinking about the teachers of your English, math, science, and social studies classes this year, for how many of these classes is each statement true?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
My teacher gives tests about facts that we studied in class.	☐	☐	☐	☐
My teacher gives tests at the beginning of a unit to see how much we already know.	☐	☐	☐	☐
My teacher gives tests that require us to use different sources of information for our answers.	☐	☐	☐	☐
My teacher gives us points on a test or homework for how we solved a problem, not just whether we got the right answer.	☐	☐	☐	☐
My teacher asks us to put together a portfolio of many different examples of our school work.	☐	☐	☐	☐
My teacher evaluates us on how well we work in groups.	☐	☐	☐	☐
My teacher asks us to evaluate ourselves on our class work.	☐	☐	☐	☐
My teacher asks us to explain our thinking.	☐	☐	☐	☐
My teacher has conferences with just me (not with my parents) so I can talk about what I'm learning in class and how well I'm doing.	☐	☐	☐	☐

Opportunities for Interdisciplinary Learning

(Source: Original, NSSE, Buck Institute for Education[10] [BIE] Project based Learning [PBL])
Rasch Reliability: .78; Cronbach's Alpha: .82

Still thinking about your English, math, science, and social studies classes this year, how often do you do the following?

	Never	Some of the Time	Most of the Time	All of the Time
I work on a project that combines more than one subject (for example, science and literature). *(Original)*	☐	☐	☐	☐
I put together ideas or concepts from different subjects for assignments or discussions. *(NSSE)*	☐	☐	☐	☐
I attend a class that two teachers from different subjects teach together (for example, a math teacher and a science teacher). *(BIE PBL)*	☐	☐	☐	☐
I use ideas or concepts from one class to help solve a problem in another class. *(Original)*	☐	☐	☐	☐

Opportunities for Real-World Connections

(Source: Various, listed below)
Rasch Reliability: .84; Cronbach's Alpha: .89

Regarding the work you do for your core academic classes (such as English, math, science, and social studies) this year, in how many classes does the following happen?

	None of My Classes	One of My Classes	Two of My Classes	Three or More of My Classes
I make observations or collect data outside of the classroom for assignments. *(BIE PBL)*	☐	☐	☐	☐
I interview or get information from family or community members. *(BIE PBL)*	☐	☐	☐	☐
We connect what we are learning to life outside the classroom. *(CCSR)*	☐	☐	☐	☐
I work on helping solve real-world problems. *(Original)*	☐	☐	☐	☐
I find information for a project from sources outside of school. *(Original)*				
We discuss how someone could use something we learned in school in a real job. *(College Student Experiences Questionnaire)*	☐	☐	☐	☐
I can apply what I learn in class to my life outside of school. *(Distance Education Learning Environments Survey[11] [DELES] personal relevance scale)*	☐	☐	☐	☐

[10] http://bie.org/

I am able to pursue topics that interest me. *(DELES personal relevance scale)*	☐	☐	☐	☐
I use my out-of-class experiences in my classes. *(DELES personal relevance scale)*	☐	☐	☐	☐
I work with real-world examples in class work. *(Original)*	☐	☐	☐	☐
I often feel my class has nothing to do with real life outside of school. *(MET study)*	☐	☐	☐	☐

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.	☐	☐	☐	☐

[11] http://tcet.unt.edu/insight/ilib/deles/actual/

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.	☐	☐	☐	☐
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)*	☐	☐	☐	☐

[12] http://www.excellenceandethics.com/assess/CEEA_v4.5_matrix.pdf
[13] http://cseq.iub.edu/pdf/cseq_whole.pdf

	Never or Almost Never True	Sometimes True	Usually True	Always or Almost Always True
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)*	☐	☐	☐	☐

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.	☐	☐	☐	☐

B. 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.[14]

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.[15]

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 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

[14] 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.

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

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.[16] 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.[17] 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.[18]

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.

---

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

[17] 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.

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

Exhibit 3.3. 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

C. Teacher Assignments

Teacher assignments were collected from ELA and mathematics teachers of the Grade 10 and Grade 12 students in our sample. For most schools, 40 students were sampled—20 from each grade. Within each grade, we categorized students into quintiles based on their predicted propensity score (described below) and randomly sampled four students per quintile. We worked with the schools to identify the ELA and mathematics teachers for each sampled student. Teachers were contacted and asked to provide a copy of the most challenging assignment given to each of the sampled students during the spring semester of the 2012–13 school year.[19] We requested the most challenging assignment in order to have an authentic measure of the most challenging types of instructional activities in which students were engaged at both network and non-network schools. We did not specifically ask for assignments that met the criteria on which we planned to rate the assignments because we wanted to determine whether these deeper learning competencies appeared in students' challenging assignments. In classes where all sampled students received the same assignment, teachers were asked to provide one additional challenging assignment. This occurred in most classes.

The assignments were scored by an ELA and mathematics content area expert using a rubric adapted from one created by the Consortium for Chicago School Research (Newmann, Secada, and Wehlage, 1995). The ELA teacher assignment rubric included ratings for four criteria and the mathematics rubric included ratings for five criteria. (See Exhibit 3.4.) The criteria were rated on either a 1 to 4 scale or a 1 to 3 scale, with a higher number indicating that the assignment was more strongly related to the specified deeper learning competency. Additional details about the score rubric criteria are available upon request. We used a Rasch model to create a composite overall score for each subject, based on the individual criteria ratings. We estimated Rasch scale scores from the individual criteria ratings using an ordered logit Rasch model implemented with the WINSTEPS software package. The resulting Rasch scale scores are in logits and have both negative and positive values. The value of zero is anchored to the average

[19] Specifically, teachers were asked "to submit at least two teacher assignments. The two assignments should be your most challenging assignments of the semester. The assignments should be the ones that you think show what the student knows and can do at a high level. In cases where students receive individualized assignments, more than two assignments per teacher will be needed." Three schools provided teacher assignments in fall 2012 and in spring 2013; the others provided assignments in spring 2013 only. Since the level and type of instruction may have changed between fall and spring, we restricted the main analyses to the spring assignments.

difficulty of the four (in ELA) or five (in mathematics) criteria. In general, an assignment with a positive score tended to exhibit above average exposure to OTL, and an assignment with a negative score tended to exhibit below average exposure.

Exhibit 3.4. Teacher Assignment Criteria and Ranges in Values

Assignment Asks Students to Show:	ELA	Mathematics
Effective Communication Skills	1 to 4	1 to 3
Critical Thinking Skills	1 to 3	1 to 4
Independent Learning Skills	1 to 4	1 to 4
Real-World Connections	1 to 4	1 to 4
Understanding of Core Content	N/A	1 to 4
Overall Rasch Scale Score	-4.9 to 1.7	-8.1 to 3.2

We collected teacher assignments from 11 of the matched pairs (22 schools). In 3 of the 22 paired schools, we were not able to get assignments from mathematics teachers. Overall, we received ELA assignments for 65 percent of the sampled students and mathematics assignments for 47 percent of the sampled students. Among the paired schools from which we received at least one ELA and one mathematics assignment, we received ELA and mathematics assignments for 65 percent of the sample students. Summary statistics for each of the rubric criteria are presented in Exhibit 3.5 (ELA) and Exhibit 3.6 (mathematics). Given the skewed distribution of the individual criteria ratings, we dichotomized the individual criteria when testing for statistically significant differences between network and non-network students: students with a 1 (the lowest score) were given a value of zero and students with any higher score were given a value of 1. For the “understanding of core content” criteria in mathematics, students with a score of 1 or 2 were given a value of zero, and students with a score of 3 or 4 were given a value of 1. To illustrate the types of assignments collected, we provide summaries of two exemplary teacher assignments (one in ELA and one in mathematics) following Exhibit 3.6.

Exhibit 3.5. Descriptive Statistics for ELA Assignments (N = 148 Assignments)

	Mean	S.D.	Min	Max
Criterion 1: Effective, Elaborated Written Communication	3.32	0.82	1.00	4.00
Criterion 2: Construction of Knowledge	2.30	0.75	1.00	3.00
Criterion 3: Connection to the World Beyond School	1.45	0.57	1.00	4.00
Criterion 4: Independent Learning	1.15	0.38	1.00	3.00
Rasch Scale Score	-1.27	1.60	-4.94	1.65

Exhibit 3.6. Descriptive Statistics for Mathematics Assignments (N = 148 Assignments)

	Mean	S.D.	Min	Max
Criterion 1: Conceptual Understanding of Core Content	2.15	0.39	1.00	4.00
Criterion 2: Critical Thinking	1.41	0.72	1.00	4.00
Criterion 3: Effective Communication About Mathematics	1.72	0.68	1.00	3.00
Criterion 4: Independent Learning	1.24	0.48	1.00	3.00
Criterion 5: Real-World Connections	1.60	0.68	1.00	3.00
Rasch Scale Score Combined TA Score (Criteria 1–5)	-3.34	3.35	-8.10	3.22

Box 1: Exemplary Assignment From an ELA Teacher

ELA Assignment Name: Interdisciplinary Food Unit—You Are What You Eat
Class: English, Grade 12
School: Network

Interdisciplinary Unit: Students participate in an interdisciplinary unit on "You Are What You Eat," studying the effects of food choices on health across all five of their classes. In English, students read "Chew On This" by Eric Schlosser and Charles Wilson—a book that discusses how the hamburger became America's favorite sandwich, the mistreatment of animals in slaughterhouses and employees in restaurants, the shocking effects too much fast food can have on growing bodies, and the impact of the industry on schools and communities. The assignment focuses on developing skills in non-fiction reading and persuasive writing.

The Task
***Part 1*:** Please select two chapters to read from the text (listed in the assignment) with a group or a partner in your class. Read with your pencil and underline important words and phrases. For each chapter you read, please choose one writing activity, one creative activity, and the viewing guide to complete. You must complete one of each, but for different chapters. The writing activities are individual and the creative activities and viewing guide should be completed with your reading pair or group.

Writing activity options:

1) Complete the strategies graphic organizer for the chapter you read.
2) Write a paragraph persuading the reader of the author's main argument in the chapter you read. Use at least three persuasive strategies.
3) Write a dialogue between two fictional characters. One of the characters must persuade the other of the argument presented in your chapter. Use at least three persuasive strategies.

Creative activity options:

1) Design a public service announcement poster based on the argument presented in the chapter you read. Use at least three persuasive strategies.
2) Create and perform a skit that includes the main argument of the chapter that you read. The skit should persuade the viewer of the argument. Use at least three persuasive strategies.
3) Design a banner for the "Be the Change" website on bannersnack.com. Your banner should persuade the reader of the main argument of the chapter and use at least three persuasive strategies.

Viewing Fast Food Nation
Watch Fast Food Nation and complete the handout/viewing guide with questions about the narratives and themes presented in the film.

Part 2: Take action and write a persuasive letter, either in favor of or against a health-related bill (three options provided) to your elected official. The thesis statement of your letter should be supported by evidence from at least three of your classes.

Assessment: The persuasive letter is scored using a rubric including evidence/analysis, counterargument, organization, and conventions.

Box 2: Exemplary Assignment From a Mathematics Teacher

Mathematics Assignment Name: Bridges and Parabolas
Class: Algebra 2, Grades 10 and 12
School: Network

Project: Students were given three different projects to choose from for their final project. This is one of these project options, and it explores the application of quadratic functions with bridge designs.

Essential Question: How can a quadratic function model the shape created by suspension bridges—specifically the Golden Gate Bridge, the Brooklyn Bridge, and a third bridge of your choice?

Background: Suspension bridges are bridges that have suspended cables between towers that help carry the weight of the bridge upon which traffic crosses. This type of bridge dates back to the early 19th century. Suspension bridges are also very aesthetically pleasing as they naturally create parabolic curves. One of the most famous suspension bridges is the Golden Gate Bridge. Another well-known suspension bridge is the Brooklyn Bridge. Photos of the bridges and specifications of the bridges are provided, including dimensions of bridges and cables.

The Task: For this Graduation Portfolio project option, your assignment is to graph the parabolas created by each bridge and find the equation for the parabola that models the main cables of the Golden Gate Bridge and the equation of the parabola that models the main cables of the Brooklyn Bridge. You will then use these representations to compare the two bridges.

Investigate the parabolic curve created by the Golden Gate Bridge and the Brooklyn Bridge. **Research** another suspension bridge and then investigate the parabolic curve created by its cables.

- Generate multiple representations of the curve formed by the main cables of each bridge. These representations must include a **table, graph, AND equation.** Explain in detail how you generated each representation for both bridges.

Analyze the mathematical representations of the three bridges and the causes for their similarities and differences.

- Compare the **similarities** and **differences** of the design of the three bridges, as represented in their equations, graphs, and tables. Determine the causes for the differences in the mathematical representations. This will require researching information about all the bridges. Include the purpose of each bridge.

Assessment: Student provides a problem statement and a written product (several pages). The written product is scored (as emerging, developing, proficient, advanced) using a rubric along five domains: problem solving, reasoning and proof, connections, communication and representation, and reflection. The first draft is peer reviewed.

D. 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.[20] Exhibit 3.7 lists the student background data we received from districts and 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 9 of the 11 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 included in the California and New York City study samples is presented in Exhibit 3.8. The descriptive statistics represent all Cohort 3 and Cohort 4 students in the study schools, before weighting had been applied to adjust for differences between network and non-network schools. As discussed later in this appendix, student background characteristics were used in the estimation of weights and as covariates in analytic models.

[20] 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.

Exhibit 3.7. Description of Student Background Data From Extant District Data

Measure	Description	Number of School Pairs With Available Data
Female	Dichotomous indicator of students' gender	11
Race/Ethnicity	Dichotomous indicators created for African American, Hispanic, white, Asian, and "other" race/ethnicities	11
Parents' Education	Categorical measure of parental education—specifically, the highest level 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
Free or Reduced-Price Lunch (FRPL) Status	Dichotomous indicator of whether the student was eligible for the free or reduced-price lunch program, typically in Grade 8	6
English Language Learner (ELL)	Dichotomous indicator of whether the student was identified as an English language learner, typically in Grade 8	11
Individualized Education Plan (IEP)	Dichotomous indicator of whether the student had an Individualized Education Plan, typically in Grade 8	11
Prior Achievement in ELA	Standardized test score in English Language Arts (ELA) prior to entering high school, from Grade 8	9
Prior Achievement in Mathematics	Standardized test score in mathematics prior to entering high school, from Grade 8, including indicators for mathematics test subject where relevant	9
Grade 8 Attendance Rate	Proportion of enrolled school days attended during Grade 8	4
Age	Age of student (in months) when first enrolled in Grade 9	4

Exhibit 3.8. 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	-0.116
Grade 8 ELA	1,061	-0.121	6,791	0.019	351[a]	0.012	354	-0.012
Attendance Rate		N/A		N/A	351[a]	91.9%	354	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.

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.

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 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 3.7 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. For students attending a non-network school, among students within the survey sample, the IPTW weight ranged between 0.002 and 4.047, with a mean of 0.35.

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 3.7, 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. For students attending network schools, attrition/non-consent weights ranged from 1.04 to 5.08, with a mean of 1.53. For students attending non-network schools, attrition/non-consent weights ranged from 1.04 to 16.02, with a mean of 2.28.

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.

[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.

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.

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

Overall, 76 percent of students selected for student survey data collection returned a survey (80 percent in network schools and 73 percent in non-network schools). To account for non-response in our analysis, we calculated a non-response weight based on the inverse probability that a student sampled for the survey completed the survey. To calculate a student's probability of completing the survey, we used the generalized boosted regression approach described above for calculating a student's probability of persisting and consenting.

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

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

With this weight, students with survey data were weighted to represent the target student sample. Among students attending network schools, non-response weights ranged from 1.02 to 3.44, with a mean of 1.2. Among students attending non-network schools, non-response weights ranged from 1 to 3.81, with a mean of 1.24.

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.[22] 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 4.1 presents summary statistics for all of the individual weights as well as the final analytic weights (before and after trimming) for the survey sample.

[22] 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 4.1. Descriptive Statistics for Individual and Combined Survey Weights, 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.

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 is provided in Exhibit 4.2. 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:

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

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 4.2. Network and Non-Network Student Balance Characteristics Before and After Weighting[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.

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.[23] 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 3.7, 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.

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

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

[23] 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.

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).[24] 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 this design would allow us to detect effect sizes of 0.07 to 0.09 standard deviations for the survey results, depending on the assumed values of key parameters. (See Exhibit 4.3.) After completing the analyses, the realized minimum detectable effect size (MDES) for the student survey OTL measures was between 0.13 and 0.15, depending on the OTL measure. For the teacher assignment analysis, which involved a much smaller sample by design, the realized MDES was 0.41 for mathematics assignments and 0.34 for ELA assignments.

Exhibit 4.3. Minimum Detectable Effect Size (MDES) Using a Fixed Effects Model, Based on Different Assumptions About the Percentage of Variance Explained by Blocking (Schools 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
- Cohort: Grade 11 (Cohort 4) versus Grade 12 (Cohort 3)
- Free/Reduced-Price Lunch Status: eligible versus not eligible
- 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[25] test score for the year in which the test was taken. Students were classified as low achieving if their

[24] 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.

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

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.

Analysis of the Relationship Between Opportunities and Outcomes

We conducted regression analyses to describe the relationship between student opportunities to engage in deeper learning and student deeper learning outcomes. For this purpose, we combined data from multiple pairs into the same analysis to maximize the sample available and to make it possible to examine variation in opportunity to learn (OTL) both across schools and within schools. Given that we were interested in the general relationships among the OTL and outcome measures (rather than the contrasts between network and non-network schools), we did not include indicators for school pairs or for the "network school attendance" designation in the analysis.

To examine the relationship between deeper learning OTL and outcomes, we selected a model that allowed us to:

- Account for measurement error in the dependent variable based on user-supplied variances[26]
- Account for the nesting of students within schools[27]

[26] We accounted for measurement error in the dependent variable to reflect the design of the PBTS. Each student received one of seven PBTS test forms, and each test form contained a subset of items on mathematics, reading, and science. Since each student's estimated score on the PBTS was based on a subset of items, the measurement error was larger than might be expected in a typical standardized test, and the magnitude of the error variance differed across students. The model we estimated took both of these issues into account. To apply a consistent model across all student outcomes, we also accounted for measurement error in the models of the association between OTL and the survey outcomes.

[27] Unlike the main analysis of the effects of attending a network school (which involved comparing a single network school and a single non-network school within a matched pair), the analyses of the association between OTL and outcomes involved data from multiple schools.

- Adjust for student-level, Grade 8 characteristics that could confound the relationship between OTL measures and outcomes
- Provide estimates for both the within-school and between-school relationship between OTL measures and outcomes

We used a three-level, variance-known, hierarchical linear model (Raudenbush and Bryk, 2002) to estimate the relationship between a given OTL measure[28] and a given outcome (Y):

Level 1: $Y_{ijk} = \pi_{0jk} + e_{ijk},\ e_{ijk} \sim N(0, \sigma^2_{ijk}),$

Level 2: $\pi_{0jk} = \beta_{00k} + \beta_{01}(OTL_{jk} - \overline{OTL}_{.k}) + \beta_{02}(X_{jk} - \overline{X}_{..}) + r_{0jk},\ r_{0jk} \sim N(0, \tau_0),$

Level 3: $\beta_{00k} = \gamma_{000} + \gamma_{001}(\overline{OTL}_{.k}) + u_{00k},\ u_{00k} \sim N(0, \tau_{00}),$

where *i* indexes outcome measures, *j* indexes students, and *k* indexes schools. σ^2_{ijk} is the known variance for a given outcome based on the measure's standard error for each student. At Level 2, OTL is centered on the measure's school mean to capture within-school variation in OTL, and X represents a vector of student-level covariates that are centered on their grand mean to adjust for differences in Grade 8 student characteristics within and between schools. At Level 3, school-mean OTL is included in the model to capture between-school variation in OTL. The models were executed using the STATA gllamm package.

Data access requirements precluded us from combining student data from California schools and New York City schools, so we ran separate models for each. In California, 1,202 students in 14 schools took the survey (Network N=493) and 747 students in 13 schools took both the survey and the PBTS (Network N=335). In New York City, 373 students in eight schools took the survey (Network N=194) and 244 students in seven schools took both the survey and the PBTS (Network N=128).[29]

Since the available data on student characteristics differ across districts, we could not include all the covariates listed in Exhibit 3.7.

For the schools in California, the following covariates were included in the X vector:

- Cohort (Grade 11 or 12)
- Gender
- Ethnicity
- IEP status
- ELL status
- Grade 8 test score in English

[28] For these analyses, each OTL measure was included in a separate model.

[29] The sample sizes reported here differ from those reported in Section II because of the treatment of the non-network schools in Pairs 5 and 11. As described in Exhibit 2.2, for the main pair-specific analyses, we included the students from three non-network schools twice because they were matched with students from more than one network school. Because this is not a pair-specific analysis, we counted each comparison student once.

For models predicting PBTS mathematics scores, the covariates also included the Grade 8 mathematics test score along with test type indicators; and for models predicting PBTS science scores, covariates included the Grade 8 science test score. We did not include a measure for parents' education or FRPL status because we did not have these measures for all schools in California.

For the schools in New York City, the following covariates were included in the *X* vector:

- Cohort (Grade 11 or 12)
- Gender
- Ethnicity
- IEP status
- ELL status
- FRPL status
- Age
- Grade 8 attendance rate

We did not include a measure of Grade 8 test scores because these data were not available for all schools in New York City.

Teacher Assignment Analysis

The analysis of teacher assignments was designed to compare the assignments received by students who attended network and non-network schools. To accomplish this, we weighted each assignment based on the number of sampled students who were represented by the assignment. For example, if an assignment was given to students who took Algebra I in the first and second class periods within a school, and eight sampled students were identified as taking Algebra I during those class periods, the assignment received a weight of 8. While each school was asked to provide the class periods when students took mathematics and ELA classes, nine schools did not provide this information. In these schools, we could identify how many students were assigned to each teacher, but we could not determine the number of students within specific class periods. If teachers in these schools handed in more than one "most difficult" assignment and/or more than one "second most difficult" assignment, all of the students assigned to that teacher were evenly divided across assignments so that the sum of the weights would be equal to the total number of sampled students taught by that teacher.

For the analysis of Rasch scale scores, the scores were standardized using the overall mean and standard deviation. The difference in the overall quality of the assignments received by students in network and non-network schools was estimated with the following three-level hierarchical linear model:

Level 1: $Y_{ijk} = \pi_{0jk} + e_{ijk},\ e_{ijk} \sim N(0, \sigma_{ijk}^2),$

Level 2: $\pi_{0jk} = \beta_{00k} + \beta_{01}M_{jk} + \beta_{02}G_{jk} + \beta_{03}I_{jk} + r_{0jk},\ r_{0jk} \sim N(0, \tau_0),$

Level 3: $\beta_{00k} = \gamma_{000} + \gamma_{001}T_k + \gamma_{002}P_k + u_{00k},\ u_{00k} \sim N(0, \tau_{00}),$

where *i* indexes the scale score, *j* indexes assignment, and *k* indexes teachers. σ^2_{ijk} is the known variance for a given scale score based on the measure's standard error for each assignment. At Level 2, we controlled for characteristics of the assignment, including: a dichotomous indicator for whether it was submitted as the "most challenging" (*M*); a set of indicators for whether the assignment was for Grade 10, Grade 12, or both (*G*); and for ELA assignments only, a dichotomous indicator for whether the assignment was an imaginative writing assignment (*I*).[30] At Level 3, we included the dichotomous indicator for whether the teacher was in a network school or non-network school (*T*), and a vector of indicators for school pair (*P*). We estimated separate models for mathematics and ELA assignments. The models were executed using the V-known (variance-known) weighting function in the HLM statistical package.

In addition to the Rasch score of overall assignment quality, we also assessed network effects on each of the specific ratings criteria for mathematics and ELA assignments. For this set of analyses, the criterion-specific ratings were dichotomized and analyzed using a two-level linear probability model specified similarly to the model for Rasch scores (the main difference being the omission of the Level 1 V-known model).[31]

In all of the analyses described above, the number of students associated with each assignment was used as a Level 1 weight.

Sensitivity Analyses

We conducted a number of analyses to check the sensitivity of the main results to specific analytic decisions.

First, as a sensitivity analysis for the main analysis of the effects of attending network schools on students' OTL measures, we removed pairs in which (1) the response rate for the student survey fell below 70 percent; and/or (2) the difference in response rates between the matched network and non-network school was larger than 10 percentage points. This resulted in the removal of 5 of the 11 matched pairs of schools. Analyses using the remaining six pairs resulted in similar findings for most of the OTL measures (i.e., opportunities for complex problem solving, opportunities for communication, opportunities for collaboration, opportunity to learn how to learn, feedback to students, interdisciplinary learning, and real-world connections). However, the differences between network and non-network students for the OTL outcomes of assessments aligned with deeper learning and opportunities for creative thinking were no longer significant within this subset of six school pairs.

In addition, we ran sensitivity analyses for the main analysis of effects of network school attendance on OTL outcomes, changing the way in which results for the 11 pairs were averaged.

[30] For ELA assignments, the rating rubric for "imaginative" tasks differed slightly from the rubric for "expository" tasks because the assignments tend to have different goals.

[31] A linear probability model was used (rather than a logit model) for binary criterion ratings to overcome the separation problem, which occurred when all assignments in the treatment or the comparison group had a value of 1 or all assignments had a value of 0 on a given criterion in some of the school pairs, and thus a logit model could not be estimated.

The main analysis averaged the 11 pair effects, weighting each based on precision, as described above. For the sensitivity analysis, we equally weighted the 11 effects. The findings in these sensitivity analyses closely resembled the main findings. Detailed results are available upon request.

We also conducted sensitivity analyses for the analyses of the association between OTL and outcomes. We compared results from the main analysis—which was conducted without including weights for attrition and non-response—with results from an analysis that included weights for these factors. We also conducted a sensitivity analysis that excluded Grade 8 test scores from the California model because Grade 8 test scores were not available in the New York City model. We did not find any meaningful differences in the results under these alternative specifications.

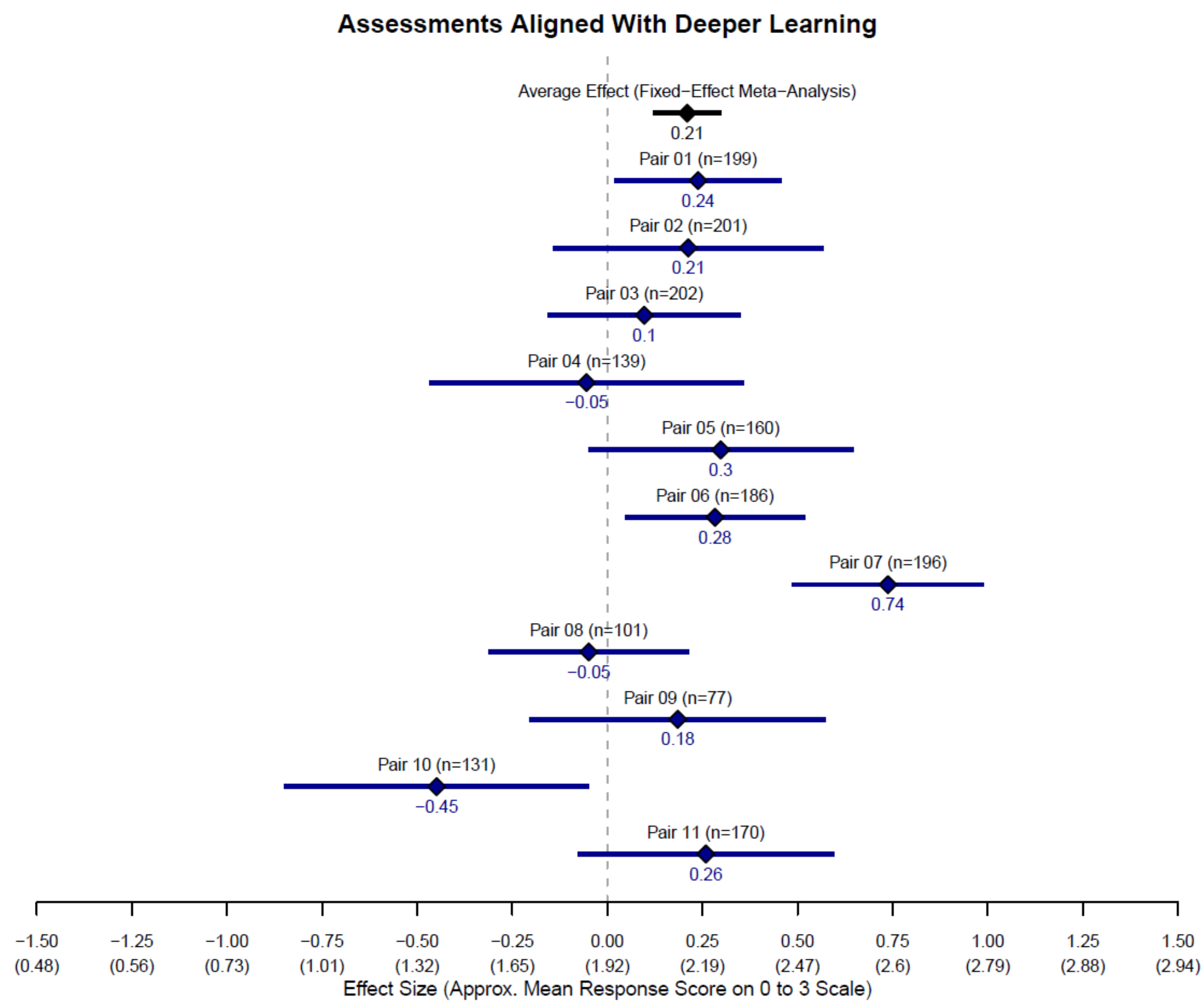
Finally, we conducted a sensitivity analysis for the effects of attending a network school on teacher assignments in ELA. In particular, we estimated a statistical model that controlled for whether the ELA assignment was coded as "imaginative." ELA assignments that were "imaginative" in nature had a distinct scoring rubric, though assignments were scored based on the same criteria and on the same scales. We did not control for "imaginative" assignments in the main analysis because this was considered a post-treatment characteristic: network students were significantly more likely to receive "imaginative" assignments than non-network students. In this sensitivity analysis, the difference between network students and non-network students in independent learning—which was significant in the main analysis—was no longer significant at the 0.05 level (though it remained significant at the 0.10 level).

V. Detailed Results

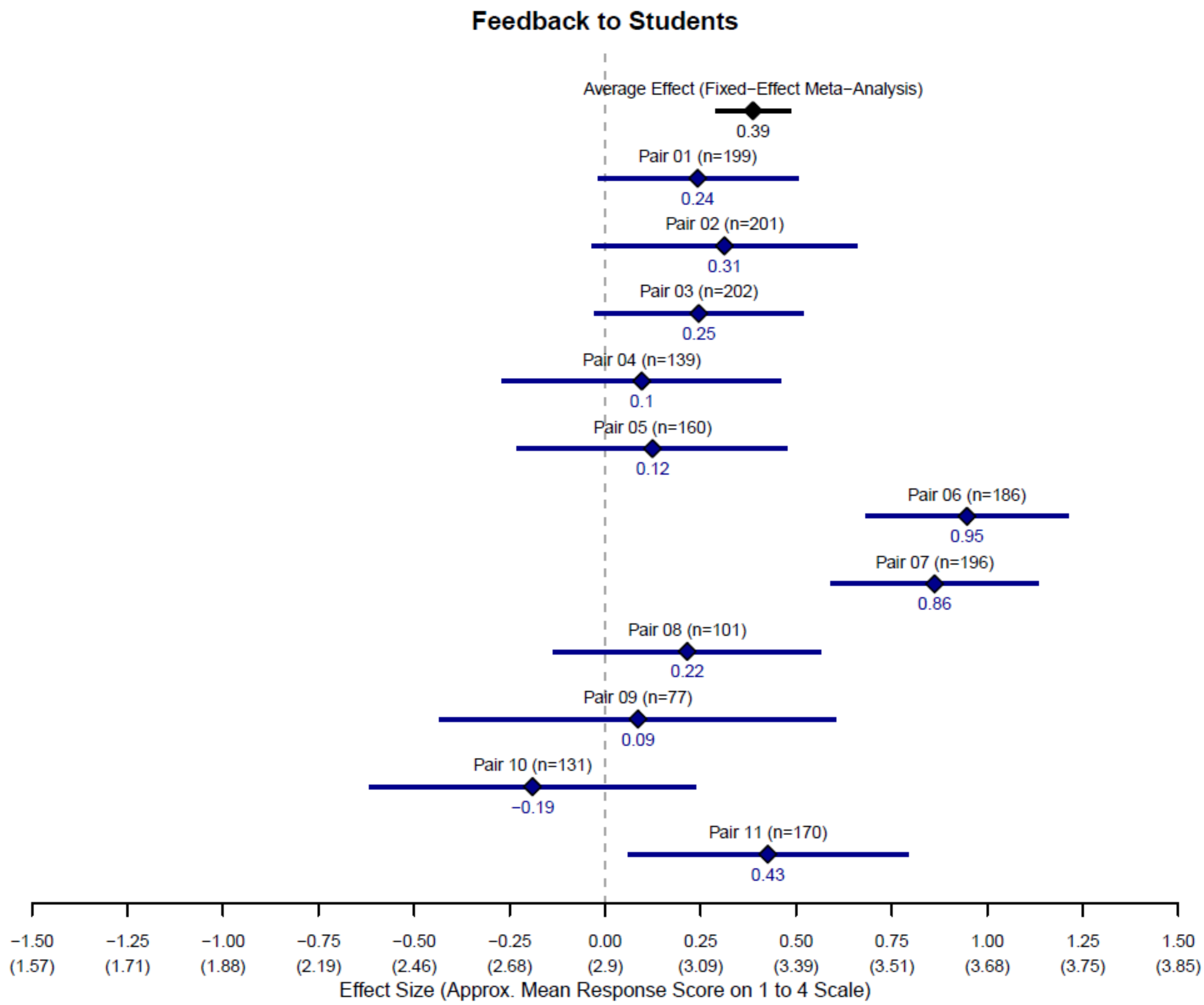
In this section, we provide supplemental figures and tables presenting more detailed information for the results described in the report.

A. OTL Effect Estimates: Pair-Specific Results

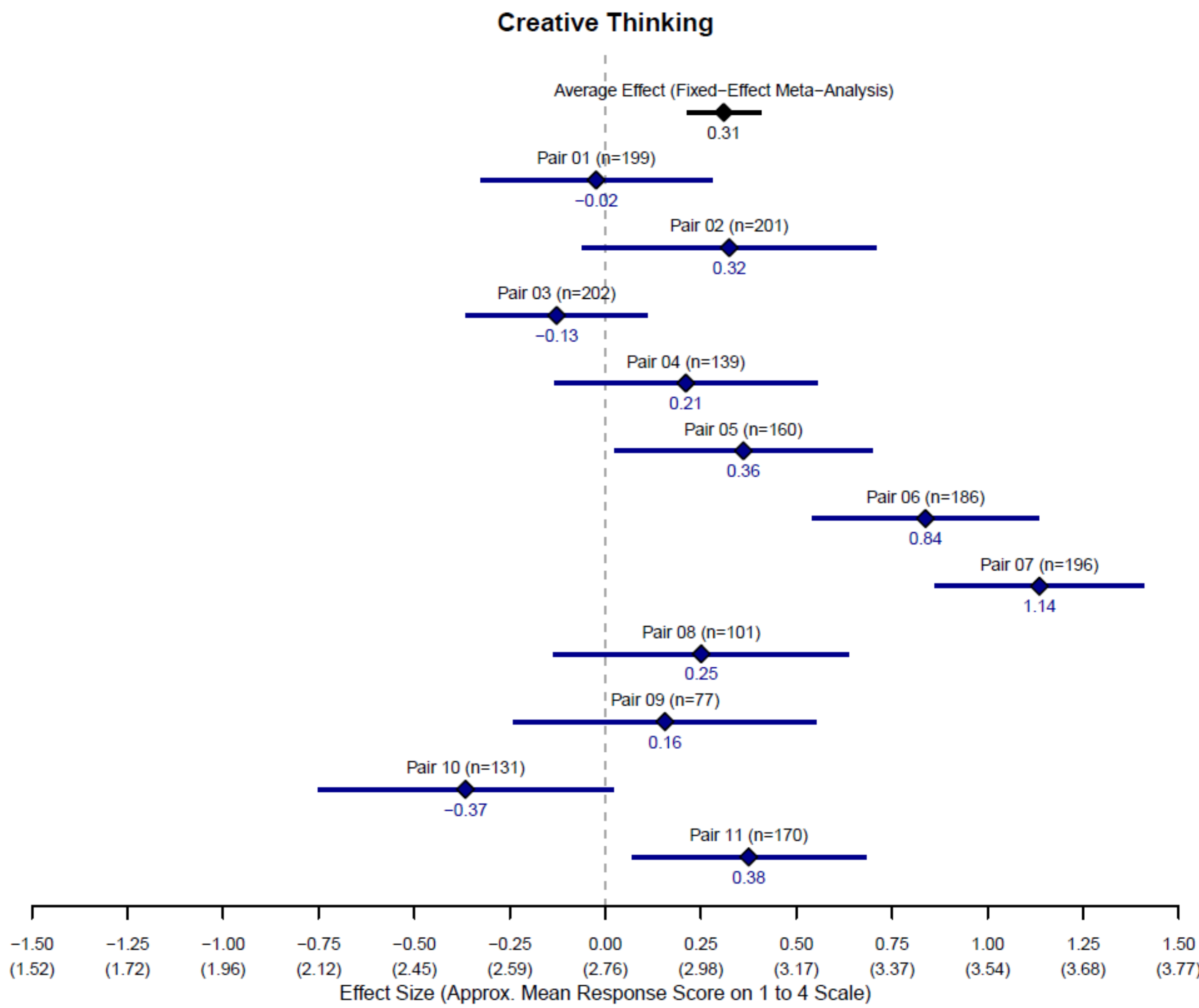
This section provides forest plots that display the meta-analytic average estimate and the pair-specific estimates of the differences in OTL measures for students attending network and non-network schools. Estimates for each measure of opportunities for deeper learning are presented in separate forest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



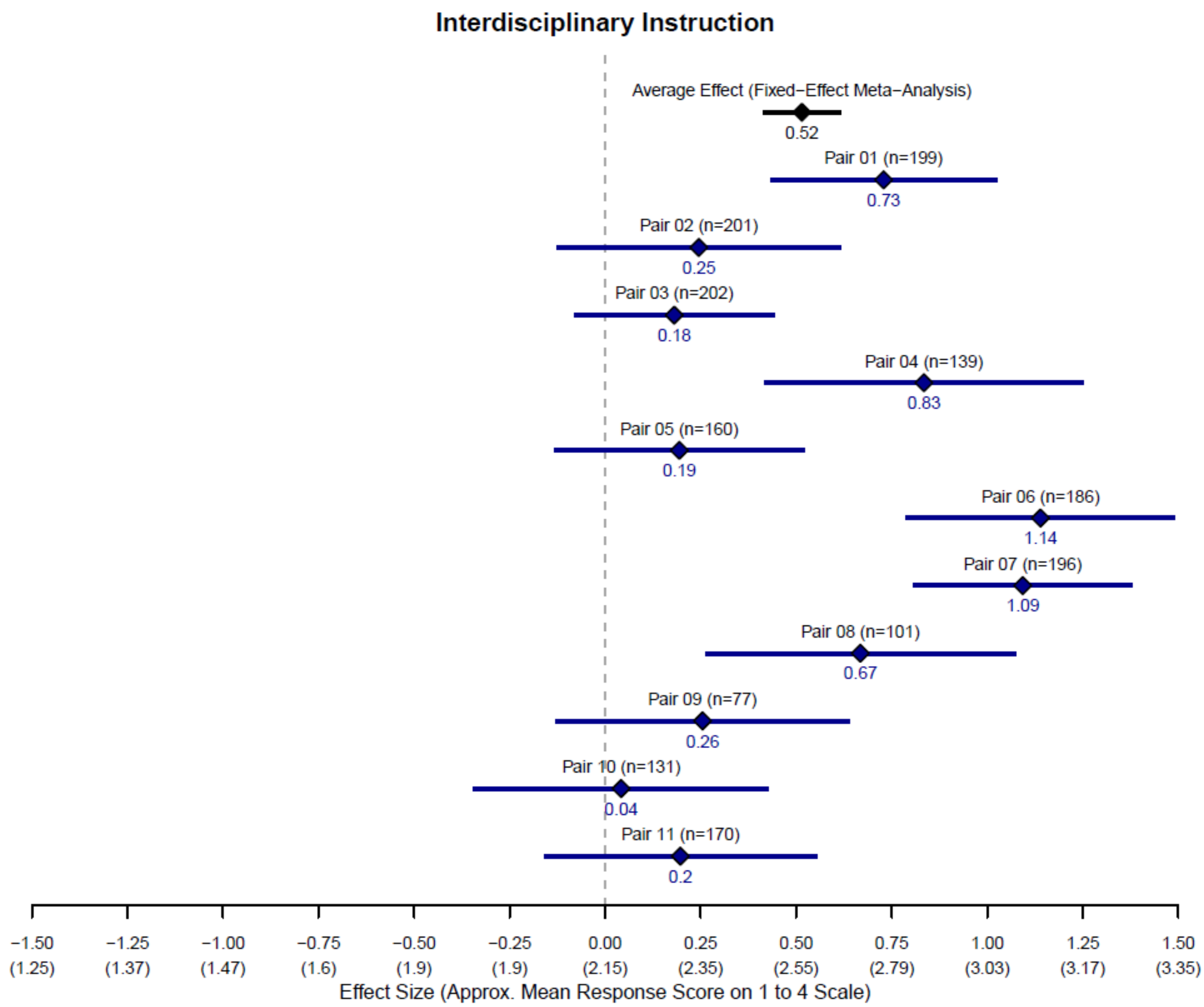
Note: Because we examined the effect of attending a network school on multiple measures of OTL, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on five OTL measures: assessments aligned with deeper learning, feedback to students, creative thinking, interdisciplinary learning, and real-world connections. The qualifying test involved examining the impact of network school attendance on a composite of the five measures. The impact on the composite measure was significant for Pairs 1, 6, 7, and 8, but not Pair 10. Thus, the result for assessments aligned with deeper learning for Pair 10 (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.



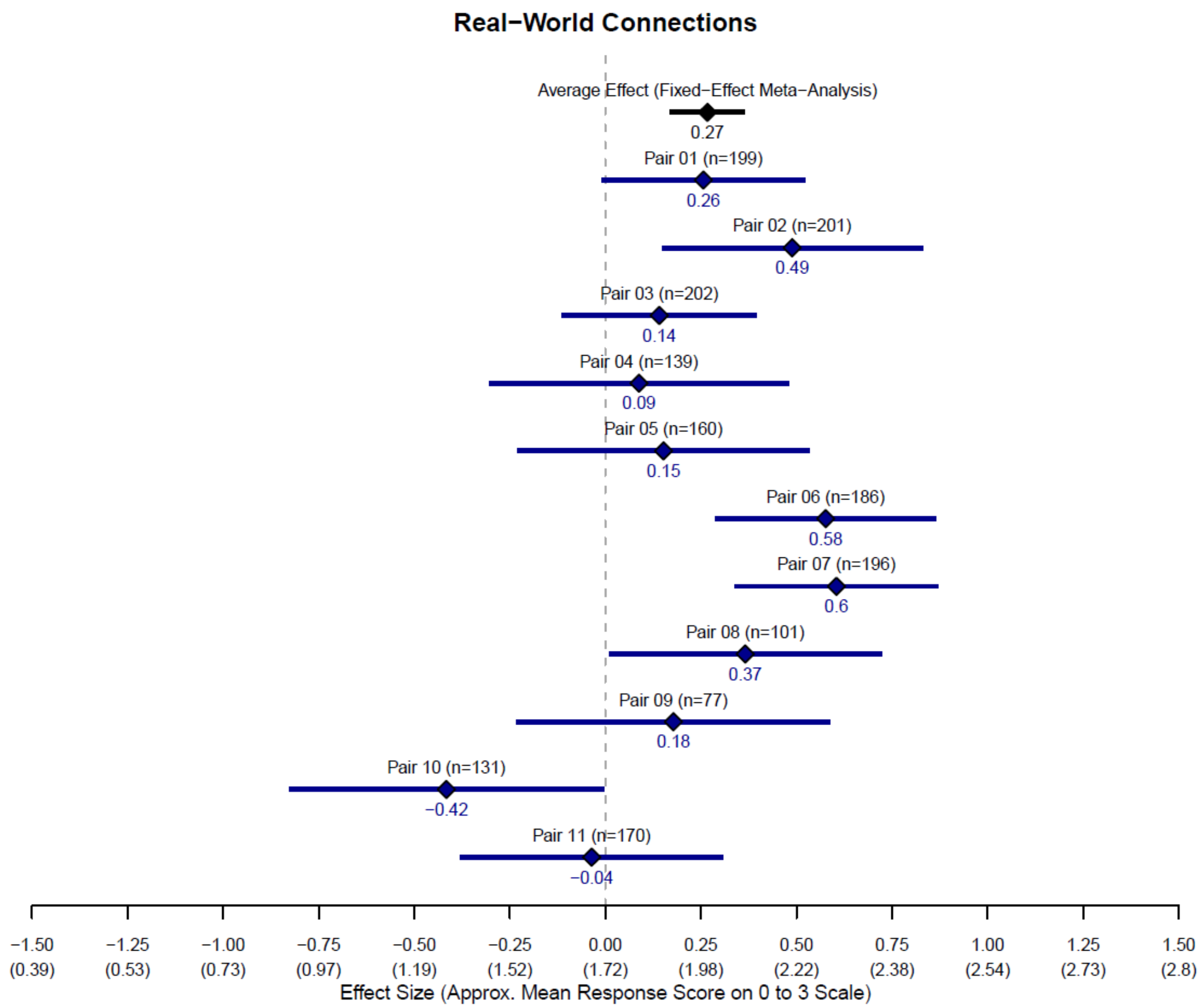
Note: Because we examined the effect of attending a network school on multiple measures of OTL, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on five OTL measures: assessments aligned with deeper learning, feedback to students, creative thinking, interdisciplinary learning, and real-world connections. The qualifying test involved examining the impact of network school attendance on a composite of the five measures. The impact on the composite measure was significant for Pairs 1, 6, 7, and 8, but not Pair 11. Thus, the result for feedback to students 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.



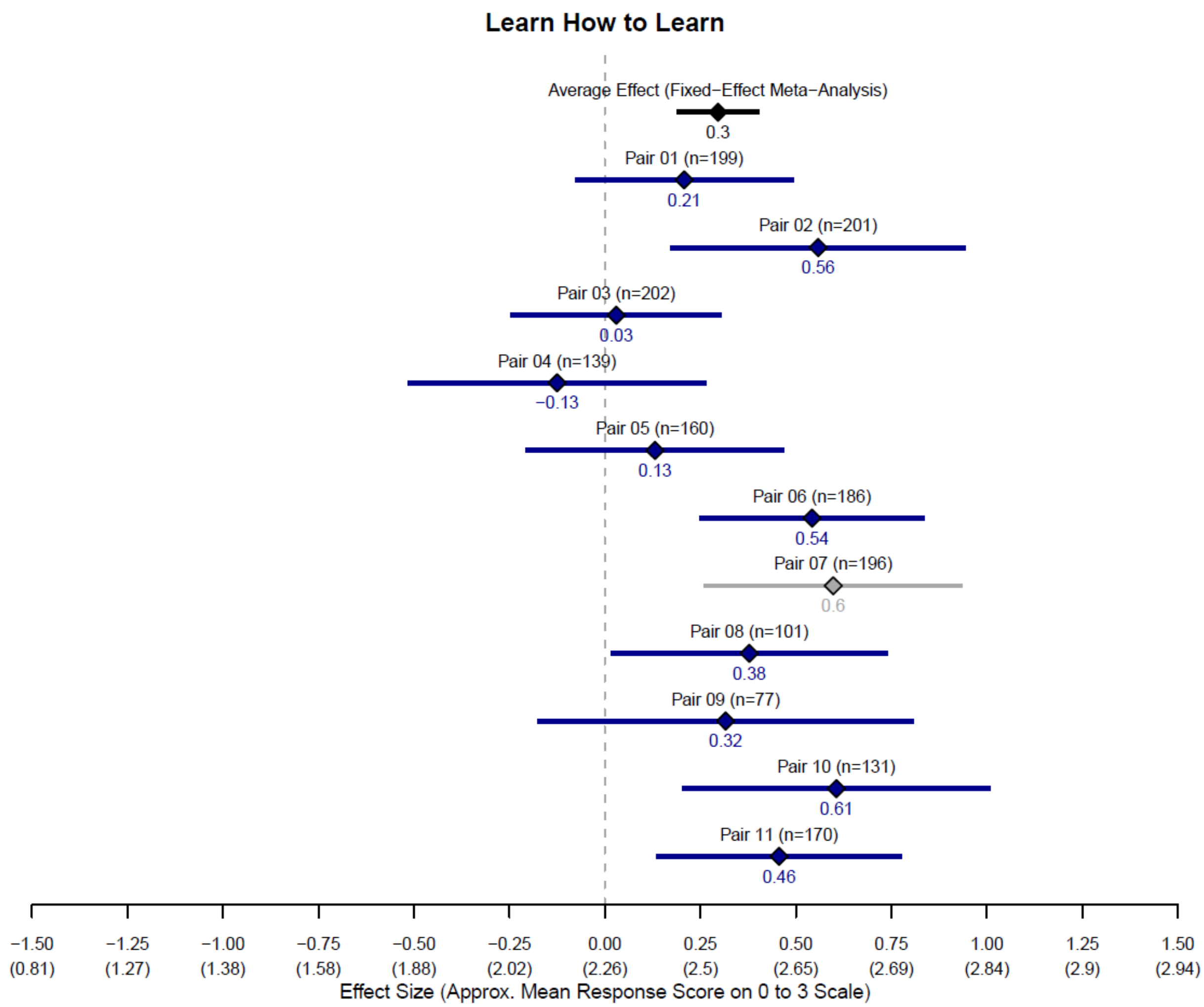
Note: Because we examined the effect of attending a network school on multiple measures of OTL, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on five OTL measures: assessments aligned with deeper learning, feedback to students, creative thinking, interdisciplinary learning, and real-world connections. The qualifying test involved examining the impact of network school attendance on a composite of the five measures. The impact on the composite measure was significant for Pairs 1, 6, 7, and 8, but not Pair 5 or Pair 11. Thus, the results for creative thinking for Pairs 5 and 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.



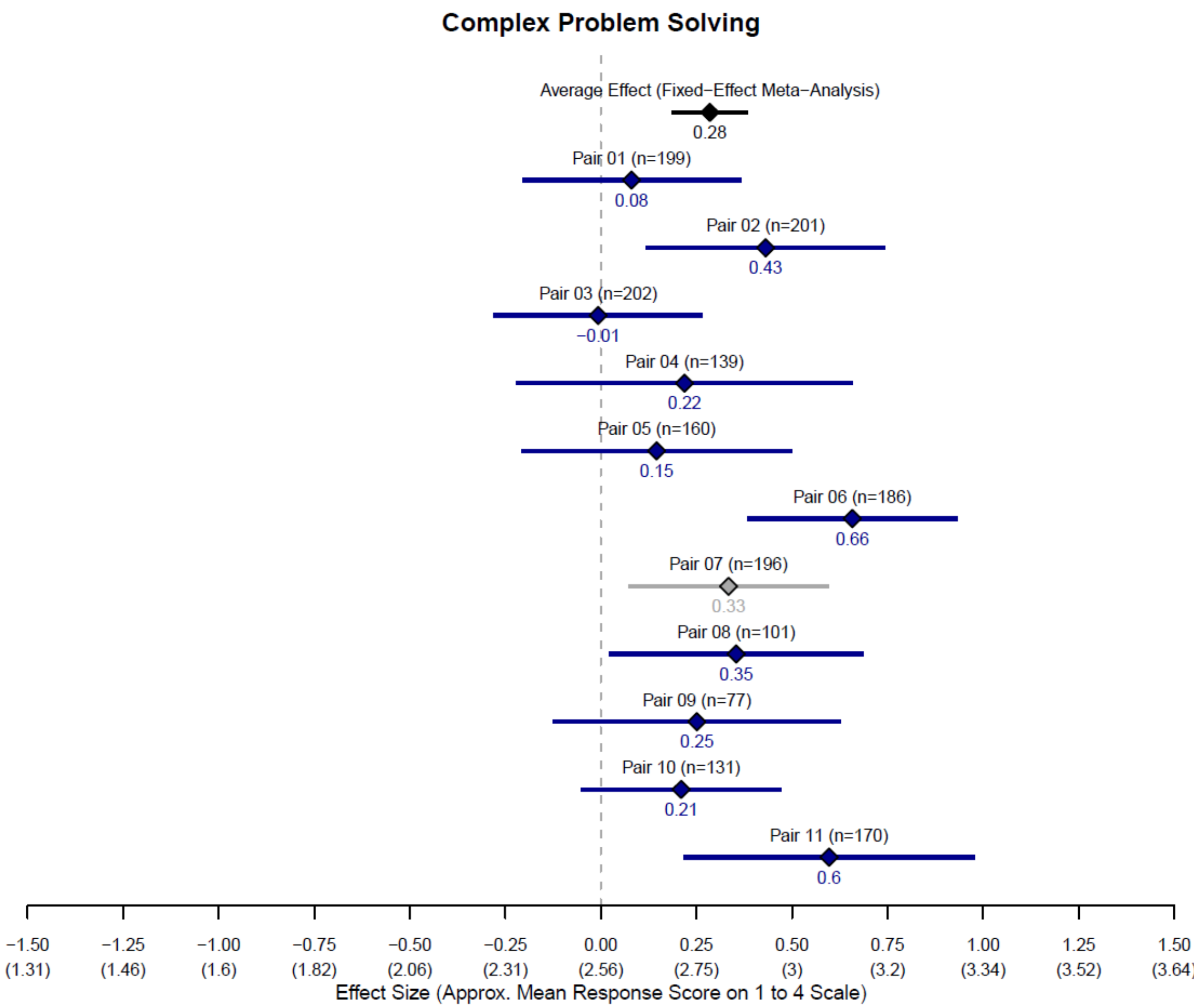
Note: Because we examined the effect of attending a network school on multiple measures of OTL, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on five OTL measures: assessments aligned with deeper learning, feedback to students, creative thinking, interdisciplinary learning, and real-world connections. The qualifying test involved examining the impact of network school attendance on a composite of the five measures. The impact on the composite measure was significant for Pairs 1, 6, 7, and 8, but not Pair 4. Thus, the result for interdisciplinary instruction for Pair 4 (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.



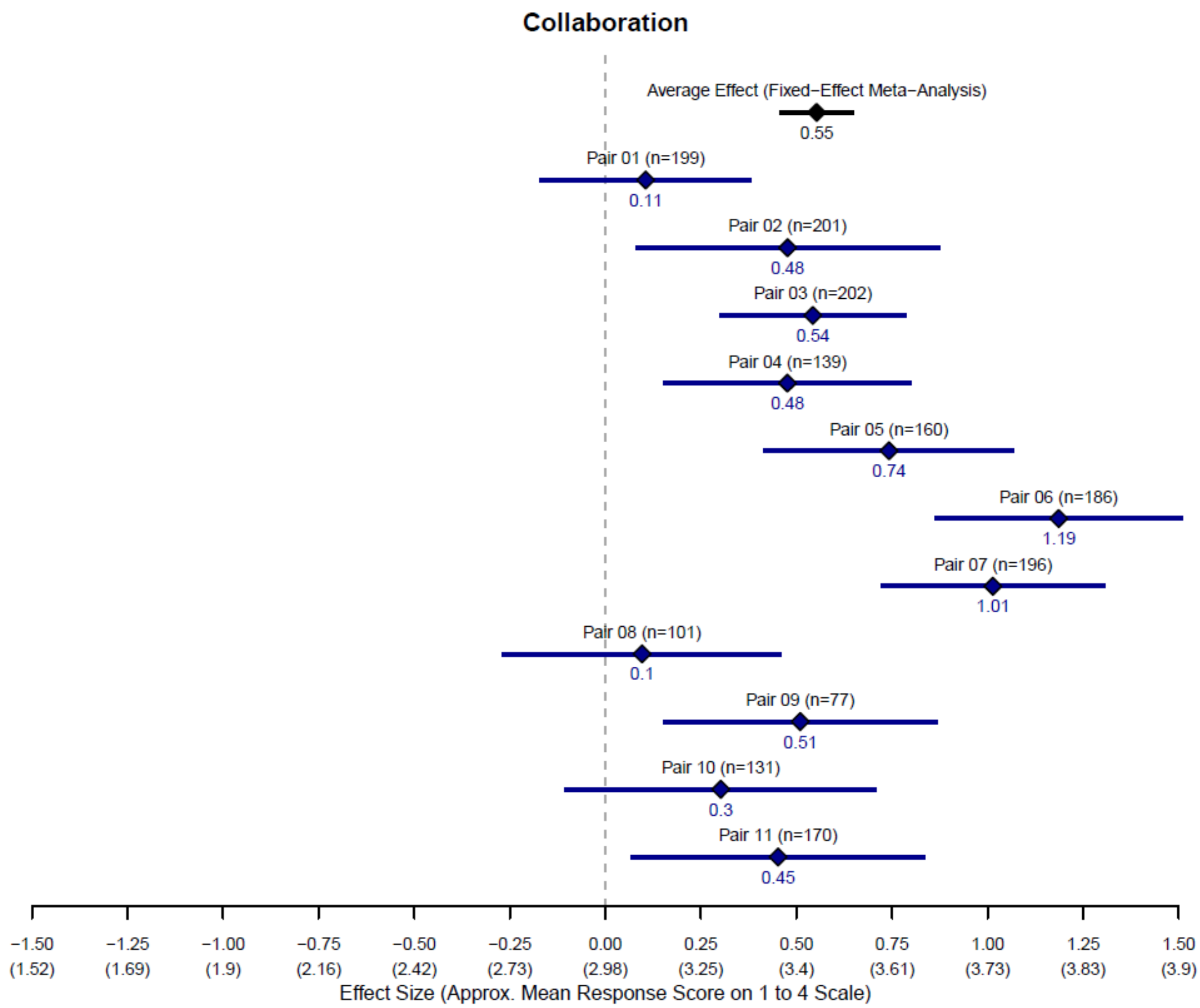
Note: Because we examined the effect of attending a network school on multiple measures of OTL, apparently significant results might have occurred simply by chance. (See Section III.A.) To check this, we performed a qualifying test prior to examining the impact of network school attendance on five OTL measures: assessments aligned with deeper learning, feedback to students, creative thinking, interdisciplinary learning, and real-world connections. The qualifying test involved examining the impact of network school attendance on a composite of the five measures. The impact on the composite measure was significant for Pairs 1, 6, 7, and 8, but not Pair 2 or Pair 10. Thus, the results for real-world connections for Pair 2 and Pair 10 (shown above) could be due to chance. 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.



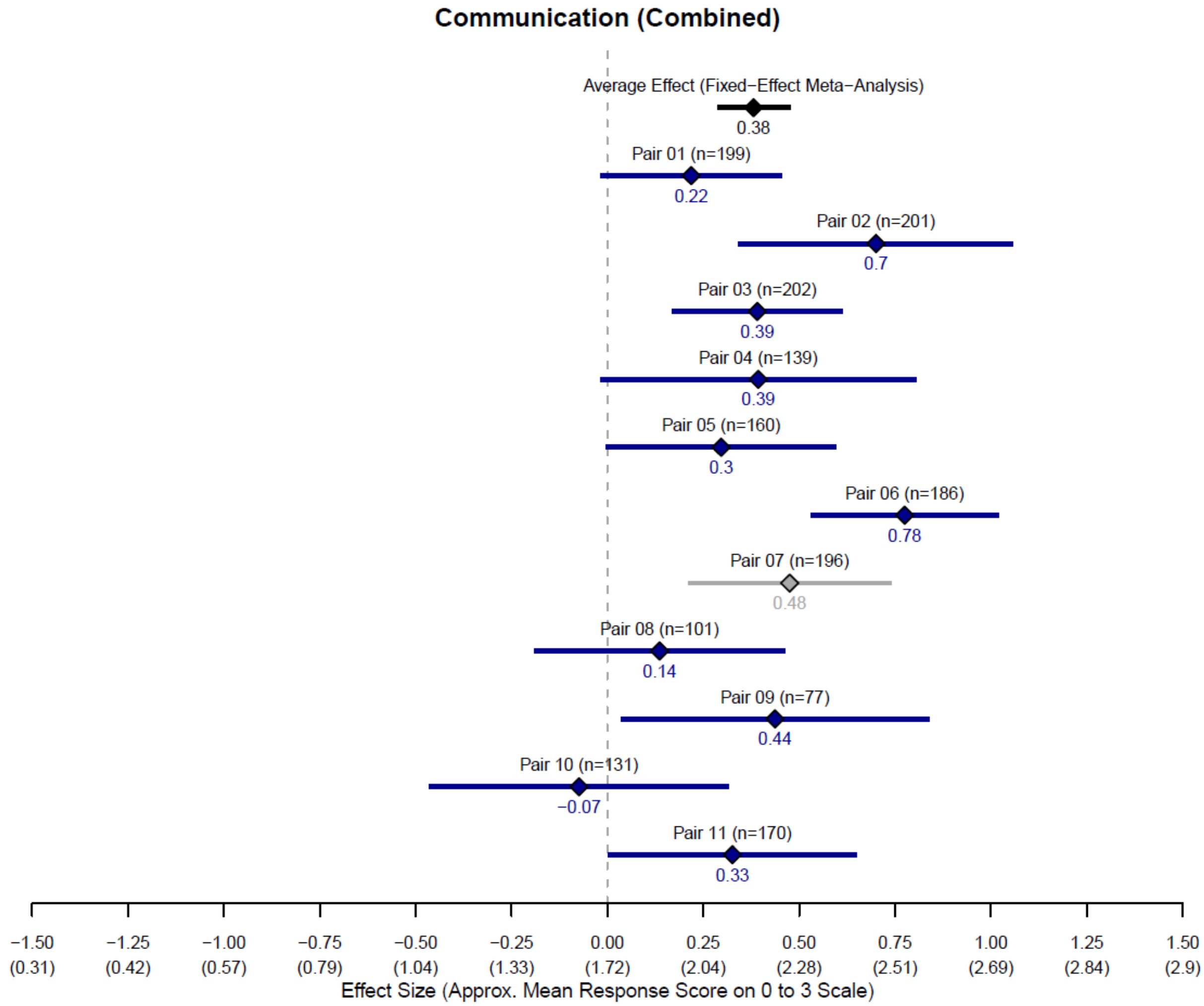
Note: A significant value of i-squared for this outcome (p = 0.045) indicates that the effect of attending a network school significantly differs across school pairs.



Note: A significant value of i-squared for this outcome (p = 0.034) 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: A significant value of i-squared for this outcome (p = 0.007) indicates that the effect of attending a network school significantly differs across school pairs.

B. OTL Effect Estimates: Subgroup Results

As described in the report, the relationships between attending a deeper learning network school and the reported exposure to opportunities for deeper learning were more positive among females than among males. In addition, for some of these OTL outcomes, Grade 12 students experienced a more positive impact from attending a network school relative to Grade 11 students. Exhibit 5.1 presents the coefficients associated with the interaction term between network school enrollment and subgroup membership. Exhibit 5.2 and Exhibit 5.3 provide more information about the size of the estimated treatment effect for each subgroup (i.e., males versus females; Grade 11 students versus Grade 12 students) only for outcomes in which the interaction term was significant.

Exhibit 5.1. Meta-Analytic Results for Subgroup Analyses for OTL Outcomes: Interaction Terms Reported as Effect Sizes

Network School	Opportunity for Complex Problem Solving	Opportunity for Communication	Opportunities for Collaboration	Opportunity to Learn How to Learn	**QT: Five OTL Outcomes[a]**	Assessments Aligned With Deeper Learning	Feedback to Students	Interdisciplinary Learning	Opportunities for Creative Thinking	Real-World Connections
Gender: Female Versus Male	0.251	0.302	0.259	0.356	0.257	0.197	0.330	ns +	0.251	0.293
Cohort: Grade 12 Versus Grade 11	ns +	ns +	ns +	0.198	0.227	ns +	0.279	ns +	0.312	0.212
FRPL: Yes Versus No	ns -	ns +	ns +	ns +	ns -	0.351	ns -	ns -	ns -	ns -
Achievement: Low Versus High (Absolute Reference)	ns -	ns +	ns -	ns -	ns +	ns -	ns -	0.220	ns -	ns -
Achievement: Low Versus High (Pair Reference)	ns -	ns +	ns +	ns -	ns +	ns +	ns +	ns +	ns +	ns +

Note: ns + denotes a non-significant positive interaction effect; ns - denotes a non-significant negative interaction effect.

[a] The qualifying test (QT) refers to the composite measure of five OTL measures that did not fit perfectly into predefined domains of deeper learning. (See Section III.A.) To reduce the probability of making a Type I error, we do not report results for the five individual OTL measures (assessments aligned with deeper learning, feedback to students, opportunities for creative thinking, interdisciplinary learning, and real-world connections) in this report unless the impact on the composite measure was statistically significant.

Exhibit 5.2. Results for Subgroup Analyses (Gender, Female Coded as 1) for OTL Outcomes: Interpretation of Significant Interaction Terms, Estimated Treatment Effects for Each Subgroup

	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)
Opportunities for Complex Problem Solving	0.186*	0.373*	0.251*
Opportunities for Communication	0.246*	0.559*	0.302*
Opportunities for Collaboration	0.497*	0.625*	0.259*
Opportunity to Learn How to Learn	0.105	0.455*	0.356*
Assessments Aligned With Deeper Learning	0.158*	0.250*	0.197*
Feedback to Students	0.220*	0.501*	0.330*
Opportunities for Creative Thinking	0.217*	0.346*	0.251*
Real-World Connections	0.120	0.367*	0.293*

* $p < 0.05$

Exhibit 5.3. Results for Subgroup Analyses (Cohort, Grade 12 Coded as 1) for OTL Outcomes: Interpretation of Significant Interaction Terms, Estimated Treatment Effects for Each Subgroup

	Treatment Effect Among Grade 11 (B_1)	Treatment Effect Among Grade 12 (B_1+B_2)	Difference in Treatment Effect (B_1-[B_1+B_2]=B_2)
Opportunities to Learn How to Learn	0.204*	0.375*	0.198*
Feedback to Students	0.272*	0.524*	0.279*
Opportunities for Creative Thinking	0.171*	0.459*	0.312*
Real-World Connections	0.150*	0.361*	0.212*

* $p < 0.05$

C. Teacher Assignment Results

Exhibit 5.4 and Exhibit 5.5 present the results of analyses of teacher assignments in mathematics and ELA. The coefficients in these tables represent the differences (in standard deviation units) in the quality of the assignments received by students in network and non-network schools. Each assignment was weighted based on the number of sampled students who received the assignment.

Exhibit 5.4. Estimates of Differences in the Quality of Mathematics Assignments Among Students Who Attended Network and Non-Network Schools

Outcome	Coefficient	SE	P-Value
Rasch Scale Score (Standardized)	0.1916	0.1461	0.195
Understanding of Core Content (Binary)	-0.0735	0.0924	0.430
Critical Thinking Skills (Binary)	0.1052	0.1248	0.403
Effective Communication Skills (Binary)	0.1828	0.1207	0.136
Independent Learning Skills (Binary)	0.2618	0.1284	0.046*
Real-World Connections (Binary)	0.0425	0.1141	0.711

Note. * p<.05

Exhibit 5.5. Estimates of Differences in the Quality of ELA Assignments Among Students Who Attended Network and Non-Network Schools

Outcome	Coefficient	SE	P-Value
Rasch Scale Score (Standardized)	0.1182	0.1228	0.340
Effective Communication Skills (Binary)	0.0059	0.0127	0.644
Critical Thinking Skills (Binary)	0.0765	0.0697	0.277
Real-World Connections (Binary)	0.1872	0.0767	0.018*
Independent Learning Skills (Binary)	-0.0008	0.0681	0.990

Note. * p<.05

D. Results of the Analysis of the Association Between OTL and Outcomes

Exhibit 5.6 and Exhibit 5.7 present the results of the analyses examining the association between opportunities for deeper learning and students' cognitive, interpersonal, and intrapersonal competency outcomes. Exhibit 5.6 presents results for schools in California, while Exhibit 5.7 presents results for New York City. Similarly, Exhibit 5.8 and Exhibit 5.9 present the results of the analyses examining the association between opportunities for deeper learning and test scores on the PBTS in California and schools in New York City.

Exhibit 5.6. Estimated Relationship Between OTL Measures and Interpersonal and Intrapersonal Competency Outcomes: California Schools

OTL Measure	Outcome	Within-School Relationship Est.	Within-School SE	Within-School P-Value	Between-School Relationship Est.	Between-School SE	Between-School P-Value
Opportunities for Complex Problem Solving	Collaboration	0.427	0.032	0.000	0.447	0.114	0.000
	Creative Thinking	0.407	0.042	0.000	0.048	0.125	0.703
	Grit	0.417	0.037	0.000	0.058	0.117	0.619
	Self-Management	0.386	0.027	0.000	-0.041	0.115	0.723
	Academic Engagement	0.402	0.034	0.000	0.409	0.158	0.010
	Motivation to Learn	0.459	0.026	0.000	0.151	0.110	0.171
	Self-Efficacy	0.418	0.027	0.000	0.129	0.145	0.373
	Locus of Control	0.370	0.028	0.000	-0.020	0.115	0.864
Opportunities for Communication	Collaboration	0.325	0.049	0.000	0.230	0.141	0.101
	Creative Thinking	0.388	0.049	0.000	-0.012	0.081	0.884
	Grit	0.371	0.053	0.000	0.002	0.086	0.983
	Self-Management	0.353	0.041	0.000	-0.023	0.097	0.817
	Academic Engagement	0.348	0.043	0.000	0.376	0.130	0.004
	Motivation to Learn	0.405	0.045	0.000	0.166	0.084	0.049
	Self-Efficacy	0.326	0.054	0.000	0.193	0.102	0.059
	Locus of Control	0.347	0.055	0.000	-0.008	0.080	0.921
Opportunities for Collaboration	Collaboration	0.293	0.037	0.000	0.203	0.083	0.015
	Creative Thinking	0.254	0.044	0.000	0.002	0.061	0.979
	Grit	0.300	0.037	0.000	-0.005	0.062	0.941
	Self-Management	0.266	0.031	0.000	-0.020	0.068	0.772
	Academic Engagement	0.337	0.029	0.000	0.275	0.079	0.001
	Motivation to Learn	0.346	0.035	0.000	0.132	0.067	0.048
	Self-Efficacy	0.290	0.048	0.000	0.129	0.066	0.051
	Locus of Control	0.313	0.037	0.000	0.015	0.070	0.825
Opportunities to Learn How to Learn	Collaboration	0.287	0.028	0.000	0.447	0.128	0.000
	Creative Thinking	0.243	0.027	0.000	0.104	0.124	0.403
	Grit	0.293	0.032	0.000	-0.014	0.134	0.918
	Self-Management	0.243	0.023	0.000	-0.125	0.136	0.360
	Academic Engagement	0.345	0.027	0.000	0.477	0.151	0.002
	Motivation to Learn	0.331	0.017	0.000	0.148	0.144	0.304
	Self-Efficacy	0.271	0.028	0.000	0.170	0.163	0.297
	Locus of Control	0.262	0.031	0.000	-0.075	0.108	0.489
Assessments Aligned With Deeper Learning	Collaboration	0.297	0.033	0.000	0.261	0.191	0.171
	Creative Thinking	0.300	0.040	0.000	0.040	0.126	0.749
	Grit	0.328	0.037	0.000	-0.064	0.122	0.599
	Self-Management	0.311	0.030	0.000	-0.052	0.141	0.711
Assessments Aligned With	Academic Engagement	0.350	0.024	0.000	0.398	0.149	0.008

OTL Measure	Outcome	Within-School Relationship			Between-School Relationship		
		Est.	**SE**	**P-Value**	**Est.**	**SE**	**P-Value**
Deeper Learning	Motivation to Learn	0.387	0.030	0.000	0.173	0.145	0.232
	Self-Efficacy	0.325	0.038	0.000	0.195	0.137	0.154
	Locus of Control	0.346	0.044	0.000	0.015	0.118	0.902
Feedback to Students	Collaboration	0.347	0.031	0.000	0.224	0.141	0.113
	Creative Thinking	0.296	0.042	0.000	0.018	0.093	0.848
	Grit	0.363	0.038	0.000	-0.029	0.100	0.772
	Self-Management	0.323	0.032	0.000	-0.046	0.104	0.661
	Academic Engagement	0.411	0.035	0.000	0.284	0.107	0.008
	Motivation to Learn	0.420	0.033	0.000	0.153	0.108	0.157
	Self-Efficacy	0.410	0.041	0.000	0.160	0.105	0.128
	Locus of Control	0.370	0.034	0.000	0.024	0.108	0.824
Interdisciplinary Learning	Collaboration	0.271	0.030	0.000	0.084	0.121	0.487
	Creative Thinking	0.266	0.034	0.000	0.069	0.054	0.197
	Grit	0.241	0.036	0.000	0.059	0.048	0.223
	Self-Management	0.243	0.028	0.000	0.087	0.051	0.088
	Academic Engagement	0.244	0.029	0.000	0.124	0.090	0.171
	Motivation to Learn	0.310	0.027	0.000	0.176	0.045	0.000
	Self-Efficacy	0.271	0.025	0.000	0.204	0.037	0.000
	Locus of Control	0.255	0.030	0.000	0.074	0.034	0.028
Opportunities for Creative Thinking	Collaboration	0.326	0.036	0.000	0.154	0.211	0.466
	Creative Thinking	0.357	0.038	0.000	0.097	0.075	0.200
	Grit	0.335	0.047	0.000	-0.002	0.103	0.984
	Self-Management	0.295	0.035	0.000	0.019	0.059	0.748
	Academic Engagement	0.332	0.040	0.000	0.328	0.167	0.049
	Motivation to Learn	0.366	0.037	0.000	0.192	0.102	0.060
	Self-Efficacy	0.360	0.052	0.000	0.226	0.069	0.001
	Locus of Control	0.344	0.051	0.000	0.065	0.066	0.330
Real-World Connections	Collaboration	0.375	0.028	0.000	0.195	0.222	0.381
	Creative Thinking	0.333	0.037	0.000	0.104	0.107	0.334
	Grit	0.371	0.036	0.000	-0.040	0.109	0.710
	Self-Management	0.363	0.033	0.000	-0.013	0.103	0.899
	Academic Engagement	0.400	0.033	0.000	0.520	0.174	0.003
	Motivation to Learn	0.450	0.033	0.000	0.267	0.111	0.016
	Self-Efficacy	0.462	0.038	0.000	0.306	0.102	0.003
	Locus of Control	0.440	0.035	0.000	0.024	0.104	0.821

Note: Based on students who attended California schools that administered the student survey (N = 14), with the student sample size ranging from 1,105 to 1,189, depending on the variables included in the model.

Exhibit 5.7. Estimated Relationship Between OTL Measures and Interpersonal and Intrapersonal Competency Outcomes: New York City Schools

OTL		Within-School Relationship			Between-School Relationship		
Measure	**Outcome**	**Est.**	**SE**	**P-Value**	**Est.**	**SE**	**P-Value**
Opportunities for Complex Problem Solving	Collaboration	0.507	0.039	0.000	0.654	0.138	0.000
	Creative Thinking	0.427	0.030	0.000	0.449	0.134	0.001
	Grit	0.476	0.028	0.000	0.529	0.143	0.000
	Self-Management	0.454	0.042	0.000	0.338	0.135	0.012
	Academic Engagement	0.472	0.060	0.000	0.631	0.119	0.000
	Motivation to Learn	0.515	0.041	0.000	0.715	0.186	0.000
	Self-Efficacy	0.452	0.055	0.000	0.633	0.135	0.000
	Locus of Control	0.428	0.063	0.000	0.467	0.105	0.000
Opportunities for Communication	Collaboration	0.251	0.035	0.000	0.803	0.198	0.000
	Creative Thinking	0.299	0.047	0.000	0.716	0.143	0.000
	Grit	0.280	0.064	0.000	0.795	0.185	0.000
	Self-Management	0.310	0.055	0.000	0.561	0.149	0.000
	Academic Engagement	0.268	0.040	0.000	0.757	0.198	0.000
	Motivation to Learn	0.295	0.066	0.000	1.011	0.191	0.000
	Self-Efficacy	0.287	0.058	0.000	0.857	0.130	0.000
	Locus of Control	0.299	0.054	0.000	0.759	0.140	0.000
Opportunities for Collaboration	Collaboration	0.356	0.046	0.000	0.514	0.102	0.000
	Creative Thinking	0.286	0.039	0.000	0.467	0.080	0.000
	Grit	0.268	0.068	0.000	0.579	0.082	0.000
	Self-Management	0.312	0.052	0.000	0.382	0.080	0.000
	Academic Engagement	0.319	0.082	0.000	0.573	0.093	0.000
	Motivation to Learn	0.310	0.072	0.000	0.773	0.051	0.000
	Self-Efficacy	0.308	0.053	0.000	0.598	0.058	0.000
	Locus of Control	0.338	0.063	0.000	0.532	0.094	0.000
Opportunities to Learn How to Learn	Collaboration	0.293	0.035	0.000	0.654	0.223	0.003
	Creative Thinking	0.184	0.037	0.000	0.465	0.174	0.008
	Grit	0.180	0.048	0.000	0.576	0.182	0.002
	Self-Management	0.248	0.055	0.000	0.350	0.195	0.072
	Academic Engagement	0.345	0.057	0.000	0.783	0.140	0.000
	Motivation to Learn	0.237	0.065	0.000	0.878	0.249	0.000
	Self-Efficacy	0.253	0.061	0.000	0.698	0.173	0.000
	Locus of Control	0.273	0.103	0.008	0.536	0.127	0.000
Assessments Aligned With Deeper Learning	Collaboration	0.244	0.080	0.002	0.507	0.310	0.102
	Creative Thinking	0.177	0.063	0.005	0.686	0.195	0.000
	Grit	0.218	0.087	0.012	0.657	0.280	0.019
	Self-Management	0.253	0.079	0.001	0.392	0.164	0.017
	Academic Engagement	0.273	0.048	0.000	0.716	0.306	0.019
	Motivation to Learn	0.244	0.087	0.005	0.918	0.360	0.011
	Self-Efficacy	0.276	0.072	0.000	0.731	0.232	0.002
	Locus of Control	0.323	0.087	0.000	0.655	0.173	0.000
Feedback to	Collaboration	0.325	0.051	0.000	0.502	0.228	0.028

OTL Measure	Outcome	Within-School Relationship			Between-School Relationship		
		Est.	SE	P-Value	Est.	SE	P-Value
Students	Creative Thinking	0.245	0.041	0.000	0.682	0.161	0.000
	Grit	0.297	0.089	0.001	0.663	0.178	0.000
	Self-Management	0.324	0.073	0.000	0.418	0.133	0.002
	Academic Engagement	0.356	0.061	0.000	0.834	0.154	0.000
	Motivation to Learn	0.341	0.085	0.000	0.975	0.181	0.000
	Self-Efficacy	0.358	0.061	0.000	0.809	0.163	0.000
	Locus of Control	0.372	0.087	0.000	0.679	0.124	0.000
Interdisciplinary Learning	Collaboration	0.118	0.039	0.003	0.412	0.123	0.001
	Creative Thinking	0.163	0.048	0.001	0.430	0.125	0.001
	Grit	0.134	0.063	0.034	0.406	0.172	0.018
	Self-Management	0.266	0.058	0.000	0.333	0.116	0.004
	Academic Engagement	0.117	0.057	0.039	0.568	0.105	0.000
	Motivation to Learn	0.221	0.047	0.000	0.607	0.186	0.001
	Self-Efficacy	0.297	0.019	0.000	0.566	0.089	0.000
	Locus of Control	0.263	0.032	0.000	0.345	0.125	0.006
Opportunities for Creative Thinking	Collaboration	0.297	0.058	0.000	0.370	0.313	0.237
	Creative Thinking	0.288	0.039	0.000	0.607	0.144	0.000
	Grit	0.325	0.048	0.000	0.580	0.211	0.006
	Self-Management	0.320	0.061	0.000	0.321	0.167	0.054
	Academic Engagement	0.308	0.055	0.000	0.861	0.177	0.000
	Motivation to Learn	0.338	0.075	0.000	0.928	0.241	0.000
	Self-Efficacy	0.378	0.044	0.000	0.726	0.160	0.000
	Locus of Control	0.393	0.056	0.000	0.581	0.116	0.000
Real-World Connections	Collaboration	0.302	0.074	0.000	0.613	0.416	0.141
	Creative Thinking	0.246	0.039	0.000	0.814	0.158	0.000
	Grit	0.290	0.064	0.000	0.819	0.282	0.004
	Self-Management	0.336	0.061	0.000	0.629	0.221	0.004
	Academic Engagement	0.353	0.087	0.000	1.037	0.211	0.000
	Motivation to Learn	0.357	0.087	0.000	1.175	0.333	0.000
	Self-Efficacy	0.358	0.068	0.000	1.010	0.151	0.000
	Locus of Control	0.364	0.077	0.000	0.768	0.197	0.000

Note: Based on students in New York City who attended schools that administered the student survey (N = 8), with the student sample size ranging from 367 to 373, depending on the variables in the model.

Exhibit 5.8. Estimated Relationship Between OTL Measures and PBTS Scores: California Schools

OTL		Within-School Relationship			Between-School Relationship		
Measure	**Outcome**	**Est.**	**SE**	**P-Value**	**Est.**	**SE**	**P-Value**
Opportunities for Complex Problem Solving (CPS)	Reading	0.025	0.033	0.460	0.671	0.284	0.018
	Math	0.053	0.020	0.010	0.220	0.209	0.293
	Science	0.079	0.024	0.001	0.500	0.220	0.023
CPS in Reading	Reading	0.017	0.028	0.543	0.412	0.181	0.023
CPS in Math	Math	0.065	0.027	0.014	-0.035	0.216	0.873
CPS in Science	Science	0.094	0.020	0.000	0.416	0.214	0.052
Opportunities for Communication	Reading	-0.003	0.021	0.878	0.129	0.188	0.494
	Math	-0.008	0.021	0.704	-0.102	0.125	0.416
	Science	0.017	0.030	0.562	0.114	0.134	0.394
Opportunities for Collaboration	Reading	-0.005	0.023	0.831	0.206	0.145	0.155
	Math	0.022	0.027	0.422	0.020	0.087	0.820
	Science	0.020	0.021	0.353	0.145	0.134	0.281
Opportunities to Learn How to Learn	Reading	0.026	0.031	0.416	0.390	0.274	0.155
	Math	-0.004	0.026	0.867	0.021	0.136	0.875
	Science	0.022	0.022	0.332	0.305	0.238	0.199
Assessments Aligned With Deeper Learning	Reading	-0.026	0.031	0.402	0.114	0.225	0.613
	Math	0.018	0.029	0.546	-0.067	0.110	0.544
	Science	0.023	0.037	0.532	0.018	0.210	0.931
Feedback to Students	Reading	0.007	0.030	0.802	0.153	0.262	0.560
	Math	0.008	0.031	0.807	-0.135	0.097	0.162
	Science	0.016	0.035	0.648	0.022	0.236	0.925
Interdisciplinary Learning	Reading	-0.020	0.025	0.429	0.248	0.210	0.237
	Math	0.013	0.026	0.632	0.021	0.112	0.853
	Science	0.019	0.034	0.570	0.147	0.196	0.454
Opportunities for Creative Thinking	Reading	-0.046	0.027	0.089	0.025	0.111	0.824
	Math	0.009	0.025	0.735	-0.063	0.099	0.522
	Science	0.003	0.027	0.919	-0.014	0.103	0.892
Real-World Connections	Reading	-0.014	0.025	0.570	0.045	0.182	0.804
	Math	-0.004	0.020	0.857	-0.021	0.097	0.830
	Science	0.009	0.034	0.786	-0.026	0.180	0.883

Note: Based on students who attended California schools that administered the student survey and the PBTS (N = 13), with the student sample size ranging from 588 to 640, depending on the variables in the model.

Exhibit 5.9. Estimated Relationship Between OTL Measures and PBTS Scores: New York City Schools

OTL		Within-School Relationship			Between-School Relationship		
Measure	**Outcome**	**Est.**	**SE**	**P-Value**	**Est.**	**SE**	**P-Value**
Opportunities for Complex Problem Solving (CPS)	Reading	0.024	0.045	0.586	0.008	0.207	0.969
	Math	0.144	0.041	0.000	0.116	0.063	0.066
	Science	0.117	0.034	0.001	-0.128	0.104	0.221
CPS in Reading	Reading	0.059	0.033	0.071	0.213	0.213	0.317
CPS in Math	Math	0.130	0.067	0.054	-0.091	0.172	0.595
CPS in Science	Science	0.081	0.023	0.000	-0.011	0.106	0.915
Opportunities for Communication	Reading	-0.019	0.053	0.717	0.271	0.326	0.406
	Math	0.044	0.051	0.390	0.182	0.069	0.008
	Science	-0.001	0.037	0.983	-0.157	0.144	0.275
Opportunities for Collaboration	Reading	-0.061	0.038	0.104	0.152	0.190	0.426
	Math	0.042	0.027	0.124	0.057	0.101	0.575
	Science	0.016	0.024	0.508	-0.176	0.088	0.045
Opportunities to Learn How to Learn	Reading	-0.044	0.043	0.309	-0.258	0.139	0.064
	Math	-0.025	0.027	0.352	0.110	0.099	0.269
	Science	-0.009	0.037	0.817	-0.139	0.204	0.496
Assessments Aligned With Deeper Learning	Reading	-0.102	0.068	0.135	-0.248	0.129	0.055
	Math	0.025	0.056	0.651	0.071	0.143	0.620
	Science	-0.008	0.053	0.877	-0.230	0.192	0.230
Feedback to Students	Reading	-0.107	0.046	0.021	-0.310	0.219	0.156
	Math	-0.027	0.071	0.705	0.098	0.200	0.624
	Science	-0.050	0.043	0.243	-0.520	0.224	0.021
Interdisciplinary Learning	Reading	-0.040	0.065	0.542	-0.343	0.167	0.040
	Math	0.039	0.046	0.391	0.015	0.192	0.936
	Science	0.001	0.045	0.979	-0.398	0.178	0.025
Opportunities for Creative Thinking	Reading	-0.097	0.052	0.063	-0.073	0.124	0.555
	Math	-0.010	0.043	0.809	0.088	0.061	0.149
	Science	-0.023	0.033	0.492	0.000	0.154	0.998
Real-World Connections	Reading	-0.035	0.070	0.616	-0.307	0.093	0.001
	Math	0.034	0.051	0.500	-0.020	0.149	0.893
	Science	0.011	0.056	0.843	-0.254	0.179	0.155

Note: Based on students who attended New York City schools that administered the student survey and the PBTS (N = 7), with the student sample size ranging from 181 to 211, depending on the variables in the model.

References

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.

Newmann, F. M., Secada, W. G., & Wehlage, G. (1995). *A guide to authentic instruction and assessment: Vision, standards and scoring*. Madison, WI: Wisconsin Center for Education Research.

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*.