WORKING
PAPER

Is Reform-Oriented Teaching Related to Mathematics and Science Achievement?

VI-NHUAN LE, J.R. LOCKWOOD, BRIAN STECHER,
LAURA HAMILTON, VALERIE WILLIAMS,
ABBY ROBYN, GERY RYAN, ALICIA ALONZO

WR-166-EDU

April 2004

Prepared for the AERA conference

EDUCATION

DRAFT

Paper presented at the annual meeting of the American Education Research Association, San Diego, California, April 2004

This material is based on work supported by the National Science Foundation under Grant No. ESI-9986612. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Recent large-scale efforts to improve mathematics and science education have focused on changing teachers' instructional practices to be more aligned with the teaching standards put forth by professional organizations such as the National Council of Teachers of Mathematics or the American Association for the Advancement of Science. These national organizations, as well as other local reform efforts, advocate an approach to mathematics and science instruction that places less emphasis on the acquisition of discrete skills and factual knowledge, and greater weight on conceptual understanding, inquiry, and application and communication of mathematical or scientific ideas (National Research Council, 1996; American Association for the Advancement of Science, 1993; National Council of Teachers of Mathematics, 1989; 2000). This approach, commonly referred to as reform-oriented instruction, is intended to engage students as active participants in their own learning and to promote the development of complex cognitive skills and processes. Although advocates of this approach do not dispute the importance of computational skills and factual knowledge, they argue that traditional curricula have often emphasized these outcomes to the exclusion of more complex problem-solving and reasoning skills, and as a result, students are often poorly prepared for careers that require the use of higher-level mathematics and science skills and knowledge.

Research shows that many teachers have begun to incorporate these approaches in their classrooms (Kim, Crasco, Blank, & Smithson, 2001), but the evidence supporting the use of these practices in mathematics and science is relatively weak. Studies that have examined the relationship between student achievement and teachers' reports of the frequency with which they engaged in reform-based instruction suggest that these practices may contribute to student achievement, but in most cases, the effects appear to be quite small. Mayer (1998) observed small positive or null relationships between reform-based practices and student scores on a standardized multiple-choice test. Similar results are described by Wenglinsky in mathematics (2002) and Smerdon, Burkam, and Lee (1999) in science. A synthesis of data from eleven NSF-funded Systemic Reform Initiatives found a mixture of null and small positive results on both multiple-choice and open-response assessments (Hamilton et al., 2003).

Of particular interest is whether reform-oriented teaching can enhance performance of scientific or mathematical communication, problem solving, or other "higher-order thinking" skills. Advocates of reform teaching believe these are areas in which reform practices might be especially effective. There is some evidence that reform instruction is positively related to these types of skills, albeit weakly. A study by Cohen and Hill (2000) revealed an association between reform instruction and scores on the California Learning Assessment System (CLAS) mathematics test, a performance-based assessment designed to measure students' understanding of mathematics problems and procedures. Thompson and Senk (2001) found that reform practices, in conjunction with a reform-oriented curriculum, correlated positively with mathematics achievement, especially on multistep problems, and problems involving applications or graphical representations in mathematics. Similarly, Saxe, Gearhart, and Seltzer (1999) found that reform instruction was associated with student mathematics problem-solving scores.

The small relationships between reform pedagogy and achievement may be partially attributable to inadequate measurement of reform-oriented instruction. Many of these studies used surveys that asked teachers about the frequency with which they engaged in particular practices. While researchers have successfully used these data to explore relationships between instructional practices and student achievement (Cohen & Hill, 2000; Gamoran, Porter, Smithson, & White, 1997; Wenglinsky, 2002), surveys nonetheless have problems that limit their utility (Rowan, Correnti, & Miller, 2002). Surveys, for example, are designed to be applicable to a range of settings and focus on long-term patterns of behavior. Thus, they often ignore important variations in teaching strategies related to grade level or content (Mullens & Gayler, 1999). Surveys also cannot capture subtleties in how teachers understand terminology or implement practices. Mayer (1999) showed that two teachers could report the same frequency of open-ended discussion, but the nature of the discussion differed in important ways between the classrooms. He concluded that survey data can distinguish between teachers who frequently use certain types of practices from teachers who do not, but it cannot provide information about how those practices are substantively implemented.

The purpose of this study is to explore the relationship between mathematics and science achievement and reform teaching. We use data gathered from various teacher sources (to be described later), and from district records of student test scores to explore this relationship. A unique feature of our analysis entails innovative methods for measuring instructional practice in the hopes of capturing aspects of reform teaching that may not be well measured by surveys. In addition, our analysis explores the relationship between reform instruction and "higher-order thinking" skills.

Methodology

Sample Participants

Three school districts that were implementing NSF-sponsored Local Systemic Change (LSC) projects participated in this study. The LSC program is one of a series of initiatives designed to promote reform-oriented, systemic reform of mathematics and science teaching. A large portion of the initiatives' funds is devoted to training to increase teachers' use of classroom practices that are consistent with the reform principles recognized by some national professional organizations (e.g., the National Council for Teachers of Mathematics, the American Association for the Advancement of Science).

Students from these school districts have been grouped into five cohorts based on school district, grade level and subject content, and their student achievement data is being tracked for three years. The districts and subject-grade combinations were chosen to meet several criteria, including an established record of teacher participation in reform-related professional development, a willingness to administer student achievement tests as well as the instructional practice measures, and the availability of a data system that would allow us to track individual students over time and link students to their math or science teachers. The sample includes two cohorts for middle school mathematics, and one each for elementary mathematics, elementary science, and middle school science. This resulted in an initial year sample consisting of third-grade mathematics, third-grade

science, sixth-grade mathematics, sixth-grade science, and seventh-grade mathematics. We followed these students for an additional two years, resulting in a longitudinal sample of third through fifth grade, and sixth through eighth grade for mathematics and science, and seventh through ninth grade for mathematics. This paper presents results only for the initial year.

Table 1 indicates the total number of responding teachers, and the number of students in the district-provided dataset who were linked to those teachers. Response rates varied from 56 to 84 percent for teachers, and from 77 to 87 percent for students.

Table 1. Description of Participating Cohorts in Year 1

Student Cohorts	School District	Subject	Grade	Teacher Response (n/N)	Student Response (n/N)
Cohort 1	District 1	Mathematics	3	68/81	1264/1642
Cohort 2	District 1	Mathematics	7	43/59	3054/3520
Cohort 3	District 2	Science	3	61/109	805/960
Cohort 4	District 2	Mathematics	6	55/79	1750/1851
Cohort 5	District 3	Science	6	63/83	5059/5808

For each site, we received student demographic information, including racial/ethnic group, gender, and eligibility for free-or reduced-price lunches. In many of the sites, we received additional information, such as age at testing, gifted status, and limited English proficient (LEP) status. Student demographics for each of the sites are provided in Appendix A.

Measures

For each subject-grade combination, different methods were used to measure classroom practice: a survey, a set of vignette-based questions, and daily classroom logs. The measures were designed to provide overlapping evidence about the extent of reform-oriented teaching that occurred in each class. For each measure, teachers were provided with a set of responses designed to capture a range of behaviors, from less reform-oriented to more reform-oriented. The measures are described in detail in the following sections.

Teacher Survey. The survey included questions about the teacher's educational background and experience, the mathematics curriculum taught in the class, and the use of a variety of teaching practices. Teachers indicated how much class time was spent on various mathematical topics (e.g., multiplication/division of whole numbers, patterns/functions/algebra). They indicated the frequency with which they engaged in particular instructional activities (e.g., lecture or introduce content through formal presentations, encourage students to explore alternative methods for solutions). They also indicated the frequency with which students took part in specific learning activities (e.g., practice computational skills, work on extended mathematical investigations). Questions of these types have been used extensively in research on mathematics instruction (see for example, Cohen & Hill, 2000; Hamilton et al., 2003; Wenglinsky, 2002).

Teacher Logs. Teachers were asked to fill out a daily log describing specific activities that occurred during their mathematics lesson. While the surveys focused on long-term patterns of behavior, the logs focused on activities during a specific two-day period. Teachers indicated how much time students spent on selected activities (e.g., use manipulatives to solve problems, complete worksheets or problem sets from text). Similarly, they indicated how much time they devoted to selected behaviors (e.g., monitor students as they work, ask questions of individuals to test for understanding). Finally, teachers indicated whether or not certain activities occurred at all during the lesson (e.g., students engaged in a debate or discussion of ways to solve a problem, teacher or student connected today's math topic to another subject (e.g., social studies)).

Vignette-Based Items. As part of the survey, teachers responded to two vignettes that contained descriptions of realistic classroom settings and events and asked teachers to indicate how they would respond in each setting. Each vignette presented the teachers with four instructional problems that provided teachers with hypothetical situations at different stages within the unit. The first problem focused on the manner in which the teacher would introduce a new unit. The second problem dealt with how the teacher responds to mistakes from students. The third problem involved teachers' reactions to students who gave two approaches to solving a problem, both of which were correct, but differed in their efficiency. The final problem asked teachers about their emphasis on different learning objectives (see Appendix B for the vignettes from sixth-/seventh-grade mathematics).

For each problem, teachers were provided with a set of responses that were designed to capture a range of behaviors, from less reform-oriented to more reform-oriented. Teachers were asked to rate the likelihood of engaging in each option using a four-point scale from "very unlikely" to "very likely" or, in the case of questions of emphasis, from "no emphasis" to "great emphasis." Teachers' responses provided an indication of their intent to teach in a less-reform or more-reform manner.

The topics for the vignettes were selected to be appropriate to the grade and curriculum of the site. In creating the vignettes, we followed a template so that the hypothetical situations were roughly comparable within and across grades for a given subject. In other words, the vignettes were designed to be "parallel" in the sense that we presented teachers with similar instructional problems and response options, but situated in different topics. However, the different setting for each vignette meant that some response options had to be modified to fit the specific mathematical or scientific context. Thus, while the vignettes are highly similar, they cannot be considered strictly parallel.

Achievement Measures. In addition to information gathered from teachers, we obtained achievement scores for students. The specific tests varied by site, but typically included multiple measures (see Table 2). Additionally, for all but one site, we received prior year test scores in various subjects.

Table 2. Achievement Measures for Each Site

Student Cohorts	Subject	Grade	Current Achievement Tests	Prior Achievement Tests
Cohort 1	Mathematics	3	SAT-9 MC Math District CRT Math	District CRT Math and Reading
Cohort 2	Mathematics	7	SAT-9 MC Math District CRT Math	District CRT Math and Reading
Cohort 3	Science	3	SAT-9 MC Science	----
Cohort 4	Mathematics	6	SAT-9 OE Math SAT-9 MC Math SAT-9 MC Math, problem solving NALT Math	MCA Reading, Math, and Writing
Cohort 5	Science	6	SAT-9 MC Science	District CRT Math and Reading MSPAP

Notes.
MC refers to multiple-choice.
OE refers to open-ended.
SAT-9 refers to the Stanford Achievement Test, 9th Version
NALT refers to the Northwest Achievement Level Test
MCA refers to the Minnesota Comprehensive Assessments
CRT refers to the local district test named the Criterion Reference Test
MSPAP refers to the Maryland School Performance Assessment Program

Analysis

Scale Development

Scales Derived from the Vignettes. We scored the vignettes separately for each grade and subject. Our goal in scoring the vignette-based questions was to characterize teachers along a single dimension of reform from highly reform-oriented to not very reform-oriented. This necessitated deciding which of the options described behaviors that reflect reform teaching. We used a judgmental process to assign a value from 1 (low reform) to 4 (high reform) to each response option. Members of the research team independently rated each response option, then convened to reconcile differences. The panel of expert mathematicians rated a comparable set of scenarios in a previous year, and we used the decision guidelines they established. For the purposes of analysis, we considered options that had been rated a 3 or 4 to be indicative of high-reform pedagogy, and options that had been rated a 1 or 2 to be indicative of low-reform teaching (See Appendix C for reform ratings for sixth/seventh-grade mathematics).

We used a multidimensional scaling process in which both the high- and low-reform items were included as part of the scoring procedures. From the *n*-teacher by *N*-item matrix, we created an *n* x *n* teacher similarity matrix, and used multidimensional scaling to plot the teachers in three-dimensional space. We added into this plot two additional points corresponding to a simulated "ideal" high-reform teacher (whose responses corresponded exactly to our judgments of reform orientation), and a simulated "ideal" low-reform teacher (whose responses were just the opposite). An examination of the teachers' responses showed that teachers were generally more similar to our idealized

high-reform teacher than they were to our idealized low-reform teacher, although there was also variation in responses. We then used the *n*-teacher by *N* response option matrix to construct and plot a *N* by *N* response option similarity matrix in 3-dimensional space. The results suggested that teachers were consistent in how they were responding to the items. That is, high-reform teachers indicated that they were likely to engage in many of the high-reform options and few of the low-reform options, whereas low-reform teachers showed the opposite pattern.

On the basis of these similarity analyses, we created a scale of reform-oriented instruction by calculating the Euclidean distance between each teacher and the ideal high-reform teacher. We scaled this measure, which we refer to as Euclid, so that teachers who are closer to the ideal high-reform teacher receive lower scores. That is, smaller values of Euclid are associated with teachers whose responses are more like our ideal high-reform teacher. (Readers interested in the relationships between teachers' responses to the vignettes and to the surveys and logs are referred to Le et al., 2003).

Scales Derived from Survey and Daily Logs. Several additional scales were derived from teachers' responses to the survey and the daily log. The scales were created using a combination of empirical analyses (e.g., factor analysis), prior results with similar items in other studies, and expert judgments based on the underlying response options. We created scales that we thought were key indicators of low- and high-reform instruction, scales that described teacher background, and scales that described classroom context. Each is described below.

Table 3 contains brief descriptions of the scales and examples of illustrative items. (The complete set of items that comprise each scale is presented in Appendix D.) Of the instructional practices scale, two were derived from the surveys. Strategies measured the number of strategies used at least moderately when teaching about multiplication (third grade) or decimals and fractions (sixth- and seventh-grade). Reform Practices, measured the frequency with which teachers engaged in nine specific reform-oriented instructional practices. The items on this scale are similar to those used in other research on mathematics and science reform, including some national longitudinal surveys (Cohen & Hill, 2000; Hamilton et al., 2003; Swanson & Stevenson, 2002; Wenglinsky, 2002).

The remaining instructional practices scales were derived from the daily logs. The Discussion scale was based on the amount of time teachers reported that the class was engaged in a dialogue about mathematical/scientific thinking and understanding. Groupwork entailed the amount of time the class spent in groups as a whole, whereas Mixed-ability Groupwork measured the amount of that time in which students worked in mixed-ability groups. Problem-solving Groupwork assessed the amount of time in which students collaboratively solved new problems. The Reform Activities scale measured the number of reform behaviors that occurred during the lesson, and the Seatwork scale described the amount of time students spent reading from the textbook, completing worksheets or problem sets, or other low-reform activities. Additionally, Hands-on described how much class time was spent on hands-on science activities, and Number of Problems Solved described the number of problems the typical student solved per minute.

There were three scales relating to curriculum coverage. Operations measured the extent to which teaches covered operations with whole numbers, which is thought to be a more traditional area of math, and Proof and Patterns assessed the extent to which they focused on proof/justification/verification and patterns/function/algebra. These topics represent more reform-oriented areas of math. In science, the Science Content scale measured the number of lessons teachers spent on selected science topics.

There were also six scales, all derived from the surveys, about teacher background. The Certification scale indicated whether the teacher held a standard certification, and the Confidence scale assessed whether teachers felt very confident in their mathematics or science knowledge that they were asked to teach. Masters assessed whether teachers held a masters degree in any subject, and Math Degree indicated whether the teacher held a major or minor in mathematics. Professional Development measured the amount of subject-specific in-service training received in the past 12 months. The Experience scale indicated the total number of years teachers taught on a full-time basis, and Experience at Grade scale indicated the total number of years teachers taught at grade level.

The final set of scales was designed to provide context about classroom conditions. The Percent Learned Concepts represented teachers' estimation of the percent of students who understood the lesson, and the Time on Task scale measured how much of the class time was spent effectively on the mathematics or science lesson.

Table 3. Summary of Scales

Scale	Description	Illustrative Items	Subject	Source	Variable Name
Instructional Practices					
Reform Inclinations	Standardized sum of teachers' answers to the high-reform response options across the two scenarios	Have a classroom discussion about the differences between the two approaches	Mathematics Science	Vignette	Allhigh
Euclid	Euclidean distance of the teacher from the ideal high reform teacher	Tell them they are both right and move on to the next problem	Mathematics Science	Vignette	Euclid
Reform Practices	Frequency with which the teacher engaged in reformed instructional practices	How often does the teacher encourage students to explore alternative methods for solutions?	Mathematics Science	Survey	Reform
Strategies	Number of strategies used at least moderately when teaching specified topic	Overall, during these lessons, what portion of the time was devoted to using manipulatives to show equivalence between fractions and decimals (sixth-/ seventh-grade)?	Mathematics	Survey	Strategies
Discussion	Amount of time the class engaged in dialogue about mathematical/ scientific thinking and understanding	How much time did students spend explaining their thinking about mathematical/scientific problems?	Mathematics Science	Log	Discuss
Groupwork	Amount of time the class spent in groupwork	How long did students work in groups during today's mathematics/science lesson?	Mathematics Science	Log	Groupwrk
Mixed-ability Groupwork	Amount of time the class spent working in mixed-ability groups	If groups were used, what share of the group time was used working in groups of mixed ability?	Mathematics Science	Log	Absmxd
Problem-solving Groupwork	Amount of time the class spent solving new problems together as a group	If groups were used, what share of the group time was used solving new problems together as a group?	Mathematics Science	Log	Absprob

Reform Activities	Presence of selected reform activities	In today's mathematics science lesson, did a student restate another student's ideas in different words?	Mathematics Science	Log	Refact
Seatwork	Extent to which students engaged in seatwork and other low-reform activities	How much time did students spend reading from textbook or other materials?	Mathematics Science	Log	Traditional
Hands-on	Amount of time the class spent on hands-on activities	How much time did students spend doing hands-on science investigations?	Science	Log	Handson
Number of Problems Solved	Number of problems solved per minute	How many math problems did the typical student work on in class today?	Mathematics	Log	Numprob
Curriculum					
Operations	Weeks spent on operations with whole numbers	Indicate the approximate amount of time you will spend on operations with signed whole numbers	Mathematics	Survey	Operations
Proofs and Patterns	Weeks spent on proof/justification/verification and patterns/functions/algebra	Indicate the approximate amount of time you will spend on patterns/ functions?	Mathematics	Survey	Proof.patterns
Science Content	Average number of lessons spent on selected science topics	How much time will you spend teaching patterns and relationships this school year?	Science	Survey	Scicontent
Teacher Background					
Certification	Whether the teacher holds standard certification	Which type of teaching certification do you hold?	Mathematics Science	Survey	--
Confidence	Whether the teacher is very confident in his/her mathematics/science knowledge	With respect to the mathematics/ science that you are asked to teach, how confident are you in your mathematical/scientific knowledge?	Mathematics Science	Survey	Confidence
Masters	Whether the teacher holds at least a masters degree (any subject)	What is the highest degree you hold?	Mathematics Science	Survey	Mastdeg
Math Degree	Whether the teacher holds a math-intensive undergraduate degree (major or minor)	Did you major in mathematics or a mathematics-intensive field for your Bachelor's degree?	Mathematics	Survey	Mathdeg

Professional Development	Amount of professional development received	In the past 12 months, how much time have you spent on professional development activities that focused on in-depth study of mathematics/science content?	Mathematics Science	Survey	Pdtot
Experience	Number of years of experience	Including this year, how many years have you taught on a full-time basis?	Mathematics Science	Survey	Yrstch
Experience at Grade	Number of years teaching at grade level	Including this year, how many years have you taught third grade? (third grade version)	Mathematics	Survey	Yrs.grade
Classroom Context					
Percent Learned Concepts	Teacher-estimated percent of students who learned the concepts	About what percent of the students learned the concepts or skills you expected them to learn today?	Mathematics Science	Log	Concepts
Time on Task	Number of minutes effectively spent on mathematics or science (i.e., disregarding disruptions)	How long was today's mathematics/science lesson?	Mathematics Science	Log	Efftime

Most of the scales were dichotomous or continuous, but Reform Practices, Discussion, Professional Development, and the scales relating to curriculum (i.e., Operations, Proof and Patterns, and Science Content) were on a 5-point metric, and Concepts was on a 6-point metric. For these scales, the score was simply the average response across items, with higher scores denoting more frequent use of the practices measured by that scale. For example, a score of 5 on the Reform Practices scale indicates that teachers spent much time on reform-oriented activities. In contrast, a score of 1 on a scale such as Discussion indicates that little class time was devoted to talking about mathematical thinking.

Modeling Approach

To examine the relationship between reform teaching and student achievement, we started by computing student-level regression equations separately for each of the five sites. Our equations controlled for differences in both teacher and student characteristics. For teachers, we controlled for professional development, degree, total years teaching, confidence in mathematics/science knowledge, and whether the teacher had a masters degree. For students, we controlled for as many characteristics as possible, including (when available) prior year achievement scores, race/ethnicity, gender, limited English proficiency status, and whether the student participated in a gifted program, a special education program, and/or a free or reduced price lunch program. In addition, we included variables representing different aspects of instruction, many of which were reform-oriented. The dependent variable was standardized multiple-choice achievement scores from the Stanford-9.

We fit a hierarchical linear model, in which we included fixed-effects for the teacher-level variables as well as random effects for teachers. This was sufficient to capture any remaining classroom-level heterogeneity, and resulted in the proper standard errors for fixed effects.

After conducting five separate regression equations (corresponding to each of our sites), we synthesized the results across sites through a meta-analysis. For each subject area, we estimated the composite standardized coefficients as a weighted average of the individual standardized coefficients, where the weight for each was proportional to its estimated precision (as indicated by the standard errors from the model fits).[1] Significance testing was conducted individually for each variable relating to instructional practices. Specifically, we pooled the estimated standardized coefficients across sites from the models that controlled for both teacher and student variables, then added the particular variable of interest. (Readers who are interested in more details of the modeling approaches are referred to Appendix E).

Results

[1] For the standardized coefficients, we used a conservative standardization method based on the marginal standard deviation of the scores across all students in the sample. Other less conservative methods would have affected the absolute size of our "effect sizes," but not their statistical significance.

Distribution of the Scales. Table 4 provides the percent of teachers answering "yes" to a dichotomous item. Tables 5 and 6 present the descriptive statistics for selected scales in mathematics and science, respectively.

Results showed that the scales were reasonably consistent, with alpha values ranging from .30 to .91 in mathematics and .42to .92 in science. Most of the scales showed moderate variation, with the exception of Certification (which was subsequently dropped from further analysis). There were some differences across sites in the score ranges and variability. Although these differences could influence the likelihood of detecting relationships with achievement, the results we discuss in later sections show no clear patterns with respect to these differences.

Table 4. Frequency Distribution for Selected Scales

Scale	Mathematics			Science	
	Cohort 1	Cohort 2	Cohort 4	Cohort 3	Cohort 5
Certification	89.55	90.70	89.09	95.08	92.06
Masters	52.94	48.84	66.13	48.53	74.58
Confidence	50.00	83.72	59.68	18.33	55.74

Table 5. Descriptive Statistics for Selected Mathematics Scales

Scale	Cohort 1			Cohort 2			Cohort 4		
	Mean	SD	Alpha	Mean	SD	Alpha	Mean	SD	Alpha
Instructional Practices									
Reform Inclinations	0.00	1.00	.75	0.00	1.00	.78	0.00	1.00	.81
Euclid	1.76	.59	.85	1.66	.57	.84	1.27	.55	.91
Reform Practices	3.81	.47	.72	3.54	.39	.57	4.14	.54	.82
Strategies	5.69	1.80	.72	4.31	1.99	.68	5.15	2.41	.78
Discussion	2.45	.59	.77	1.96	.45	.79	2.50	13.23	.76
Groupwork	11.27	9.26	--	4.18	5.48	--	22.59	13.23	--
Mixed-ability Groupwork	6.38	7.92	--	1.96	5.07	--	14.97	14.38	--
Problem-solving Groupwork	5.51	7.76	--	1.41	4.11	--	13.14	11.42	--
Reform Activities	3.39	1.33	.54	2.65	1.47	.38	4.05	1.44	.63
Seatwork	2.58	.77	.58	2.42	.78	.67	2.74	.73	.55
Strategies	5.69	1.80	.72	4.31	1.99	.68	5.15	2.41	.78
Number of Problems Solved	.38	.22	--	1.08	1.40	--	.25	.24	--
Curriculum									
Operations	3.78	.82	.64	3.27	.87	--	3.76	.90	--
Proofs and Patterns	2.61	.95	.72	2.79	.64	.30	2.91	1.01	.35
Teacher Background									
Professional Development	2.53	.87	.88	2.83	.91	.85	2.85	1.07	.89
Experience	13.28	9.61	--	14.70	10.18	--	10.51	8.52	--
Experience at Grade	5.54	4.51	--	9.65	8.25	--	5.98	5.38	--
Math Degree[a]	--	--	--	.63	.79	--	.35	.75	--

Classroom Context									
Percent Learned Concepts	5.27	.66	--	5.47	.39	--	4.96	.79	--
Time on Task	43.03	11.91	--	26.47	15.20	--	52.17	17.05	--

Note.
[a] Math degree takes on a value of 2 if the teacher holds a major, 1 if the teacher holds a minor, and 0 otherwise.

Table 5. Descriptive Statistics for Selected Science Scales

Scale	Cohort 3			Cohort 5		
	Mean	SD	Alpha	Mean	SD	Alpha
Instructional Practices						
Reform Inclinations	0.00	1.00	.85	0.00	1.00	.84
Euclid	1.02	.56	.91	.91	.50	.89
Reform Practices	3.52	.51	.78	3.86	.43	.76
Discussion	2.41	.51	.68	2.56	.65	.84
Groupwork	27.27	22.46	--	22.19	14.52	--
Mixed-ability Groupwork	21.23	14.54	--	14.22	15.99	--
Problem-solving Groupwork	18.07	14.35	--	12.73	14.91	--
Reform Activities	3.95	1.32	.42	4.18	1.53	.59
Seatwork	1.63	.54	.57	2.06	.67	.64
Hands-on	28.92	21.25	--	18.37	12.16	--
Curriculum						
Science Content	3.31	.92	.74	3.49	.59	.45
Teacher Background						
Professional Development	2.38	1.03	.92	2.74	1.02	.91
Experience	10.87	9.74	--	12.34	9.11	--
Experience at Grade	5.66	5.33	--	6.78	5.93	--
Classroom Context						
Percent Learned Concepts	5.22	.66	--	5.12	.88	--
Time on Task	42.44	26.67	--	54.66	19.33	--

Relationships between Teacher-Reported Practices and Student Achievement. As indicated earlier, we examined relationships between teacher-reported instructional practices and student achievement using regression models that controlled for prior achievement and student background characteristics. We estimated separate models for each of our sites, then conducted a pooled analysis across sites within a subject. Readers interested in individual site-model results are referred to Appendix F.

Figures 1 and 2 shows the relationship between teacher-reported reform practices and achievement pooled across our three mathematics and two science sites, respectively. We report standardized coefficients, which represent the expected difference in test score standard deviations for a one standard deviation unit increase in scores on the instructional practices scales. On the figure, the dark dot represents the point estimate for the coefficient and the bar represents the 95% confidence interval for that point estimate.

Figure 1 shows that few of the teacher-related variables in mathematics reached statistical significance, including our more innovative vignette-based measures. The amount of time the teacher spent covering proofs and patterns and teacher experience, both as a whole and at grade level, was positively related to student achievement. Of the student characteristics, classrooms with greater proportions of African-Americans and of students who were eligible for free-or reduced-price lunches were negatively related to mathematics achievements. In contrast, classrooms with greater proportions of White students showed positive relationships to SAT-9 scores.

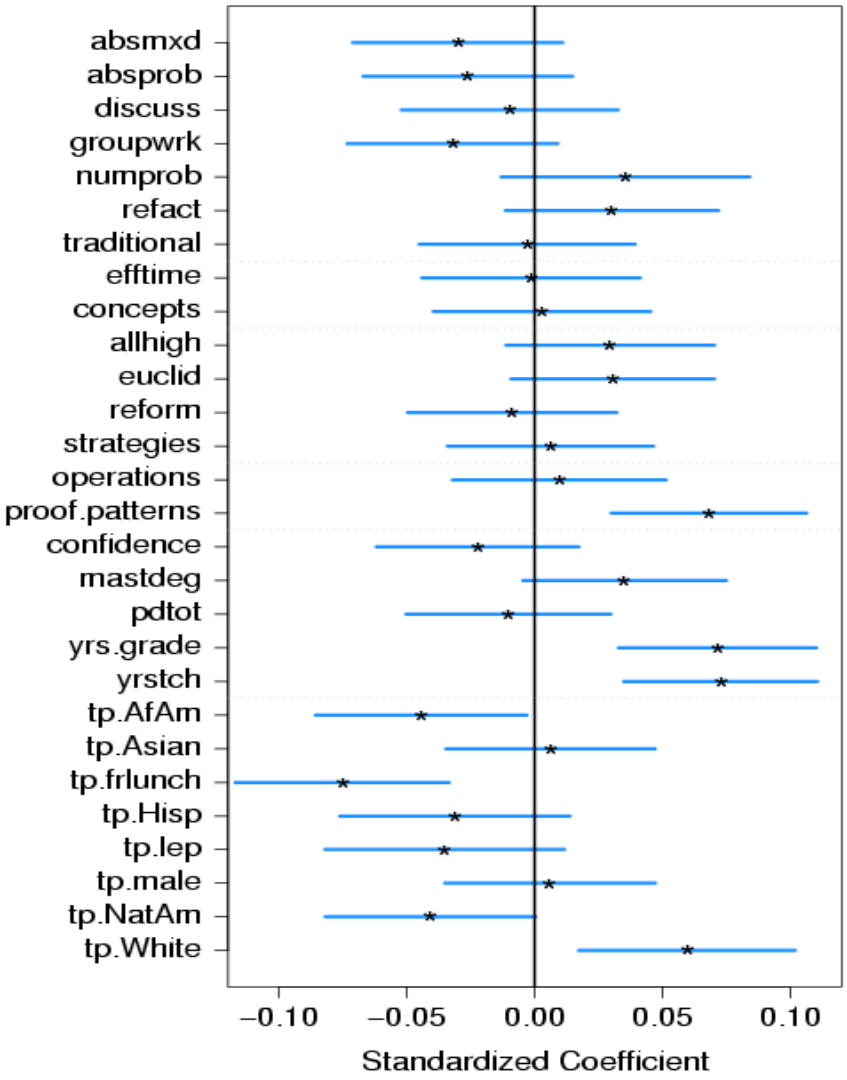

Figure 1. Pooled Estimates and 95% Confidence Intervals for Standardized Coefficients of Teacher Variables (Mathematics)

In science, teachers who reported more classroom discussion and more experience teaching at grade level had students who demonstrated higher achievement (see Figure 2). The effects, however, are marginally significant. None of the other measures of teaching practices or student characteristics showed a significant relationship with science test scores.

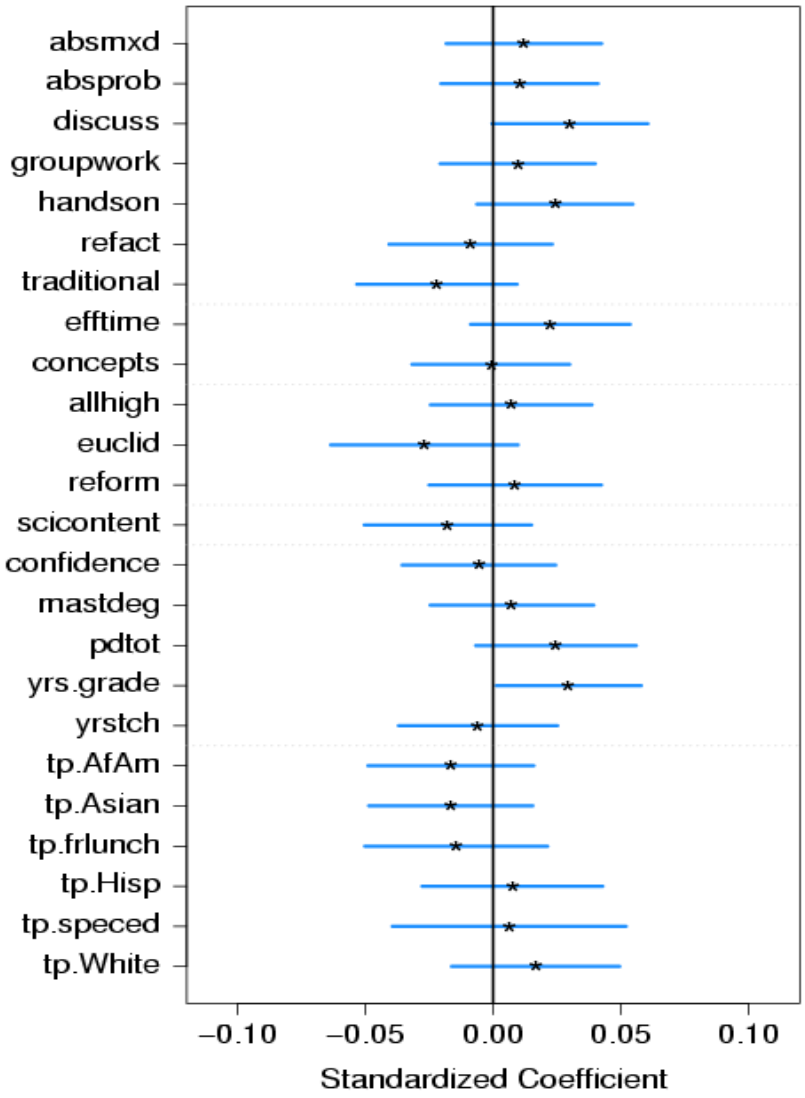

Figure 2. Pooled Estimates and 95% Confidence Intervals for Standardized Coefficients of Teacher Variables (Science)

Of particular interest is whether reform-oriented instructional practices bear any relationships with higher-order thinking skills. Figures 3 and 4 present the relationship

between reform teaching and achievement on open-ended mathematics and problem-solving items, respectively.[2] As with the earlier findings, very few measures of reform teaching are associated with student achievement. Groupwork and experience teaching at grade level are positively associated with higher scores on open-ended mathematics items (see Figure 3). Of the classroom and student characteristics, teachers with classrooms that learned many of the concepts, and classrooms with greater proportions of Hispanic and White students showed positive relations to open-ended test scores.

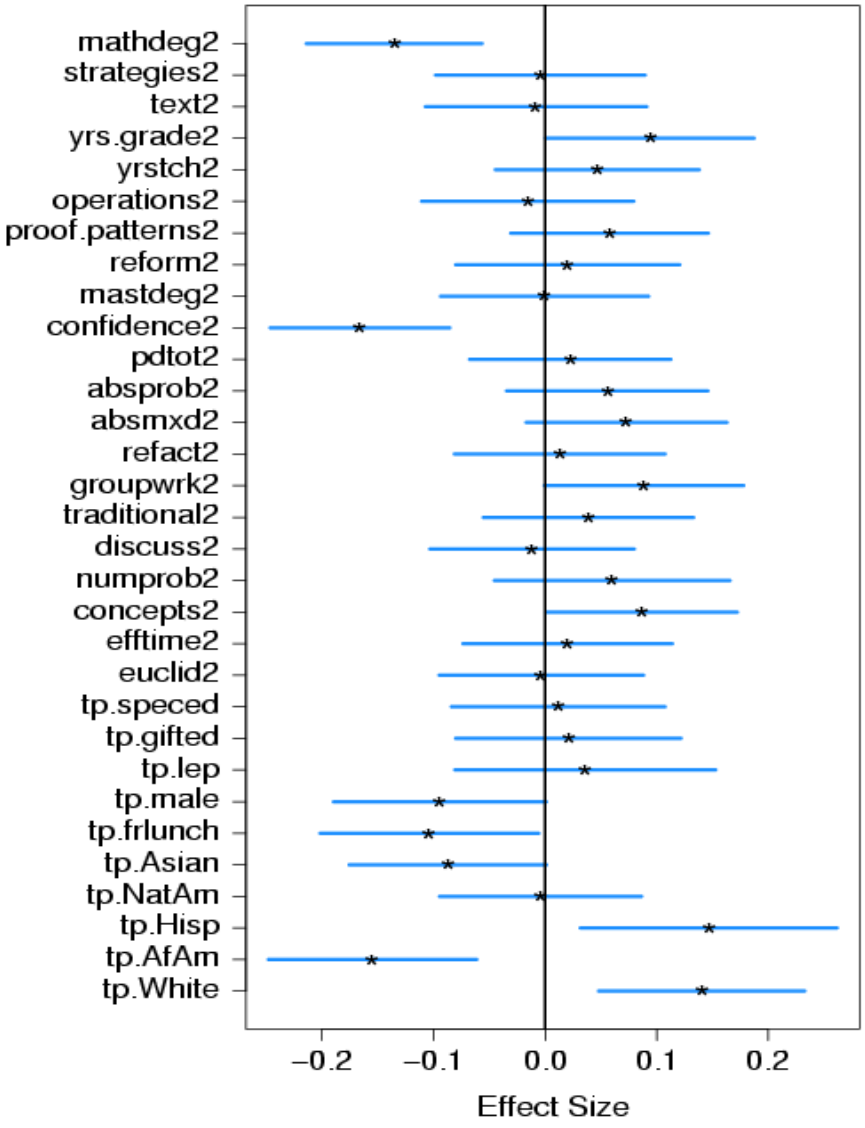

Figure 3. Estimated Point Estimates and 95% Confidence Intervals for Standardized Coefficients of Teacher Variables (Open-Ended Mathematics)

[2] This represents results for Cohort 4 only, which is the only site that had information on open-ended and problem-solving items.

Surprisingly, students whose teachers reported having a mathematics degree showed lower scores on the open-ended items. This trend was also observed for students whose teachers reported greater confidence in their mathematics knowledge. Additionally, some student characteristics were negatively related to open-ended performance. These included classrooms with larger percentages of Asians, African-Americans, males, and students who were eligible for free- and reduced-price lunches.

The relationships between instructional practices and performance on mathematics problem-solving items were similar to that observed for open-ended mathematics (see Figure 4). As with open-ended mathematics, students whose teachers had a math degree and expressed greater confidence in their mathematics knowledge showed lower scores on problem-solving items. Similarly, greater proportions of males and of students who were eligible for free- or reduced-price lunches were negatively related to problem-solving achievement. In contrast, greater proportion of White students was positively associated with problem-solving scores, although only marginally so.

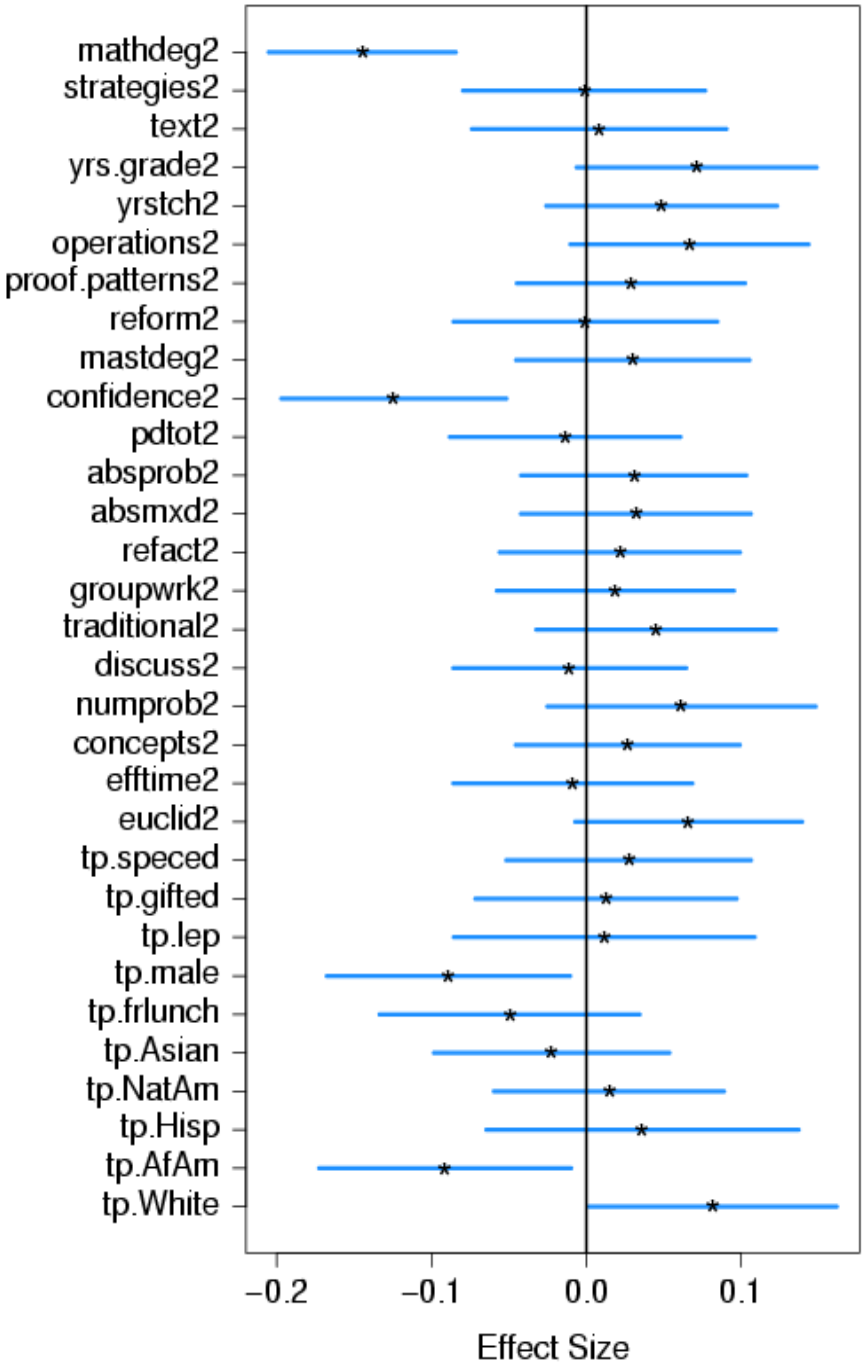

Figure 4. Estimated Point Estimates and 95% Confidence Intervals for Standardized Coefficients of Teacher Variables (Problem-Solving Mathematics)

Discussion

The purpose of this study was to examine the relationships between teachers' reported use of instructional practices in mathematics and science and student achievement. Previous research that has explored this issue has found that reform teaching is positively associated with student achievement, but the effects are small. One possible explanation underlying the small effects is the limitations arising from the use of surveys to measure instruction. In our study, we examined the relationship between instruction and achievement using a variety of measures, including vignettes, which we hoped captured aspects of teaching that could not be captured via surveys.

Regardless of whether instruction was assessed via surveys or vignettes, few of the instructional practices measures showed significant relationships with student achievement. Across both mathematics and science, number of years teaching at grade level was positively associated with test scores. In mathematics, total number of years teaching and the amount of time teachers spent covering proof and patterns was also positively related with performance. In science, students whose teachers encouraged classroom discussion showed slightly higher performance.

Because reform teaching is thought to have the greatest effect on "higher order thinking" skills and on open-ended tests, we undertook additional analysis in which we examined the relationship between reform pedagogy and performance on open-ended and problem-solving items in mathematics. The use of groupwork was associated with higher scores on open-ended items, but there was no relationship between any of the measures of reform instruction and student problem-solving scores. Additionally, we found students whose teachers had a mathematics degree and expressed more confidence in mathematics performed more poorly on open-ended and problem-solving items. It is unclear why these latter findings arise, but they are counterintuitive to our expectations. We are pursuing additional analysis to better understand whether self-selection effects or other factors may account for the anomalous results.

Taken together, the findings provide some support that certain aspects of reform pedagogy (e.g., discussion or groupwork) are associated with improved student achievement, but the effects are quite small. As mentioned earlier, small effects may stem from lack of measurement quality from our indicators of reform instruction. Despite our attempts to use vignettes to measure instruction in innovative ways, we could not address how reform instruction was actually implemented. Other methods that provide more refined measures of instruction, such as observations, may find stronger relationships between reform practices and student outcomes. We are currently undertaking an analysis in which we explore how observers' ratings of the "reformedness" of the classrooms relate to student achievement.

Small effects from our study may also stem from the fact that our analysis focused on students' exposure to practices during a single academic year. It is widely believed that students must be exposed to reform practices for more than a single year before the effects of these practices on achievement can become clearly evident. We have just completed our data collection efforts in which we followed students over three years, and

collected information about the instructional practices used by their teachers during each year. This data will allow us to explore whether the degree of exposure to the practices is related to student growth in achievement over a longer period of time.

References

American Association for the Advancement of Science (1993). Benchmarks for science literacy. New York: Oxford University Press.

Cohen, D.K., & Hill, H.C. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), 294-343.

Gamoran, A., Porter, A.C., Smithson, J., & White, P. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. Educational Evaluation and Policy Analysis, 19 (4), 325-338.

Hamilton, L.S., McCaffrey, D.F., Stecher, B.M., Klein, S.P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. Educational Evaluation and Policy Analysis, 25(1), 1-29.

Kim, J., Crasco, L., Blank, R., Smithson, J. (2001) Survey results of urban school classroom practices in mathematics and science: 2000 Report. Norwood: Systemic Research, Inc.

Le, V., Stecher, B.S., Hamilton, L.S., Ryan, G., Williams, V., Robyn, A., Alonzo, A. (2003). Vignette-based surveys and the Mosaic II project. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Mayer, D.P. (1998). Do new teaching standards undermine performance on old tests? Educational Evaluation and Policy Analysis, 20, 53-73.

Mullens, J., & Gayler, K. (1999). Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction. (NCES 1999-08). Washington, DC: National Center for Education Statistics.

National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2000). Principles and Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1996). National Science Standards. Washington, DC. National Academy Press.

Rowan, B., Correnti, R., & Miller, R.J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. The Teachers College Record, 104 (8), 1525-1567.

Saxe, G.B., Gearhart, M., & Seltzer, M (1999). Relations between classroom practices and student learning in the domain of fractions. Cognition and Instruction, 17(1), 1-24.

Stein, M.K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. Educational Research and Evaluation, 2, 50-80.

Swanson, C.B., & Stevenson, D.L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. Educational Evaluation and Policy Analysis, 24(1), 1-27.

Thompson, D, & Senk, S. (2001). The Effects of Curriculum on Achievement in Second Year Algebra: The Example of the University of Chicago Mathematics Project. Journal for Research in Mathematics Education, 32 (1), 58-84.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. Education Policy Analysis Archives, 10(12), http://epaa.asu.edu/epaa/v10n12/.

Appendix A: Student Demographic Characteristics for Each Site

Mathematics:

Demographic Characteristics	Cohort 1	Cohort 2	Cohort 4
Racial/Ethnic Group			
African-American	3.78	3.27	45.60
Hispanic	38.73	23.47	8.00
Other	5.05	6.36	22.00
White	52.44	66.90	24.40
Age			
Less than modal age	28.01	31.14	32.51
Exactly modal age	67.42	63.66	63.26
Greater than modal age	4.57	5.20	4.23
Female	49.51	50.00	51.03
Limited English Proficient	21.86	9.46	20.63
Two-Parent Family	73.93	76.53	--
Eligible for Free or Reduced Price Lunches	56.88	36.22	71.94
Special education	--	--	11.09
Gifted	--	--	18.34
Attendance Rate	--	--	
Less than 90 percent	--	--	22.57

Science:

Demographic Characteristics	Cohort 3	Cohort 5
Racial/Ethnic Group		
African-American	36.76	23.26
Hispanic	9.75	13.88
Other	21.87	12.53
White	31.62	50.33
Age		
Less than modal age	32.80	--
Exactly modal age	64.09	--
Greater than modal age	3.11	--
Female	45.02	48.40
Limited English Proficient	23.15	5.15
Eligible for Free or Reduced Price Lunches	63.45	20.88
Special education	9.75	12.72
Gifted	21.97	--

Appendix B: Sixth- and Seventh-Grade Mathematics Vignettes

Teaching Scenarios

Instructions. The following questions contain brief "scenarios" or stories that describe teaching situations and ask how you would respond in each case. We know there are many ways to teach mathematics, and you may not organize your lessons in the manner that is presented. Please answer as if you were in the situation that is described.

The scenarios are brief and do not describe every detail. Assume that other features are similar to your current school and your current students.

Please do the following:

a. Read the scenario.
b. Read the first possible option.
c. Circle the response that shows how likely you would be to do this option.
d. Read the next option and circle your response.
e. Repeat the process until you have responded to all the options.
f. Please evaluate each of the options independently of the others. In other words, you may select as many 1's (or 2's or 3's or 4's) as you like.

SCENARIO I: U.S. STANDARD MEASUREMENT UNITS (4 QUESTIONS)

Imagine you are teaching a sixth-grade class. You are about to begin a week-long unit on converting units of length within the U.S. standard measurement system. Your students have had experience using rulers to measure objects in feet and inches, and are also familiar with yards and miles as units of measurement.

1. You are ready to start the unit on conversion. How likely are you to do each of the following activities **to introduce** the unit?

	(Circle One Response in Each Row)			
	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask students what they know about inches and feet	1	2	3	4
b. Have students use rulers / yardsticks to measure lengths of objects in the classroom(e.g., desks or chairs)	1	2	3	4
c. Demonstrate how to solve problems such as converting 22 inches into feet and inches	1	2	3	4

d.	Display an equivalence table on the board that provides conversions among inches, feet, yards, and miles	1	2	3	4
e.	Have students solve a problem such as estimating the width of the classroom in inches	1	2	3	4
f.	Explain the procedures for converting units (e.g., multiply by 12 when converting feet into inches)	1	2	3	4
g.	Lead a classroom discussion about the problems of measuring if you only had one unit of measurement (e.g., foot)	1	2	3	4
h.	Have students work in pairs or groups to measure the size of each other's feet	1	2	3	4

2. You are at the midpoint of your unit on conversion, and most students appear to understand the procedures. Next, you pose more complex problems. You ask your students how many inches are in 9 yards, 2 feet.

   When most students appear to have completed the task, you ask Joey if he will share his solution. He replies that 9 yards, 2 feet is close to 10 yards, which is 360 inches, so he subtracted, and found the answer to be 358 inches.

   You know, however, that the correct answer is 348 inches.

   After praising Joey for knowing that 9 yards, 2 feet is close to 10 yards, what do you do next? How likely are you to do each of the following?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask Joey, "How did you get from 10 yards to 358 inches?"	1	2	3	4
b. Pose another similar problem for the class	1	2	3	4
c. Suggest that Joey use a ruler to solve the problem	1	2	3	4
d. Tell Joey that he was close, but the answer is 348	1	2	3	4
e. Call on another student who you expect will give you the right answer	1	2	3	4
f. Tell Joey that his answer is close, and ask if anyone can help him with his solution	1	2	3	4
g. Ask the class, "Did anyone else use a similar method but get a different answer?"	1	2	3	4

h. Explain that one foot (12 inches) should have been subtracted	1	2	3	4
i. Ask the class, "Are there any other answers?"	1	2	3	4
j. Give Joey another problem similar to this one, and ask him to solve it	1	2	3	4

3. You are almost at the end of the unit on conversion. You ask students to work in pairs or groups to solve the following problem.

$$\begin{array}{r} 5 \text{ ft } 3 \text{ in} \\ -\ 3 \text{ ft } 6 \text{ in} \end{array}$$

After working on the problem for a while, you ask each group if they will share their work.

The first group responds that the answer is 1 feet 9 inches. They explain that they converted 5 feet 3 inches to 4 feet 15 inches, then subtracted.

The second group gives the same answer, and explains that they drew the distances on the floor using a yardstick and measured the non-overlapping portion.

How likely are you to do each of the following in response to these two explanations?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask the class if they can think of other ways to solve the problem	1	2	3	4
b. Think of a new problem in which the two methods are not equally effective and ask the groups to solve it	1	2	3	4
c. Tell them that they are both right and move on to the next problem	1	2	3	4
d. Tell them that it is better to use the first group's method because it can be applied to any similar distance problems	1	2	3	4
e. Have a classroom discussion about the differences between the two approaches	1	2	3	4

4. If you were to teach a unit on conversion of lengths to the target class, how much emphasis would you place on each of the following learning objectives?

(Circle One Response in Each Row)

	No emphasis	**Slight emphasis**	**Moderate emphasis**	**Great emphasis**
a. Students will understand that similar principles of conversion apply in other situations (e.g., when measuring area, volume)	1	2	3	4
b. Students will be able to use rulers and yardsticks to solve conversion problems (e.g., show why there are 48 inches in 1 yard, 1 foot)	1	2	3	4
c. Students will be able to solve mixed-unit problems (e.g., converting 1 yard 2 feet to inches)	1	2	3	4
d. Students will be able to estimate the lengths of objects in their neighborhoods (e.g., cars)	1	2	3	4
e. Students will know how to convert among inches, feet, and yards	1	2	3	4
f. Students will know which units of measurement are appropriate for measuring objects or distances of differing length	1	2	3	4

SCENARIO II: DIVIDING BY FRACTIONS (4 questions)

Instructions. Please imagine that you are in the situation that is described, and indicate how likely or unlikely you would be to give each of the possible options. We understand that you may not actually organize your lessons in the manner that is presented here, but try to respond as if you were in the situation.

Read each option and circle the number that shows how likely you would be to do that option. Remember, you may select as many 1's (or 2's or 3's or 4's) as you like.

Imagine you are teaching a sixth-grade class. The students are familiar with division of whole numbers, and you are about to begin a week-long unit on dividing by fractions (e.g., 1/5 divided by 2/3).

5. You are ready to start the unit on dividing by fractions. How likely are you to do each of the following activities **to introduce** the unit?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Have students work in pairs or groups to solve a problem such as dividing 8 by 1/4	1	2	3	4
b. Explain the procedures for dividing fractions (e.g., multiply the first number by the reciprocal of the second number)	1	2	3	4
c. Have students use fraction bars or drawings to divide fractions	1	2	3	4
d. Ask students what they know about dividing fractions	1	2	3	4
e. Ask students to solve a problem such as finding how many batches of cookies could be made if a cookie recipe used 1/5 cups of flour and there were 3 cups of flour	1	2	3	4
f. Remind students of ways to think about dividing whole numbers	1	2	3	4

g.	Lead a classroom discussion about dividing by fractions using real-world examples	1	2	3	4
h.	Demonstrate how to solve problems such as dividing 6 by 1/2	1	2	3	4

6. You are at the midpoint of your unit on dividing by fractions, and most students appear to understand the procedures. Next, you pose more complex problems. You ask students to divide 4/5 by 3/4.

When most students appear to have completed the task, you ask Martina if she will share her solution.
She replies that she converted 4/5 into 16/20 and 3/4 into 15/20 and then divided to get 15/16.

You know, however, that the correct answer is 16/15.

After praising Martina for knowing how to find a common denominator, what do you do next? How likely are you to do the following in response to Martina?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask the class, "Are there any other answers?"	1	2	3	4
b. Tell Martina that she was close, but the answer is 16/15	1	2	3	4
c. Ask Martina, "How did you get from 16/20 and 15/20 to 15/16?"	1	2	3	4
d. Pose another problem for the class	1	2	3	4
e. Suggest that Martina use manipulatives (e.g., counters) to solve the problem	1	2	3	4
f. Call on another student who you expect will give you the right answer	1	2	3	4
g. Ask the class, "Did anyone else use a similar method but get a different answer?"	1	2	3	4
h. Explain that once a common denominator is found, the answer is the quotient of the numerators	1	2	3	4
i. Give Martina another problem similar to this one, and ask her to solve it	1	2	3	4

	1	2	3	4
j. Tell Martha that her answer was close and ask if anyone can help her with her solution				

6. You are almost at the end of the unit on dividing by fractions. You ask students to work in pairs or groups to divide 10/4 by 1/6.

   After working on the problem for a while, you ask each group if they will share their work.

   > The first group says the answer is 15. They explain that they multiplied 10/4 by the reciprocal of 1/6, and reduced.
   >
   > The second group gives the same answer but gives a different explanation. First, they reduced 10/4 to 2 1/2. They knew there were 12 one-sixths in 2 and 3 one-sixths in 1/2, so together there would be 15 one-sixths in 2 1/2.

   How likely are you to do each of the following in response to these two explanations?

   (Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Tell them that it is better to use the first group's method because it is faster	1	2	3	4
b. Think of a new problem in which the two methods are not equally effective and ask the groups to solve it	1	2	3	4
c. Have a classroom discussion about the differences between the two approaches	1	2	3	4
d. Ask the class if they can think of other ways to solve the problem	1	2	3	4
e. Tell them that they are both right and move on to the next problem	1	2	3	4

8. If you were to design a unit on dividing by fractions <u>to the target class</u>, how much emphasis would you place on each of the following learning objectives?

(Circle One Response in Each Row)

	No emphasis	**Slight emphasis**	**Moderate emphasis**	**Great emphasis**
a. Students will understand that division of fractions entails finding the number of times the second number goes into the first number	1	2	3	4
b. Students will know the procedures for dividing by fractions (e.g., multiply the first number by the reciprocal of the second number)	1	2	3	4
c. Students will be able to solve problems such as dividing 5/4 by 2/3	1	2	3	4
d. Students will understand that operations that are possible with whole numbers are possible with fractions	1	2	3	4
e. Students will recognize when a situation calls for dividing by a fraction	1	2	3	4
f. Students will be able to use fraction bars to divide by a fraction	1	2	3	4

Appendix C: Judgmental Reform Ratings for the Response Options from Sixth-and Seventh-Grade Mathematics

Item No.	Reform Rating
Vignette 1 (Measurement System)	
1a	3
1b	3
1c	1
1d	1
1e	3
1f	1
1g	4
1h	3
2a	4
2b	2
2c	2
2d	1
2e	1
2f	2
2g	3
2h	1
2i	--
2j	2
3a	3
3b	4
3c	1
3d	2
3e	4
4a	4
4b	2
4c	--
4d	3
4e	--
4f	4
Vignette 1 (Division of Fractions)	
5a	3
5b	1
5c	3
5d	3
5e	4
5f	2
5g	4
5h	1
6a	--
6b	1
6c	4
6d	2
6e	--
6f	1

6g	3
6h	1
6i	2
6j	2
7a	1
7b	4
7c	4
7d	3
7e	1
8a	2
8b	--
8c	--
8d	3
8e	4
8f	2

Appendix D: Survey and Log Scale Items

Survey

Instructional Practices

Reform Practices

On average throughout the year, approximately how often do you employ the following teaching strategies during your mathematics lessons?

- Use open-ended questions
- Require students to explain their reasoning when giving an answer
- Encourage students to communicate mathematically
- Encourage students to explore alternative methods for solutions
- Help students see connections between mathematics and other disciplines

On average throughout the year, approximately how often do your students take part in the following activities as part of their mathematics lessons?

- Share ideas or solve problems with each other in small groups
- Engage in hands-on mathematics activities
- Work on extended mathematics investigations (a week or more in duration)
- Record, represent, or analyze data

Strategies

Overall, during these lessons, what portion of the time was devoted to each of the following activities? (Third-grade version)

- Reviewing single-digit multiplication facts
- Using manipulatives to show multiplication as adding groups of objects
- Graphically representing multiplication in the context of rows and columns
- Identifying shorthand multiplication rules (e.g., to multiply by 10, add a zero to the end of the number)
- Practicing traditional multiplication algorithms
- Investigating patterns in the multiplication table
- Discussing multiplication in the context of the distributive property (e.g., 26 x 4 is 20 x 4 plus 6 x 4)
- Discussing multiplication as repeated addition

(Sixth- and seventh-grade version)

- Reviewing the names of the tenths, hundredths, and thousandths places
- Developing an equivalence table showing common fractions and their decimal equivalents
- Representing decimals and fractions as parts of the same figures (e.g., grids)
- Converting fractions to decimals by dividing the numerator by the denominator
- Converting decimals to fractions by finding a power of 10 to represent the denominator
- Learning common equivalents (e.g., ½ = .5; ¼ = .25)
- Identifying shorthand conversion strategies (e.g., if 1/8 is .125, then 3/8 is .375)
- Using manipulatives to show equivalence between fractions and decimals
- Using a calculator to convert fractions to decimals

Teacher Background

Certification

What type of teaching certification do you hold? (Circle one)

- Not certified
- Temporary, provisional, or emergency certification (requires additional coursework before regular certification can be obtained)
- Probationary certification (the initial certification issued after satisfying all requirements except the completion of a probationary period)
- Regular or standard certification

Confidence

With respect to the mathematics/science that you are asked to teach, how confident are you in your mathematical/scientific knowledge? (Circle one)

- Not confident at all
- Somewhat confident

Moderately confident
Very confident

Masters Degree
What is the highest degree you hold?
BA or BS
MA or MS
Multiple MA or MS
PhD or EdD
Other

Math Degree
Did you major in mathematics/science or a mathematics/science-intensive field for your Bachelor's degree?
Did you minor in mathematics or a mathematics/science-intensive field for your Bachelor's degree?

Professional Development
In the past 12 months, how much time have you spent on professional development activities that focused on the following aspects of teaching mathematics?
In-depth study of mathematics/science content
Methods of teaching mathematics/science
Use of particular mathematics/science curricula or curriculum materials
Students' mathematical/scientific thinking
Mathematics/science standards or framework
Mathematics/science assessment/testing
Use of educational technology for mathematics/science instruction

Experience
Including this year, how many years have you taught on a full-time basis?

Experience at Grade
Including this year, how many years have you taught third-graders? (*third-grade version)3*

Curriculum
Operations
Indicate the approximate amount of time you will spend on each content area this school year?
(Third-grade version)
Addition/subtraction of whole numbers
Multiplication/division of whole numbers

(Sixth- and seventh-grade version)
Operations with signed whole numbers

Proof and Patterns
Indicate the approximate amount of time you will spend on each content area this school year?
(Third- and sixth-grade version)
Proof and justification/verification
Patterns/functions/algebra
(Seventh-grade version)
Proof and justification/verification
Patterns/functions
Algebra

Science Content
Approximately how many lessons will you present on each of the following areas of science in your class this school year? (Third-grade version)

Life Science
Earth Science
Physical Science
Science Reasoning and Technology

How much time will you spend teaching each of the following areas of science in the target class this school year?(Sixth-grade version)

Patterns and relationships
Diversity and adaptation of organisms
Chesapeake Bay populations and ecosystems
Motion and Forces

Log

Instructional Practices

Discussion

How much time did students spend on each of these activities during today's mathematics lesson?

Explain their thinking about mathematical problems
Lead discussion of a mathematics topic

How much time did you spend on each of these activities during today's mathematics lesson?

Ask open-ended questions and discuss solutions to them
Ask questions of individuals to test for understanding

Groupwork

How long did students work in groups during today's mathematics/science lesson?

Mixed-ability Groupwork

If groups were used, what share of the group time was used in the following ways?

Working in groups of mixed ability

Problem-solving Groupwork

If groups were used, what share of the group time was used in the following ways?

Solving new problems together as a group

Reform Activities

How often did the following occur during today's mathematics lesson? (Mathematics version)

Students engaged in debate/discussion about ways to solve a problem
Student restated another student's ideas in different words
Students demonstrated different ways to solve a problem
Students explored a problem different from any they had solved previously
Students worked on an activity or problem that will take more than one period
Teacher or student connected today's math topic to another subject (e.g., social studies)

How often did the following occur during today's science lesson? (Science version)

Students engaged in debate / discussion about a scientific conclusion or generalization
Student restated another student's ideas in different words
Students represented data using charts, graphs or tables
Students worked on a problem or investigation different from any done previously
Students worked on a problem or investigation that will take more than one period
Teacher or student connected today's topic to another subject (e.g., social studies)

Seatwork

How much time did students spend on each of these activities during today's mathematics lesson?

Mathematics version

Read from textbook or other materials
Complete worksheets or problem sets from text

Hands-On

How much time did students spend doing hands-on science investigations?

Number of Problems Solved
How many math problems did the typical student work on in class today?

Classroom Context
Percent Learned Concepts
About what percent of the students learned the concepts or skills you expected them to learn today?

Time on Task
How long was today's mathematics/science lesson?
How much mathematics/science time was lost to discipline issues?

Appendix E: Regression Modeling Approaches

General Strategy For Modeling Student-Level Data. The general strategy we used in each of our sites was to fit a linear mixed effects model of the following form:

$$Y_{ij} = \alpha + \boldsymbol{\beta}^T \mathbf{X}_{ij} + \boldsymbol{\gamma}^T \mathbf{U}_{ij} + \theta_i + \varepsilon_{ij} \quad (1)$$

where i indexes teachers, j indexes students within teachers, α is a grand mean, Y_{ij} is a student test score, $\mathbf{X}_{ij}$ is a vector of student background characteristics and $\boldsymbol{\beta}$ is a vector of fixed effects for the student characteristics. $\mathbf{U}_{ij}$ is a vector of polynomial terms (up to degree three) of centered and scaled prior year scores for a student, with associated fixed effects $\boldsymbol{\gamma}$. The random classroom intercepts θ_i are assumed to be iid $N(0, \sigma_\theta^2)$, and the error terms ε_{ij} are assumed to be iid $N(0, \sigma^2)$ and independent of the θ_i.

The covariates included all of the demographic variables for the students available at each site. In general, these included student ethnicity, gender, age, gifted education status, special education status, free lunch status, and LEP status. Where available we also used the year-to-date percentage attendance, whether or not English is the primary home language, and an indicator of whether the child comes from a two-parent family. All but the year-todate percentage attendance were treated as categorical (i.e. class variables). This included student age, which was categorized into approximately three classes depending on the site. In particular we wanted to control for students who were exceptionally old or young for their cohort. For all sites, to ensure a conservative analysis, we included all student covariates in all models. Given the large number of students relative to the number of covariates, this resulted in no notable loss of efficiency in estimating parameters.

In addition to the student background characteristics, we also controlled for prior achievement, which was available in some form for all but the Cohort 3 science site (for which we conservatively used a current year score as a control). When available, we used prior scores for both reading and mathematics because excluding one or the other left correlation between the residuals of the model and the excluded score, which could introduce bias into our analyses. We centered and scaled all prior year scores to have mean zero and variance one. In addition, for most outcomes, we needed to include quadratic and/or cubic functions of the standardized prior year scores on the right-hand side of the models to account for notable non-linear structure in the residuals without these terms. Such non-linearities are the result of inherently non-linear relationships between outcomes (particularly scaled scores) on different assessments.

We also examined whether the random classroom intercepts were sufficient to capture heterogeneity among classrooms by fitting (where possible) the fixed effects models to each classroom separately and examining the variations across classrooms in the estimated parameters. Fortunately, there was no strong evidence of between-classroom variation in any parameters other than the intercepts, making the random classroom intercepts model appropriate (i.e. no random slopes were necessary).

All models were estimated via restricted maximum likelihood using the package `nlme` for the R statistics package. For all models, we performed batteries of residual diagnostics to check for influential observations and to ensure that all model assumptions were adequately met. All results reported here are from models that have been checked as such. In a few cases there was evidence of approximately linearly increasing residual variance as a function of a prior year test score; in those cases we variance weighted the observations by the value of the prior year score. Also where necessary we tested the sensitivity to the inclusion or exclusion of a few classrooms that had atypically large proportions of gifted or special education students; the impacts on substantive results were negligible, so we left all such classrooms in the analysis. A handful of either certainly or likely miscoded student observations were removed from the analysis. Because these numbers were relatively small percentages of the total numbers of students, and because they were distributed across approximately half of the total number of teachers, we eliminated these students from the analysis and expect little impact.

We explored a diverse array of outcomes as response variables, including all of the different assessments for the subject in question, as well as different transformations of the same assessment. The different transformations of a given outcome included, when available, the raw score, the scaled score, and the arcsine square-root transformation of the percentage correct (a popular variance stabilization transformation for proportion data). In all cases, the substantive results about the relationships between teacher characteristics/teaching practices and student achievement were essentially invariant to which of these three versions of an outcome was used as a response. The best linear unbiased predictors of the random effects from the mixed models fit to these different responses were correlated at greater than 0.95 in all cases. Because the scaled score is intended to represent a meaningful latent scale of achievement, all results we present in the remainder of this report are for scaled scores unless scaled scores were not available for a primary outcome variable, in which case we used the arcsine square-root transformation of the percent correct as a response.

While the inferences were not sensitive to different transformations of the same outcome, there were cases of notable interactions of inferences with different outcomes. We briefly discuss these in the summaries of the individual site analyses.

Adding Teacher-Level Variables to Student-Level Model. The overall goal of our analysis was to examine how well the heterogeneity in classroom intercepts could be explained by the teacher-level variables (i.e., scales relating to teacher background and classroom practices). In addition to these variables, we also considered relationships with classroom contextual variables, including classroom percentages of different student ethnic groups, male students, special education students, gifted students, LEP students, and students participating in free/reduced price lunch programs.

As an exploratory tool as well as to identify outliers and influential observations, for every model under consideration, we plotted the estimated classroom random effects (the BLUPs) for the models against each of the classroom-level variables. To test for potential significance of the effects of the teacher variables, we examined the effects of adding

each of these variables individually to each outcome model for each site. That is, for each teacher or classroom variable V, we fit the model:

$$Y_{ij} = \alpha + \beta_T X_{ij} + \gamma_T U_{ij} + \delta V_i + \theta_{.i} + \varepsilon_{ij}$$

where δ is a fixed effect, $\theta_{.i}$ are again iid normal random intercepts representing residual classroom variation after accounting for V, and all other terms are as described above. A significant value of δ implies that V helps to explain the observed heterogeneity in adjusted mean outcomes across the classrooms.

Imputing Missing Student Data. One complication to all of the analyses were missing test scores. Percentages of missingness varied from generally in the 10-20 percent range for national tests and somewhat less for district or state tests. In order to account properly for missing test scores in our analyses, we used multiple imputation techniques as implemented by Schafer *et al* in the R package `norm`. When raw scores were available, we carried out imputation on the raw scores rather than scaled scores. This was because although in general the marginal distributions of the scaled scores were closer to Gaussian, the conditional expectations of the raw scores given the other raw scores were closer to linear. This latter condition is more essential to the success of the algorithm. When raw scores were not available, we used scaled scores or percentages correct to perform the imputation, because in no case was the deviation from joint normality overly egregious. All imputed raw scores were mapped to the nearest valid raw score, and then subsequently mapped to the corresponding scaled score. Thus the imputations result in plausible realizations of both the raw score and the scaled score for every missing value on every test.

In order to preserve as much of the multivariate structure of the data as possible, the imputation model generally included all of the available outcomes (both prior and current year) in the particular data set, variables for all of the student demographic characteristics, and teacher-level fixed effects to maintain the teacher-level structure that is most important to the analysis. We checked for second-order interactions among these variables via mixed models on the complete cases and there was no evidence for them, so the first order linear imputation model was deemed sufficient. In some cases because of singularities in the imputation model resulting from a few small classrooms with some scores entirely missing for the classroom, it was necessary to combine small numbers of the smallest classes. We checked to see that such classrooms were not markedly different. Although the imputations resulted in plausible outcomes for all students on all assessments, only those students who had an observed value of a particular outcome were used in models for that outcome. The imputations were used only to account for missingness in prior year scores used as covariates in the model.

For each site requiring imputation (all but the Cohort 3 science which had no prior year scores), this procedure was carried out 10 times resulting in 10 replicate data sets. All key analyses were carried out on all of the imputed data sets. The resulting estimates, standard errors and significance calculations were based on combining quantities via the appropriate multiple imputation procedures, which accounts for the uncertainty due to missingness.

Aggregation of Results Across Sites. To synthesize our results across sites, we also performed a meta-analysis of the standardized coefficients, separately for science and math. For each site we chose the standardized coefficients from a single outcome to estimate aggregate standardized coefficients for all teacher variables that are common to all sites. For the science outcomes (SAT-9 multiple choice science) we chose the Cohort 3 science results that adjusted for a current year score to be conservative. For the math sites, we used the SAT-9 multiple choice mathematics scaled score, the only assessment common to all three sites. We estimated the composite standardized coefficients as a weighted average of the individual standardized coefficients, where the weight for each is proportional to its estimated precision (as determined from the standard errors from the model fits). This gives more weight to coefficients estimated with less variance. Specifically, for each variable, the model in site s provides an estimated standardized coefficient $\hat{a}_s$ with associated estimated variance v_s and precision $p_s = 1/v_s$. Because the estimator is a linear combination of (assumed) independent normal random variables, its variance is equal to $(\sum_s p_s)^{-1}$, which reduces to the usual variance of a sample mean when all variances are equal. Confidence interval limits are determined in a straightforward fashion from the quantiles of the appropriate normal distribution.

Appendix F: Results for Individual Sites

In this section we summarize the results for the individual teacher variables for each of the outcomes/sites. In all models, the student variables were generally significant, most notably prior achievement, ethnicity, and free/reduced price lunch status. The variables were relatively highly collinear and thus it is difficult to interpret individual coefficients. Moreover, they are not of direct interest, but rather are included only to make more defensible inferences about the relationships between teaching practices and student achievement. Thus we focus on the teacher variables and do not present any results about the student fixed effects.

We did a very thorough job of controlling for everything we could about the students, erring on the side of conservatism (e.g. by including all available information about students regardless of significance). There was still statistically significant classroom heterogeneity remaining in all cases, though the size of this variability relative to the within-classroom variability varied greatly across sites. The lowest was in Cohort 3, where the estimated between-classroom variance component $\hat{\tau}^2$ was only about 3 percent of the estimated residual (within-classroom) variance $\hat{\sigma}^2$. The highest were in Cohort 1 and Cohort 4, where on some outcomes the percentage was around 40.
The main results concerning the teacher variables are presented in Tables A1 and A2 for science and mathematics, respectively. For each classroom-level variable, we transform the coefficient α obtained from the model described previously into an estimated standardized coefficient by multiplying the coefficient by the sample standard of the variable across teachers and dividing by the marginal sample standard deviation of the outcome on the left-hand side of the model. This is a conservative metric whose scale facilitates comparison across sites where potentially different student-level covariates are included in the models. The interpretation on this scale is an increase in the teacher-level variable by one standard deviation is associated with a K population standard deviation increase in the outcome. The specific notation used in the tables is described in the captions.

For Table A1, the Cohort 3 site had no measure of prior achievement. We considered two general approaches. The first used no control for any other achievement measure, and the other used a current year score (the scaled score on the Problem Solving part of the SAT-9 math exam). The latter analysis is conservative because it likely removes some of the effects that we are trying to capture. However, it avoids the likely positive bias that would result by not controlling for prior achievement. In all for both science sites the standardized coefficients are very small, indicating that the teacher-level variables did not help to explain much of the variance in student outcomes. For both sites the relative size of the between to within classroom variance was small, so this was not surprising.

Table A1. Summary of the Standardized Coefficients for the Science Sites on the Sat-9

Scales	Cohort 3	Cohort 5
Instructional Practices		
Absmxd		
Absprob		

Discuss	+	
Groupwork		
Handson		
Refact		
Traditional		
Allhigh		
Euclid		
Reform		
Curriculum Coverage		
Chesapeake	N/A	
Chesapeake.goals	N/A	
Diversity	N/A	
Earthsci		N/A
Foodchains		N/A
Fosstot		N/A
Ideas		N/A
Lifesci		N/A
Lifesci.goals	+	N/A
Motions	N/A	
Patterns	N/A	
Patterns.goals	N/A	
Physci	-	N/A
Physci.goals		N/A
Physicssound		N/A
Scicontent		
Scitech		N/A
Sleuthgoals	N/A	
Sleuths	N/A	
Sound		N/A
Structures		N/A
Water		N/A
Teacher Background		
Confidence		
Mastdeg		
Mastchr	N/A	
Pdtot		
Resource	N/A	
Scideg	N/A	
Yrs.grade		
Yrstch		
Classroom Context		
Efftime		
Concepts		
Tp.AfAm	-	
Tp.Asian		
Tp.frlunch	-	
Tp.gifted		N/A
Tp.Hispanic		

Tp.lep	+	N/A
Tp.male	-	N/A
Tp.NatAm		N/A
Tp.speced		
Tp.White	+	

Notes.
Scales under the Curriculum Coverage section represent specific science topics, most of which were specific to the site. More information is available from the authors.
N/A indicates that a variable was not measured for the site.
One sign indicates a standardized coefficient between .05 and .10.
Two signs indicate a standardized coefficient between .10 and .15.
Three signs indicate a standardized coefficient greater than .15.
* indicates $p < .05$.
** indicates $p < .01$

Analogous results are presented for mathematics in Table A2. The table focuses on the outcome that was common to all three sites, namely, the SAT-9 multiple-choice examination. There are a larger number of larger standardized coefficients than for science, though all are less than 0.2 in absolute value. Nonetheless there is some evidence of some interesting trends. The variables that exhibit a consistent signal across sites are years teaching (both total and at the particular grade level) which is positive, and the classroom percent of students participating in free-or reduced-price lunch programs, which is negative. Other of the classroom context variables show relationships of expected directions. The two strongly negative coefficients for Confidence and Math Degree in Cohort 4 are presumably anomalous.

Table A2. Summary of the Standardized Coefficients for the Mathematics Sites on the SAT-9

Scales	Cohort 1	Cohort 2	Cohort 4
Instructional Practices			
Absmxd	--*		
Absprob	--*		
Discuss			-
Groupwork	--*		
Numprob	+		++**
Refact		+*	
Traditional	+*		
Allhigh	+		+*
Euclid			++*
Reform			
Strategies	+		
Curriculum Coverage			
Operations			
Proof.patterns		++**	+
Teacher Background			
Confidence	+		---**
Mastdeg			+
Mathdeg	N/A		---**

Pdtot	-		
Yrs.grade	+	+*	+*
Yrstch	+	+**	+
Classroom Context			
Efftime			
Concepts			
Tp.AfAm			-
Tp.Asian			
Tp.frlunch	-	-*	-
Tp.gifted	N/A	N/A	+
Tp.Hispanic			
Tp.lep	-		
Tp.male	+		-
Tp.NatAm	-		
Tp.singleparent	--*	-*	N/A
Tp.speced	N/A	N/A	
Tp.White		+	++*

Notes.
N/A indicates that a variable was not measured for the site.
One sign indicates a standardized coefficient between .05 and .10.
Two signs indicate a standardized coefficient between .10 and .15.
Three signs indicate a standardized coefficient greater than .15.
* indicates $p < .05$.
** indicates $p < .01$

The results concerning the teacher practices are mixed. There are cases of reform-oriented activities being positively, negatively, or unrelated to outcomes, and the same for more traditional activities. Aside from the Reform Inclinations scale, there seems to be some tendency for reform-oriented activities to be signed negatively more often and for more traditional practices to be signed positively more often, though the signal is not strong. While we treated the SAT-9 MC results are the primary outcome, results for additional outcomes in all three math sites are provided in Table A3. The primary trends found in the SAT-9 MC, i.e., positive effects of years teaching experience and relatively strong classroom contextual effects – are again evident. Note that the district outcome in Cohort 2 exhibits patterns substantially different from what is indicated by the SAT-9, particularly in the classroom contextual effects. This underscores the importance of the interplay between teaching practices and the assessment on which outcomes are measured.

Table A3. Summary of the Standardized Coefficients for the Mathematics Sites on the District Related Tests (Cohorts 1 and 2 Only)

Scales	Cohort 1	Cohort 2
Instructional Practices		
Absmxd	--*	-
Absprob	-	-
Discuss		++
Groupwork	-	-

Numprob		+
Refact		
Traditional	+	
Allhigh		
Euclid		
Reform		+
Strategies	+	++
Curriculum Coverage		
Operations		-
Proof.patterns		
Teacher Background		
Confidence		+
Mastdeg		
Mathdeg	N/A	
Pdtot		++
Yrs.grade	+$^{+}$	+
Yrstch	+	+
Classroom Context		
Efftime	+	-
Concepts		
Tp.AfAm	-*	+++**
Tp.Asian		--**
Tp.frlunch	-	+
Tp.Hispanic		+
Tp.lep		+
Tp.male	+	
Tp.NatAm		+
Tp.singleparent	--$^{+}$	+++**
Tp.White		--

Notes.
N/A indicates that a variable was not measured for the site.
One sign indicates a standardized coefficient between .05 and .10.
Two signs indicate a standardized coefficient between .10 and .15.
Three signs indicate a standardized coefficient greater than .15.
* indicates $p < .05$.
** indicates $p < .01$