REPRINT

Studying Large-Scale Reforms of Instructional Practice

An Example from Mathematics and Science

Laura S. Hamilton, Daniel F. McCaffrey,
Brian M. Stecher, Stephen P. Klein,
Abby Robyn, Delia Bugliari

Reprinted from *Educational Evaluation and Policy Analysis*

EDUCATION

Published 2005 by the RAND Corporation

To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Educational Evaluation and Policy Analysis
Spring 2003, Vol. 25, No. 1, pp. 1–29

Studying Large-Scale Reforms of Instructional Practice: An Example from Mathematics and Science

Laura S. Hamilton
Daniel F. McCaffrey
Brian M. Stecher
Stephen P. Klein
Abby Robyn
Delia Bugliari
RAND

A number of challenges are encountered when evaluating a large-scale, multisite educational reform aimed at changing classroom practice. The challenges include substantial variability in implementation with little information on actual practice, lack of common, appropriate outcome measures, and the need to synthesize evaluation results across multiple study sites. This article describes an approach to addressing these challenges in the context of a study of the relationships between student achievement and instructional practices in the National Science Foundation's Systemic Initiatives (SI) program. We gathered data from eleven SI sites and investigated relationships at the site level and pooled across sites using a planned meta-analytic approach. We found small but consistent positive relationships between teachers' reported use of standards-based instruction and student achievement. The article also describes the ways in which we addressed the challenges discussed, as well as a number of additional obstacles that need to be addressed to improve future evaluations of large-scale reforms.

Keywords: *program evaluation, student achievement, systemic reform*

RESEARCHERS face a number of challenges when attempting to evaluate large-scale educational reforms that are designed to affect classroom practice.[1] Such reforms are common at present—examples include the National Science Foundation's Systemic Initiatives programs, the New American Schools, and the Comprehensive School Reform Demonstration project—so these issues have wide applicability. The key feature of these reforms, from the point of view of this article, is that they are defined in terms of a particular model of school or classroom practice but, due to their size and scope, they are often implemented with variation from site to site as the models are interpreted by local agents and adapted to local conditions. The obstacles researchers face when evaluating such programs include variations in the implementation of the reform combined with a lack of information about actual changes in practice; variation in outcome measures available across sites; and a lack of a straightforward set of analytic methods applicable to replicated cases. This article uses a study of the NSF Systemic Initiatives (SI) to illustrate methods that can be used to address these obstacles.

This research was supported by a grant from the National Science Foundation, Division of Elementary, Secondary, and Informal Education. We are grateful to the editor and to three anonymous reviewers for comments that improved this article.

The first problem in studying large-scale reforms is that the program may differ from site to site in ways that are not explicit in the formal program model and that are not measured well. This occurs because large-scale reforms are often implemented in a top-down manner, with guidance coming from above but with significant local flexibility. In such cases, the local context is likely to affect the implementation in unknown ways. This is not necessarily problematic. The process of mutual adaptation has been found to foster implementation (Berman & McLaughlin, 1978). However, local flexibility means that a single reform may have multiple instantiations, making it difficult for evaluators to identify a single program "effect." Two sites may structure their teacher training programs differently, imparting slightly different visions of key reform components. Even within a single reform site there is likely to be significant variation in how the program elements are adopted in individual classrooms. Furthermore, few programs collect any information about classroom practice, making it impossible to determine the extent of variation that exists. Even though all teachers in a school participate in the same training program, they do not necessarily provide students with the same subsequent instruction.

Despite this variation in implementation, judgments about the overall *effectiveness* of a program can still be made. For example, one can use changes in student outcomes that occur after program initiation as an indicator of overall impact. While such global judgments provide a "bottom line" estimate of impact, they offer little or no information about the causes of the effect, that is they do not provide any indication of the degree of implementation or the impact of particular program elements. As such, judgments of effectiveness are likely to be less useful for program improvement than analyses that link outcomes to information about the adoption of the reform model of practice. To determine whether outcomes are associated with the desired changes in practice, one needs to measure implementation and practice directly and associate them with changes in student performance. In combination, this information tests the *efficacy* of the reform; that is, whether the fundamental model of practice that guides the reform actually works. It is important to know both whether teaching in a particular manner improves performance and whether the program is fostering these behaviors among teachers.

The second problem faced by evaluators of large-scale instructional reforms is the lack of common, appropriate measures of student achievement outcomes. In many cases, the achievement measures available to evaluators consist of state- or district-administered standardized tests (typically multiple-choice). The specific tests used vary across districts and states, making it difficult to combine information from different reform sites. In addition, existing measures may be insensitive to important outcomes of the reform. For example, many standards-based reforms emphasize skills such as writing, which may not be measured well by existing multiple-choice achievement tests. Open-response and performance assessments, which may provide better information about such skills, are rarely available.

Third, standard analytic methods are not always applicable to multisite projects in which both site-level implementation factors and student outcome measures may differ. Standard multiple regression methods, including hierarchical linear models, are useful for controlling for pre-existing differences between treatment and comparison groups, both at the individual and aggregate levels, but they do not work as well for programs that are essentially replicated case studies with local variation in implementation and outcome measures.

SI programs provide examples of large-scale reforms that seek to influence instructional practices according to a common set of standards while permitting local variation in design and implementation. States and districts received SI funds to implement math and science reforms that promote instruction consistent with national standards. Although the SI programs were intended to address the entire educational system, their effects on student achievement depend most directly on changes in instruction in the classroom. Substantial resources have been invested in the programs, but the link between student achievement and reformed practice at the classroom level has not been tested.

Our approach to studying the SI programs had three major features that address the challenges identified earlier: (a) We collected information on practices at the classroom level rather than simply comparing participating and nonparticipating schools; (b) we used existing student achievement data but supplemented them with additional

measures wherever possible in an effort to have multiple measures of student achievement; and (c) we designed the study as a planned meta-analysis in 11 distinct sites, with similar but not identical data from each site. While this study is innovative in many ways, it builds on a growing body of literature that explores relationships between teacher-reported instructional practices and student achievement. More importantly, the study suggests a number of limitations of methods that are currently used for evaluating large-scale reforms that promote instructional change, and provides some lessons for future evaluations of such efforts.

The article begins with a brief description of the SI programs. We provide background information about the initiatives, discuss prior research findings, and describe the specific challenges faced in evaluating the SI programs. We then describe our approach to studying the reforms, including samples, measures, and methods of analysis, followed by the results. The article concludes with a discussion of the strengths and limitations of our approach and directions for future research and evaluation.

Systemic Reform Initiatives

NSF's Systemic Initiatives included the Statewide Systemic Initiatives (SSI), Urban Systemic Initiatives (USI), Rural Systemic Initiatives (RSI), and Local Systemic Change (LSC) programs. The initiatives were intended to provide resources to promote system-wide change, and many sites were funded at a level of several million dollars over multiple years. The USI program, for example, funded 20 large urban districts with awards of up to $15 million each for a five-year period, focusing on cities where high proportions of children live in poverty. The program is described as a "comprehensive and systemic effort to stimulate fundamental, sweeping, and sustained improvement in the quality and level of K–12 science, mathematics, and technology (SMT) education" (Williams, 1998, p. 7). Taken together, these SI programs received approximately $100 million per year in NSF funding during the 1990s. In addition, most sites supplemented their NSF grants with additional local contributions. For example, sites have used Title I funds, corporate donations, and grants from private foundations to support and expand their systemic initiatives (Williams, 1998).

A cornerstone of the systemic reform initiatives is the alignment among all parts of the educational system, including curriculum, instruction, assessment, teacher preparation, and state and local policies such as graduation requirements. Such alignment is perceived as necessary for promoting change in the classroom and, ultimately, improving student performance (Smith & O'Day, 1991). Systemic reform efforts have resulted in part from observations that addressing one component of the educational system tended to be ineffective due to constraints imposed by other parts of the system (Hill, 1995; Knapp, 1997).

Furthermore, these initiatives involve the adoption of rigorous curriculum and performance standards and the mobilization of all components of the system to support and enable all students to reach those standards (Consortium for Policy Research in Education, 1995). Typically, sites adopted a set of standards consistent with national mathematics and science standards documents, purchased curriculum materials consistent with the standards (many of which were developed with support from NSF), and provided extensive in-service training for teachers to use the materials and teach in a particular manner. As specified in a recent program announcement, a primary objective of these programs is "to improve and/or advance . . . implementation of a standards-based, inquiry-centered science, mathematics, and technology education for all students K-12" (National Science Foundation, 2001b, p. 9). The terms "standards-based" and "inquiry-centered" refer to practices that are consistent with curriculum standards and guidelines published by the National Research Council (1996), the American Association for the Advancement of Science (1993), and the National Council of Teachers of Mathematics (2000).

Common to all of these documents is an emphasis on instruction that engages students as active participants in their own learning and that enhances the development of complex cognitive skills and processes. Specific practices that are often associated with this approach include cooperative learning groups, inquiry-based activities, use of materials and manipulatives, and open-ended assessment techniques. All of these practices are intended to support active rather than passive learning, to promote the application of critical thinking skills, and to provide opportunities to apply math and science learning to real-world contexts. For the remainder of this article

we refer to this set of practices as "standards-based" or "reform-based" instruction. Readers should keep in mind that the use of a standards-based approach to instruction does not preclude the use of more "traditional" activities such as lecture-style teaching, as well.

Teachers must adopt these reforms and ensure they take hold in the classroom in order for them to be effective (Tyack & Cuban, 1995). Thus, a primary emphasis of the systemic reform initiatives involves promotion of teaching practices that are assumed to facilitate student learning. Most initiatives offer professional development to teachers, and this component constitutes a fairly large proportion of the budget. For example, the SSI sites spent nearly one third of their first-year budgets on in-service training for teachers, more than on any other category of spending (Shields, Corcoran, & Zucker, 1994). The goal of most of this training is to increase teachers' use of classroom practices that are believed to improve achievement.

Earlier Evaluations of the SI Programs

Individual SI sites worked with outside organizations to evaluate their efforts, and NSF has commissioned additional external evaluations. There have also been several studies of relationships between student achievement and the kinds of instructional practices that are promoted by the SI program.

Evidence about overall effectiveness

To date, most of the research on the SI programs has focused either on implementation (e.g., type and frequency of professional development offered to teachers, level of participation among teachers) or on overall impact (e.g., average improvement in student test scores for participating sites). A large-scale study of SSI programs conducted by SRI International revealed small but statistically significant differences in test scores that favored participating over nonparticipating schools in four of seven sites (Laguarda, 1998). More recently, NSF published a report showing large test-score gains among students participating in the Urban Systemic Initiatives program (National Science Foundation, 2001a).

Evidence about efficacy of reform-based teaching practices

If the systemic initiatives do result in improved student achievement, it is undoubtedly due in large part to what occurs in the classroom. Witness the large roles that professional development and the promotion of standards-based instructional practices play in the SI initiatives. Research provides some evidence of the effectiveness of some of the individual practices that have been advocated by supporters of standards-based instruction. An experiment conducted by Ginsburg-Block and Fantuzzo (1998), for example, showed that low-achieving elementary students who were assigned to problem solving or peer collaboration conditions obtained higher math scores and reported higher levels of motivation than did students who received neither of these interventions. Several other studies have demonstrated the value of peer tutoring and collaboration (e.g., Fantuzzo, King, & Heller, 1992; Greenwood, Carta, & Hall, 1988; Webb & Palincsar, 1996), as well as the benefits of contextualizing instruction in real-world problems (Verschaffel & De Corte, 1997).

Other studies have focused on relationships between student achievement and teachers' use of combinations of these practices. For example, Cohen and Hill (2000) studied teacher-reported use of several practices consistent with the 1992 California Mathematics Framework and found that teacher-reported frequency of use was positively related to scores on the California Learning Assessment System (CLAS) mathematics test at the school level, after controlling for demographic characteristics. Furthermore, this research revealed that the type of professional development offered to teachers influenced the degree to which teachers reported adopting reform-based practices. Mayer (1998) found small positive or null relationships between teacher reports of reform-based practices and student scores on a standardized multiple-choice test. Von Secker (2002) used questionnaire data from the National Education Longitudinal Study of 1988 (NELS:88) to examine relationships between inquiry-based practices and achievement and found that teacher-reported use of these practices was related to higher student achievement. He also found that these practices contributed to greater inequalities in achievement between socio-economically advantaged and less advantaged students. Smerdon, Burkam, and Lee (1999) used NELS:88 teacher questionnaire data to examine "didactic" and "constructivist" practices. They found that teachers' reported use of both types of practices varied by socioeconomic status and course type and that reported hands-on learn-

ing opportunities were positively related to student achievement in science (Burkam, Lee, & Smerdon, 1997). Wenglinsky (2002), using National Assessment of Educational Progress (NAEP) data, found that teacher-reported emphasis on higher order thinking skills was associated with positive student achievement outcomes.

Studies that have used observational methods have obtained similar results. For example, Stein and Lane (1996) found positive relationships between student performance in mathematics and their exposure to instructional tasks that required complex thinking. Thus, there is some evidence that teachers' use of reform-based practices is related to higher student achievement.

There is also a related body of evidence (Briars, 2001; Schoenfeld, 2002) that implementation of inquiry-based curricula, most of which incorporate the style of practice we have been discussing (e.g., by providing extended science or mathematics investigations) is related to improved achievement. Much of the research on practices and curriculum emphasizes the importance of sustained professional development that is designed to support the adoption of the curriculum or practices (Cohen & Hill, 2000; Garet, Porter, Desimone, Birman, & Yoon, 2001).

Challenges for evaluators

The previous research suggests that reform-based instructional practices may be a promising approach to improving student achievement in mathematics and science, and that to understand the effectiveness of the SIs, it is critical to examine practices as well as outcomes. Doing so presents exactly the challenges we highlighted earlier.

First, although all SI sites are expected to provide professional development to help teachers adopt practices that are consistent with published standards and guidelines, there is extensive variation among sites in the nature and amount of professional development, in the specific instructional approaches that are adopted, and in the curriculum materials that are adopted.

Moreover, within-site and even within-school variation in classroom practice is likely: Other researchers have found that teachers' use of reform practices is influenced by a number of factors, including the nature and frequency of professional development participation (Cohen & Hill, 2000; Weiss, Montgomery, Ridgway, & Bond, 1998) and the degree to which they understand the subject matter (Cohen & Ball, 1990). Without studying these variations in practice it is impossible to know whether students who are exposed to reform-minded teachers achieve greater gains than those whose teachers have resisted the reforms. Thus attention to variation in practice should be an important part of any effort to understand the effectiveness of the SI programs. To date, however, there has been little information available to researchers who wish to understand what teachers do in SI classrooms.

The second difficulty in conducting evaluations of multisite programs in general, and of the systemic initiatives in particular, is a lack of common and appropriate measures of student achievement. Each state that implements an SI has a different testing program that has served as the primary outcome measure for sites in that state. Few SI sites supplemented the existing tests with additional measures of achievement. In addition, as discussed earlier, most state testing programs rely heavily or exclusively on multiple-choice items (Education Week, 2001), a format that may not always lend itself to measuring many of the scientific inquiry and mathematical problem solving skills encouraged by the systemic initiatives. Science reforms are particularly difficult to evaluate because not all states administer science assessments, and those that do often limit them to a few grade levels (Goertz & Duffy, 2001).

State testing programs also typically fail to provide data that can be used to track the progress of individual students over time. Many states, for example, test students only in a few grades (e.g., 4th, 8th, and 10th; see Goertz & Duffy, 2001)[2], particularly in science. This forces evaluators to focus on changes in scores of successive cohorts of students, which confound the effects of reforms with differences among the groups of students. In addition, improvements in scores over time, which are often cited as evidence of beneficial effects of reforms on student learning, may in many cases reflect inappropriate narrowing of the curriculum or teaching to the test (Koretz & Barron, 1998; Linn, 2000). This problem is especially likely when the tests are part of a high-stakes accountability system or when the same form of a test is administered multiple times. For all of these reasons, it is desirable to supplement existing state tests with additional measures whenever possible.

Finally, the SI programs were implemented in dozens of locations across the country. Each state,

district, or consortium should be treated as a separate site, because each constituted its program in somewhat different ways. Rather than simply pooling the data from the separate sites, one should adopt an approach that maintains the integrity of the site-specific relationships but also permits aggregation of results where appropriate.

The study reported here addressed these challenges—some better than others. We gathered data from 11 SI sites, including some states and some districts. We measured classroom practices at the level of the individual teacher, which allowed us to address variations in reported practice both within and across schools. In addition, we measured student achievement using both multiple-choice and open-response tests, including some hands-on science tasks that we developed and administered ourselves. Finally, we used information on student demographics and prior achievement as part of a planned meta-analysis to control for pre-existing differences among students. An earlier report (Klein et al. 2000) presents results for the first six sites, and provides additional details on methodology.

It is important to clarify that our study was not an evaluation of the systemic reform initiatives, *per se,* but an investigation of relationships between achievement and teachers' use of practices consistent with the initiatives. The effectiveness of the SI program is the subject of a comprehensive evaluation undertaken by SRI International (e.g., Corcoran, Shields, & Zucker, 1998; Shields, Marsh, & Adelman, 1998).

Methods

We collected data from 11 sites—six during the spring of 1997 and six during the spring of 1998 (one site participated both years). Our specific procedures for site selection, subject and grade-level selection, and data collection are described in the following sections.

Site, School, Subject, and Grade-Level Selection

Because we knew that it would be difficult to study the relationship between reform-based instructional practices and achievement in the absence of a reasonable degree of reform, we selected sites in a way that maximized the probability of encountering substantial numbers of teachers using reform practices. NSF staff proposed sites in which there were indications that large numbers of teachers had adopted the reform practices in their classrooms. NSF drew its recommendations from site visits and from progress reports submitted by the grant recipients.

The selected sites included SSIs, USIs, and one LSC. Most were involved in both mathematics and science reforms, but a few focused on only one of these subjects. The number of years in which sites had been involved in the SI varied, but no site was in its first year of participation. Within sites, the amounts of time teachers had been using the reform practices varied due to mobility, teachers' own preferences, and other factors. Because we were not trying to judge the impact of the SI reform per se, but to measure the relationship between reported teaching practices and achievement during a single school year, we could include data collected in different years and from sites and teachers with different amounts of SI experience.

All of the sites had been involved in the reform for more than one year, but had yet to implement the reforms in all schools in the site. The same basic research design was used at each site. School district and program personnel at each site specified the grade level(s) and subject(s) in which they believed reform practices were most pervasive, and then nominated schools to participate in the study. We asked local staff to select approximately ten schools in which there was good reason to believe mathematics and/or science reforms had been implemented, and ten demographically similar schools in which the reforms had yet to be implemented. We used the nominations only to ensure variation in teaching practices; we did not compare the high- and low-implementing schools with one another directly except for exploratory analytic purposes. Table 1 lists the grade(s) and subject(s) of data collection and the numbers of teachers and students participating at each site. It also indicates the subjects that the site focused on and whether the site was an SSI, USI, or LSC. Between 85% and 100% of selected schools participated in the study, and within sites, teacher survey completion rates ranged from 71% to 100%, with most sites achieving close to 100% participation.

Student achievement data

We obtained student scores on the mathematics and science assessments regularly administered at each site, and supplemented these with additional assessments where feasible, to provide

(text continues on page 9)

TABLE 1
Sites, Grades, Subjects, Number of Participants, and Assessments

Site	Grade	SI Type	Subject in SI	Subject in study	Number of teachers	Number of students	Tests	Added for present study?	Prior year test scores available?
1	3[a]	SSI	Math	Math	46	804	State multiple-choice math	No	
			Science				State open-response math	No	No
2	5	USI	Math	Math	100	1651–1686	State multiple-choice math	No	
							Commercial open-ended math	Yes	
			Science	Science	99	1639–1662	Commercial multiple-choice science	Yes	Yes
							Hands-on science[b]	Yes	
3	5	USI	Math	Math	73	1366–1451	Commercial multiple-choice math	No	
					74	1367–1438	Commercial open-ended math	Yes	
			Science	Science			Standards-based multiple-choice science[c]	No	Yes
							Standards-based open-ended science[c]	No	
4	5	USI		Science	45	909–932	Standards-based multiple-choice science[c]	No	Yes
							Standards-based open-ended science[c]	No	
5	7	SSI	Math	Math	48	2937–3018	State multiple-choice math	No	
							Commercial open-ended math	Yes	
			Science	Science	33	2047–2079	Commercial multiple-choice science	Yes	No
							Hands-on science[b]	Yes	
6	7	USI	Math	Math	57	3237	Commercial multiple-choice math[d]	No	
			Science	Science	52	3279	Commercial multiple-choice science[d]	No	
7	7	LSC	Math	Math	57	3127–3145	Commercial multiple-choice math	No	
							Commercial open-ended math	Yes	Yes
			Science	Science	52	2870	Commercial multiple-choice science	No	
8	5	SSI	Math	Science	37	1637–1641	Commercial multiple-choice science	No	Yes
			Science				State open-ended science	No	
9	4	USI	Math	Science	116	1783–1786	Commercial multiple-choice science	Yes	No
			Science				Commercial open-ended science	Yes	

(continued on next page)

TABLE 1 *(Continued)*

Site	Grade	SI Type	Subject in SI	Subject in study	Number of teachers	Number of students	Tests	Added for present study?	Prior year test scores available?
10	4	SSI	Math	Math	76	1244–1248	State multiple-choice math	No	No
							State open-ended math	No	
			Science	Science	76	1265–1270	State multiple-choice science	No	
							State open-ended science	No	
11	8	SSI	Math	Math	28	1163	State multiple-choice math	No	No
							State open-ended math	No	
			Science	Science	18	1033	State multiple-choice science	No	
							State open-ended science	No	
12	5	USI	Math	Math	67	1507–1592	Commercial multiple-choice math	No	Yes
							State multiple-choice math	No	
							State open-response math	No	

Note.

[a] At this site, we studied teaching practices for 3rd-grade teachers and measured the relationships with student test scores that were gathered during the following fall when students had advanced to the 4th grade.

[b] See (*reference deleted for anonymity*) for a description of tasks and scoring guides.

[c] This test was developed by a consortium of educators and researchers, and was designed to be aligned with NSF-supported reform efforts.

[d] In this site, we were unable to schedule any open-ended testing.

SSI = Statewide Systemic Initiatives, USI = Urban Systemic Initiatives, Local Systemic Change.

both multiple-choice and open-response scores. Supplementary tests were chosen in consultation with local staff, who were encouraged to select measures that they believe were reasonably well aligned with their reform efforts. Hands-on science tasks that were developed and extensively field-tested by (*RAND*) were made available, and some sites opted to use them. Mosaic project staff trained exercise administrators to administer some of the supplementary measures, including the hands-on tasks. All other tests were administered by the classroom teachers or by test administrators who worked at the local sites. Table 1 indicates the types of tests administered in each site.

In all but one site, students completed a standardized multiple-choice assessment in mathematics and/or science depending on the site designation, and an open-response test that required students to produce, rather than select, their responses. One site administered only multiple-choice tests, and we were unable to schedule additional testing due to time constraints. We used existing tests wherever possible, including state-developed tests and commercially available standardized tests. The column "Added for present study?" in Table 1 indicates whether we supplemented the district or state's testing program with additional measures or relied only on those measures already used by the sites.

It is important to acknowledge the relative strengths and weaknesses of the two types of measures we used. The multiple-choice tests tended to have good technical quality (e.g., score reliability) and were typically tests that had been widely administered. However, many of the SI personnel believed that these tests failed to capture some important aspects of the instruction to which students were exposed. On the other hand, the open-response tests, which many SI participants believed were more closely aligned with the reforms, tended to have lower degrees of reliability than the multiple-choice tests, and many of them had not been used widely prior to this study. By using both types of measures we hope to offset some of the problems, but they need to be kept in mind when interpreting our results.

To control for pre-existing differences in student achievement, we obtained district or state test scores in the relevant subject from the spring prior to our data collection. In most sites, prior year test scores were missing for 5 to 10 percent of the student sample. For each of these sites, we imputed multiple values for each prior year test score using Bayesian models for multivariate clustered data, as described in Schafer (1997). We used the PAN software for S-plus to fit the models and draw imputed values (Schafer, 1998). The imputation models predict the missing test scores as a function of the student demographic and socioeconomic (SES) variables and the instructional practice scales. The imputation models also include the current year test scores as a predictor for the missing values of the prior year scores. For each multiple imputation we created ten replicate data sets. Each replicate contained observations for every student in the sample, and each of the ten data sets contained an imputed value for every value missing from the original data. For each observation with a missing prior-year test score, the imputed values varied across the ten replicate data sets. To obtain the final estimates reported in the results section below, we repeated our analyses on each of the ten replicated data sets and then pooled the ten sets of estimates using the methods described in Rubin (1987). The multiple imputation method allowed us to use all the observed data with standard software packages, while providing statistical inferences that account for imprecision created by the missing values by adjusting statistical tests and confidence intervals to reflect between-imputation variability in the estimates. The models also accounted for the hierarchical structure of the data with students nested within classrooms.[3]

In five sites we were unable to obtain prior year test scores because the state or district did not administer tests in the relevant grade or did not maintain individual student records. In these cases we used contemporaneous reading and language scores (i.e., scores on a reading test that was administered at approximately the same time as the tests we used as outcome measures) as covariates. Both prior year and contemporaneous scores serve as measures of student achievement. Unlike prior year scores, however, contemporaneous scores are not necessarily independent of the instructional practices measured by our surveys: If instruction in math or science involves activities that promote the use of verbal skills, for example, this instruction could improve reading or language scores. Including contemporaneous scores as covariates could absorb some of the effects of instruction and result in an incorrectly estimated relationship between practices and achievement

in math or science. The alternative approach, which would involve excluding an achievement covariate altogether, is also problematic. We conducted sensitivity analyses and determined that including contemporaneous scores was the most appropriate approach, resulting in a very slight attenuation of the relationship between reform practices and achievement [for details on these analyses see Klein et al., 2000]. The last column in Table 1 indicates for which sites we had prior year test scores.

Teacher questionnaires

Our primary measure of teaching practices in each site was a questionnaire developed by Horizon Research, Inc. (HRI). This instrument is a modified version of a questionnaire that HRI developed in collaboration with SI staff. It has been validated and used extensively by HRI to evaluate the implementation of the LSC initiatives. Questionnaires were administered to all teachers in a school teaching the targeted subject and grade level. Typically, the site coordinator or assistant distributed the questionnaires either individually or at after-school meetings and then collected completed questionnaires in individual, sealed envelopes for return to us.

The surveys reflect the common beliefs about mathematics and science instruction that characterized the systemic reform initiatives. Although NSF did not mandate a particular curriculum or a specific set of teaching strategies for the Systemic Initiatives, there was an emerging consensus among math and science educators about what should be taught and how it should be presented. (National Council of Teachers of Mathematics, 2000; National Research Council, 1996). In light of this consensus, it is not surprising that the systemic reform programs adopted very similar content and instructional goals. An independent evaluation of the SSI program reported that, "across the states there was remarkable similarity in the perceived shortcomings of current practices and the set of desirable reforms in curriculum content and instructional strategies." (Shields, Marsh, & Adelman, 1998; p. 2). The shared content goals included greater emphasis on conceptual understanding of math and science concepts, the application of this knowledge to everyday situations, and the integration of concepts across subjects.

The instructional emphasis was equally distinct. Rather than viewing students as passive learners who absorb unrelated facts and procedures, the reforms sought to engage students actively in their own learning, to be sensitive to each student's learning style, to increase the use of technology, and to utilize new forms of assessment for instructional planning. In mathematics this meant more "data gathering and analysis, statistics, geometry and visualization, discovery learning and 'constructivist' approaches;" in science more "scientific processes, such as observation, comparison, experimentation, hypothesis generation, hypothesis-testing, and theory building" (New Jersey SSI Proposal, p. 7; quoted in Shields, Marsh & Adelman, 1998, p. 3).

We created separate questionnaires for mathematics and science teachers, but many of the items were identical across subjects. Questions asked teachers to report the frequency of various instructional practices and classroom activities (e.g., "record, represent, and/or analyze data," "write a description of a plan, procedure, or problem-solving process"). General topics included: amount of time spent on science/mathematics; approach to introducing a new topic; typical teacher instructional practices; typical student activities; types of written assignments; and methods of assessing student learning.

In addition, teachers completed a brief demographic section, providing information about their college degree, teaching certification, coursework in mathematics and/or science, gender, ethnicity, and years of teaching experience. In sites where science or mathematics specialists delivered instruction instead of the regular classroom teacher, we administered surveys to the specialists and also asked the respondent to clarify the teaching situation.

Student demographic data

Finally, we obtained student-level demographic data, which in most sites included race/ethnicity, gender, participation in free or reduced-price lunch programs, language background, and participation in special education or gifted programs. These data were used to verify that comparison schools were similar to implementing schools on student demographics, enrollment, and grade span, and were included as covariates in the analysis of relationships between teaching practices and student achievement. These data also enabled us to study whether these relationships varied as a function of student characteristics.

The inability to link data from students and teachers is a factor that hinders many evaluations of instructional programs. Few district or state data systems maintain these links in a readily usable form, especially at the secondary grades when students typically have a different teacher for each subject. Most of our sites lacked these links, so we collected class rosters from teachers and used these rosters to make the links.

Analysis

The primary purpose of this study was to investigate the degree to which student achievement was associated with teachers' reported use of reform-based instructional practices. We developed linear regression models to estimate the relationship between test scores and teacher-reported instructional practices. The models used individual student test scores and background characteristics, with all students in a class receiving the same value for the instructional practice scales. We used ordinary least squares to fit the model and post hoc adjustments to the standard errors to account for possible correlation among scores from students within the same class and for possible heteroscedasticity of scores (McCaffrey, Bell, & Botts, 2001). This post hoc adjustment is a nonparametric estimator of the standard error that is robust to assumptions about the correlation among scores for students from the same classroom. Standard errors from alternative methods such as hierarchical linear modeling are dependent on assumptions about the correlation among scores for students from the same class and can result in inaccurate inference if those assumptions are incorrect. Therefore hierarchical models require careful exploration of the correlation among student scores. Because the focus of the study was on the relationship between teaching practices and student test scores and not the correlation among scores from students within a classroom, we used the nonparametric standard error estimates. As noted above, because we used the multiple imputation procedure described by Rubin (1987), the standard errors also account for the increased variability in the estimates that results from the fact that some prior year test scores were missing. All statistical tests were conducted using approximate Wald T- and F-tests based on the adjusted standard errors.

At each site, we conducted separate analyses for mathematics and science and for open-response and multiple-choice tests. We estimated models with and without the teacher background variables (e.g., teacher education level) but here we report only the models that exclude those variables because they did not provide any additional explanatory power. The specific student background information available for inclusion in the models varied by site. For each site we conducted exploratory analysis to select which of these predictors to include in each model. We included predictors with significant bivariate relationships with scores and used backward deletion to select the final model in sites with many available covariates. We also used graphical methods, residual and added variable plots (Neder, Kutner, & Nachtsheim, 1996) to check model fit, and in particular to check for nonlinearities in the relationship between instructional practices and test scores.

The use of data from multiple sites provides an opportunity to conduct a planned meta-analysis. We therefore also conducted pooled analyses that combined data from all six sites to produce a single estimate of the coefficient relating teaching practices to student achievement. This approach provides results that are similar to what would be obtained by pooling individual scores and fitting a random coefficients model with interactions between sites and the covariates, but it permits the pooling of data across sites without requiring identical models in every site (Goldstein, 1995). We conducted separate analyses by subject (math or science) and test format (multiple-choice or open-response).

Results

We first present summaries of teachers' reported use of instructional practices. We then present our findings with regard to the relationships between reported use of these practices and student achievement for each site, followed by a discussion of a cross-site pooled analysis. Finally, we describe results of an analysis of differences between open-response and multiple-choice achievement measures.

Distributions of teaching practices

We conducted exploratory factor analyses of the questionnaire items in each site. Although the specific factor loadings varied across sites, we consistently found that items clustered into two factors that represented similar constructs across grade levels and subjects. We used these findings

to create two scales for each site. The first scale consisted of 22 items that reflected "reform practices." Some of these items described teacher behaviors, such as "arrange seating to facilitate student discussion," "use open-ended questions," and "require students to explain their reasoning when giving an answer." Other items described student behaviors, such as "work in cooperative learning groups," "make formal presentations to the class," or "work on solving a real-world problem." We also created a 5-item "traditional practices" scale based on items that measured the amount of time teachers or students spent in traditional activities (such as textbook work, lectures, and short-answer tests). These scales match quite closely those used in the LSC evaluation (Weiss, Montgomery, Ridgeway, & Bond, 1998). Furthermore, this distinction between reform-related practices and more traditional practices is consistent with the kinds of definitions used in other research on math and science reform (e.g., Cohen & Hill, 2000; Smerdon, Burkam, & Lee, 1999). The score for each teacher was simply the average item response across items. All items used a 5-point Likert scale, so teachers' scores could range from 1 *rarely or never using any of the practices* to 5 *engaging in all practices daily or almost daily.* Across sites and subjects, the average alpha coefficient was 0.92 for reform practices and 0.70 for traditional practices.[4] The appendix lists the items comprising the reform and traditional practice scales for each subject.

It is important to note that the two scales are not opposites of one another. Correlations between the two scales ranged across sites from moderately negative to moderately positive, with many close to zero. It is possible for teachers to receive high scores on both scales because the scale scores do not indicate the total amount of time spent on these practices, but rather the frequency with which they are used. Thus a teacher who intersperses lecture-style teaching with opportunities for student discussion in every lesson might score high on both scales. In addition, there are other activities that are not addressed by either scale, so it is possible for teachers to receive low scores on both. Finally, as we discussed earlier, the adoption of a reform-based approach to instruction does not preclude the use of activities such as lecture and worksheets.

In each site we found a wide range of scores on both the reform and the traditional scales. Table 2 provides descriptive information for each combination of site and subject (math or science). There are 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. Inspection of the distributions of scores suggests that range restriction is not a problem in any of our sites.

Although it is not shown in this table, we observed large variability within schools, regardless of whether they were originally classified by site staff as high- or low-implementing. High-implementing schools were likely to include at least one teacher who reported infrequent use of reform practices, and low-implementing schools often had teachers who reported using reform practices liberally. This underscores the importance of linking student outcomes directly to his or her teacher rather than to a school-wide average. We discuss this issue further in the final section.

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 the reform and traditional scales for each of the four subject-by-test-format combinations (math multiple-choice, math open-response, science multiple-choice, and science open-response). Below we provide detailed results only for the reform practices scale, partly because it is more directly relevant to the questions that motivated this study, but also because the five-item traditional scale did not always function well in our analyses: There were significant nonlinearities for several sites, making it difficult to pool results across sites, and the standard errors for the traditional practices coefficients tended to be large, due in part to the relatively low reliability of the 5-item scale.[5] When we pooled results across models for which a linear term was appropriate, in none of the four pooled-analyses was the coefficient for traditional practices significantly different from zero.

Figures 1 through 4 provide an overview of our reform practices analyses in each site, as well as the pooled results across sites.[6] The relationships

TABLE 2
Descriptive Statistics on Teaching Practices Scales, by Site

Site	Subject	Scale	*M*	*SD*	Minimum	Maximum
1	Math	Reform	3.20	.56	1.82	4.50
2	Math	Reform	3.61	.58	2.05	4.64
3	Math	Reform	3.38	.56	1.68	4.68
5	Math	Reform	3.01	.59	2.00	4.64
6	Math	Reform	3.34	.59	1.50	4.73
7	Math	Reform	3.32	.54	2.00	4.50
10	Math	Reform	3.79	.83	1.60	4.90
11	Math	Reform	3.25	.97	1.60	4.60
12	Math	Reform	3.63	.45	2.76	4.88
1	Math	Traditional	3.73	.66	2.00	5.00
2	Math	Traditional	3.63	.66	2.00	4.80
3	Math	Traditional	3.33	.57	2.20	5.00
5	Math	Traditional	3.40	.65	1.80	4.80
6	Math	Traditional	3.73	.63	2.20	5.00
7	Math	Traditional	3.80	.51	2.60	5.00
10	Math	Traditional	3.25	1.08	1.00	5.00
11	Math	Traditional	3.07	1.06	1.40	4.60
12	Math	Traditional	4.08	.56	2.40	5.00
2	Science	Reform	3.27	.68	1.00	4.55
3	Science	Reform	3.33	.64	1.64	4.41
4	Science	Reform	3.57	.53	1.95	4.36
5	Science	Reform	3.22	.69	1.64	4.53
6	Science	Reform	3.28	.69	1.45	5.00
7	Science	Reform	3.31	.58	1.55	4.27
8	Science	Reform	3.44	.48	2.56	4.81
9	Science	Reform	3.00	.58	1.00	4.19
10	Science	Reform	3.66	.85	1.80	4.90
11	Science	Reform	3.33	.77	1.80	4.60
2	Science	Traditional	3.34	.75	1.00	5.00
3	Science	Traditional	2.86	.60	2.20	4.40
4	Science	Traditional	2.65	.70	1.20	4.00
5	Science	Traditional	3.78	.67	2.60	4.90
6	Science	Traditional	3.62	.59	2.00	5.00
7	Science	Traditional	3.66	.58	2.20	4.80
8	Science	Traditional	3.31	.63	1.60	4.40
9	Science	Traditional	2.43	.54	1.00	4.00
10	Science	Traditional	2.72	1.02	1.00	5.00
11	Science	Traditional	3.26	.79	2.20	4.60

Note. All scores are averages across items on the 5-point scale described in the text of the article.

depicted in the figures are the estimated coefficients from our regression models for the reform practices scale. We report standardized coefficients, which represent the expected difference in test score standard deviation units for a one standard deviation unit increase in scores on the reform scale. The dark dot represents the point estimate for the coefficient and the gray bar represents 95% confidence interval for that point estimate. The bottom bar in each figure shows the estimated coefficient from the pooled analysis, described later.

Figure 1, which shows relationships between teacher-reported use of reform practices and achievement on open-response math tests, indicates that in seven of the eight sites where we had open-response mathematics tests, higher test scores were associated with greater reported use

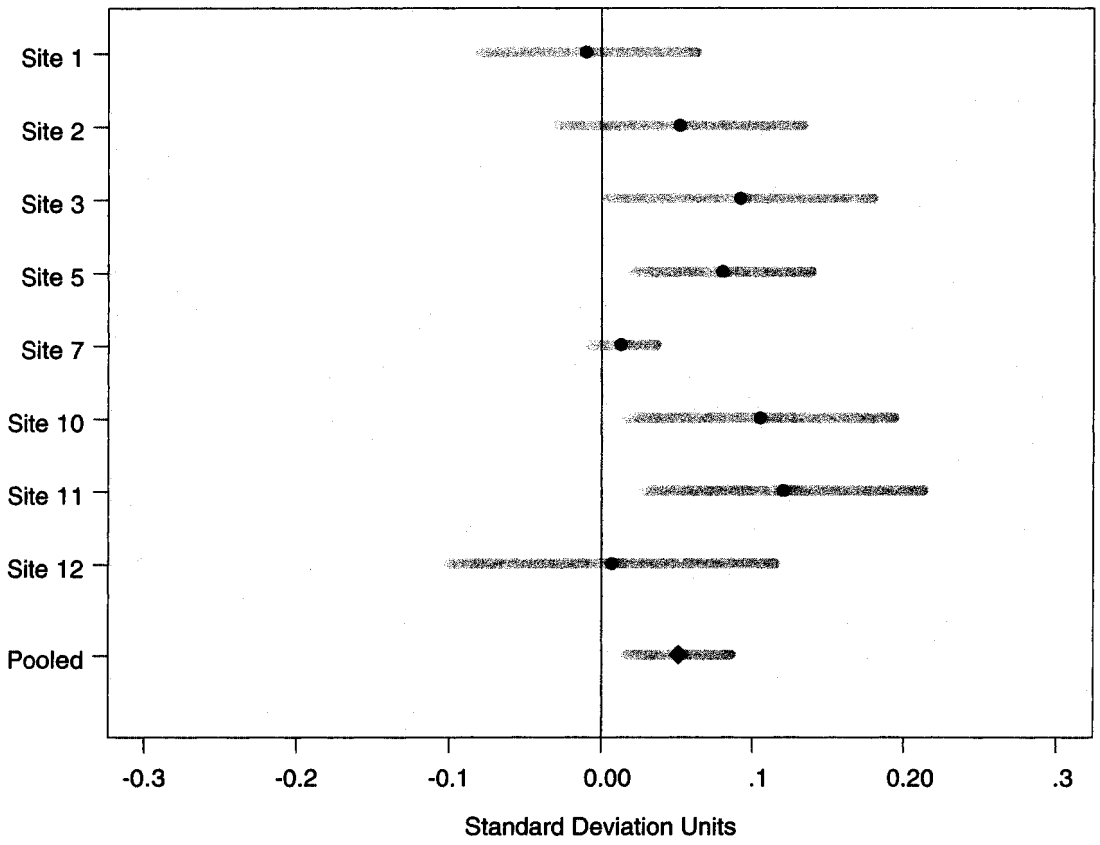

FIGURE 1. *Coefficients for reform practices, open-response math.*

of reform practices (i.e., the estimated coefficient was positive). However, as indicated by the confidence intervals, the coefficients were statistically significantly greater than zero in only four of these sites. Similarly, Figure 2 shows that for almost all of the participating sites, higher multiple-choice test scores in mathematics were associated with greater reported use of reform practices, although only one of the estimates was statistically significantly different from zero.

Figures 3 and 4 show that greater reported use of reform practices in science was associated with higher test scores on both open-response and multiple-choice measures in science. Again most of the estimated coefficients were extremely small and were not statistically significantly different from zero, even though an inspection of coefficients across sites shows a consistent pattern of a weak positive relationship between the reform practice scale and test scores.

As shown in Figures 1 through 4, the relationship between reported use of reform practices and test scores is at most small in almost all our models. For example, one of the larger coefficients was 0.09, an estimate of the relationship between reform practices and open-response science tests in Site 2 (see Figure 3). In this site, our model suggests that for a teacher who reported using all of the reform practices monthly, the average student was predicted to score at about the 48th percentile in the site on the test, while for a teacher who reported using all of the reform practices weekly we would predict that a similar student would score at about the 54th percentile.[7] Smaller changes in percentiles would be expected in most of the other sites. Compared with the coefficients for most of the student background characteristics (e.g., an average coefficient of approximately 0.5 across sites for participation in free- or reduced-price lunch programs), all of the relationships we observed may be considered small.

Open-response measures are often perceived as more appropriate indicators of student achieve-

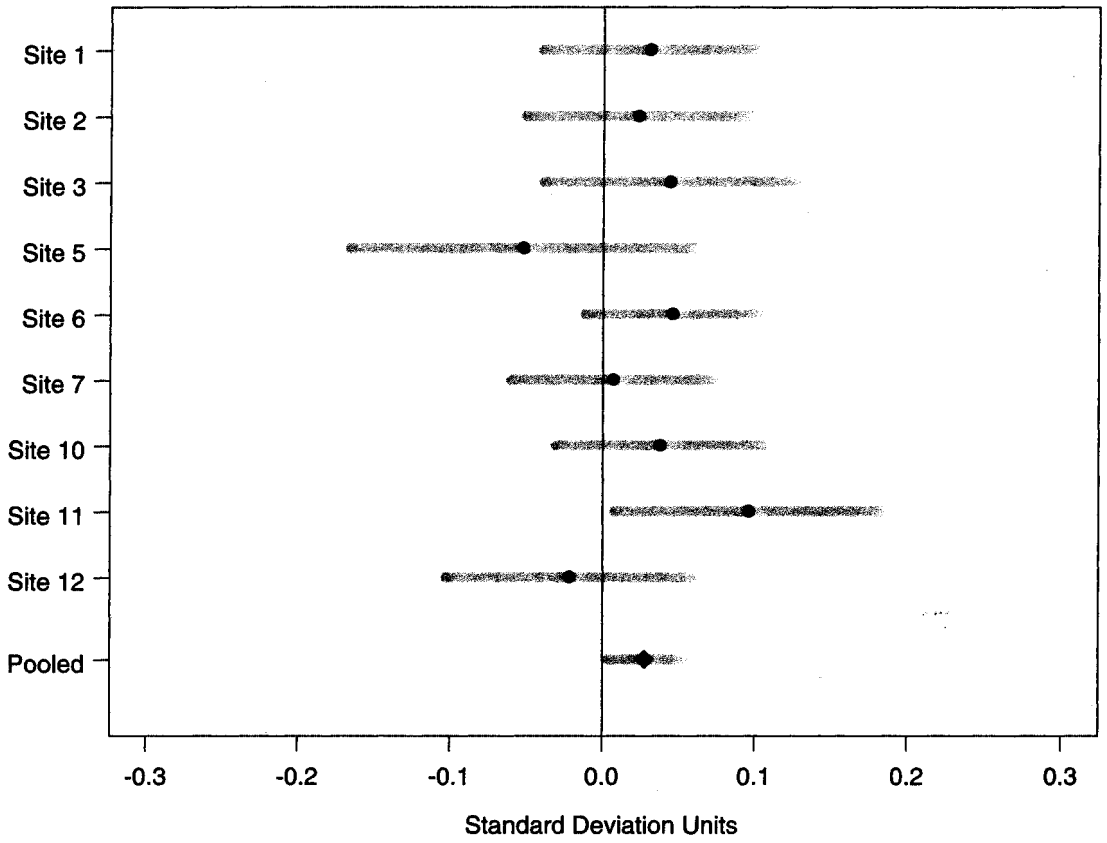

FIGURE 2. *Coefficients for reform practices, multiple-choice math.*

ment in the context of standards-based teaching than are multiple-choice measures, in part because they appear to tap skills that are similar to those that are emphasized in the classroom (e.g., open-ended problem solving). Inspection of the regression coefficients suggests that there may be a difference in the strength of the relationships between reported instructional practices and the two types of tests, but it is small. Later we discuss a test of the statistical significance of this difference.

The bottom bars in Figures 1 through 4 show the pooled estimates of the standardized regression coefficients for each of the four analyses. The coefficients and confidence interval bounds are also presented in Table 3. The confidence intervals exclude zero in all four cases, though the lower bounds are very close to zero. Nevertheless, taking all of our data into account, we observe small, positive relationships between reported use of reform-based instruction and student achievement in math and science, measured by both multiple-choice and open-response tests.

Differences between test formats

Consistent with the individual site results, inspection of the coefficients from the pooled analyses suggested slightly larger relationships between open-response scores and reported use of reform teaching practices than between multiple-choice scores and reported use of reform teaching practices, especially in math. This finding is consistent with the hypothesis that the former type of test is more closely aligned with the reforms and therefore better able to detect effects. To test the statistical significance of this difference, we calculated the difference in standard deviation units between each student's score on the open-response test and his or her score on the multiple-choice test in the same subject. We then modeled these differences as a function of teaching practices and student background covariates. The analysis was repeated for both subjects and for all sites.

The difference between formats was statistically significant in only a few sites, and it was not significant when we pooled results across sites.

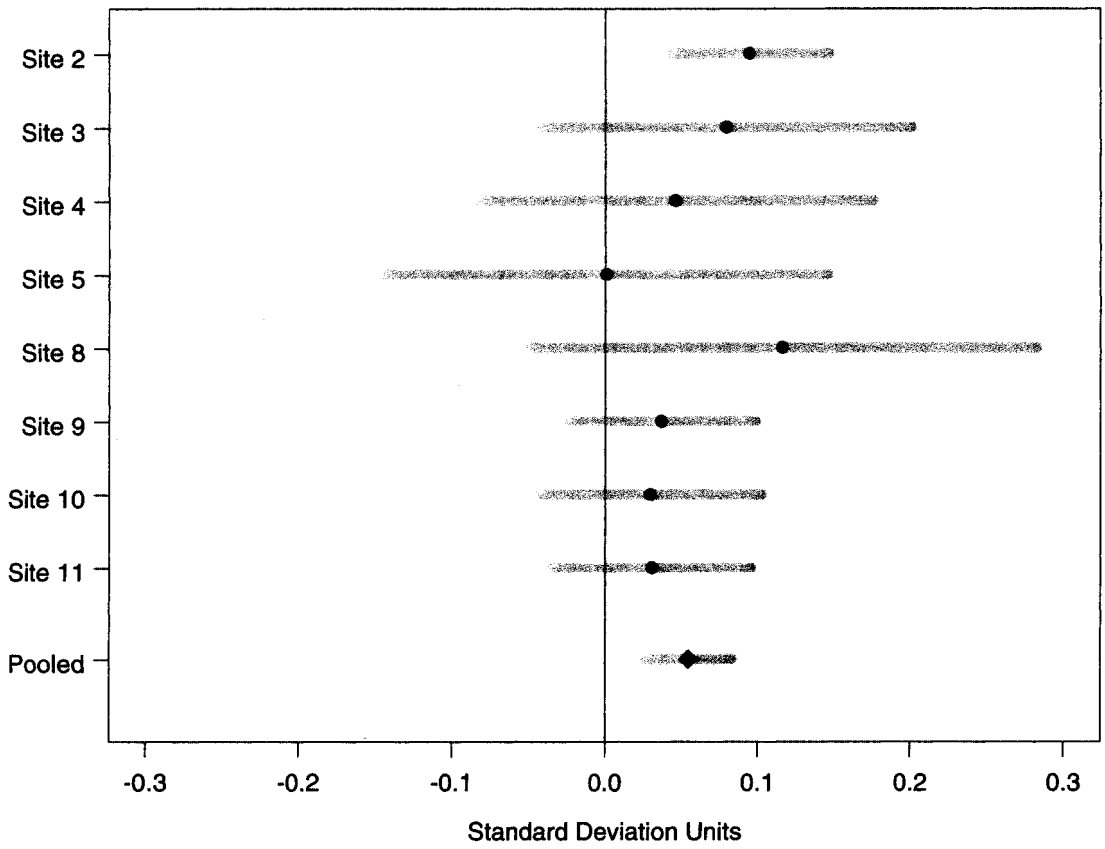

FIGURE 3. *Coefficients for reform practices, open-response science.*

Table 4 presents the coefficients from the pooled analysis, along with 95% confidence intervals. The coefficient for math was 0.031. This implies that across sites, the expected increase in student math scores for a unit increase in a teacher's score on the reform scale was 0.031 standard deviation units higher for open-response tests than for multiple-choice tests. However, our estimate was not statistically significantly different from zero. The estimate for science was even smaller. In addition, we found a relatively large between-site variance in these estimated differences, even after controlling for sampling error within site. In other words, we found that the difference in the sensitivity of open-response and multiple-choice tests varied from site to site. This variation is to be expected, given the large variations in test type within a format (e.g., open-ended science tests included both hands-on and paper-and-pencil, short-answer measures).

Thus, although inspection of regression coefficients suggests that open-response tests functioned differently from multiple-choice tests, our data do not provide sufficient evidence to support the claim that the formats differ in their relationships with teacher-reported practices. Even so, the consistency in the patterns we observed, and that fact that educators involved in these reforms often assert that open-response tests are generally more closely aligned with their efforts, suggest that further investigation of format differences is appropriate and warranted. As states continue to develop standards-based assessments, and as results from these assessments are increasingly used in evaluations of educational programs as well as for high-stakes accountability purposes, it is critical that the instructional sensitivity of different test formats be examined (for an example of an analysis of instructional sensitivity see Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002).

Discussion

The goal of the study described in this article was to investigate relationships between stu-

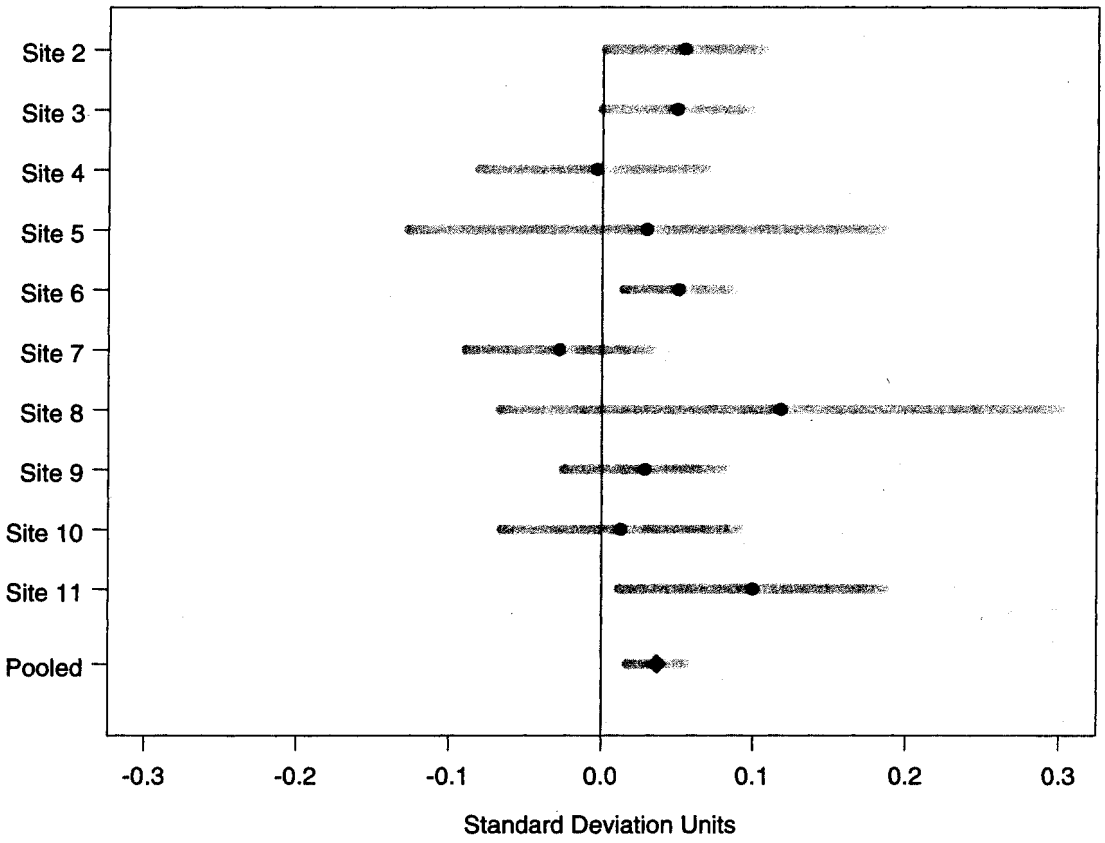

FIGURE 4. *Coefficients for reform practices, multiple-choice science.*

dent achievement and teachers' reported use of reform-based instructional practices in mathematics and science. We believe this study is important for two reasons. First, despite some limitations in measurement and design, the results indicate that there was not a strong relationship between teacher-reported instructional practices and student achievement during a given school year. At a minimum, this suggests that policymakers may need to reconsider their beliefs about large-scale reforms. For example, it might make sense to focus first on understanding what it takes to bring about changes in achievement on a small scale and then determine what steps would be necessary to instill those changes more broadly. Second, the limitations in measurement and design are themselves important leverage points for researchers. We faced challenges similar to those encountered in many educational evaluations, including the need to rely heavily on existing measures of

TABLE 3
Standardized Regression Coefficients for Reform Practices, Pooled Across Sites

Subject	Test Format	Weighted average coefficient	Lower bound of 95% confidence interval	Upper bound of 95% confidence interval
Math	OR	0.051	0.017	0.085
Math	MC	0.028	0.003	0.053
Science	OR	0.054	0.025	0.083
Science	MC	0.037	0.017	0.057

TABLE 4
Pooled Estimates of Differences between Formats

Subject	Weighted average coefficient	Lower bound of 95% confidence interval	Upper bound of 95% confidence interval
Math	0.031	-0.001	0.064
Science	0.010	-0.023	0.042

achievement and self-reported indicators of practices, as well as the need to combine information across sites. While conducting this research we had to confront these limitations, and our experience may provide an example of techniques to address the challenges inherent in evaluating reforms that attempt to change what occurs in the classroom. While we did not solve all the problems, our experience provides some guidance for others trying to do large-scale educational policy research under less than ideal circumstances.

This section provides a brief summary of the results, along with a discussion of the extent to which the study did and did not overcome the challenges we raised at the beginning of the article. We conclude with a few thoughts about ways to facilitate effective large-scale program evaluations in the future.

Summary of findings

As illustrated by Figures 1–4, the relationships between student achievement and teachers' reported use of reform practices tend to be positive but small, particularly in comparison to relationships between achievement and student background characteristics such as socioeconomic status and ethnicity. These findings are consistent with earlier studies of the systemic initiatives as well as with other studies that have examined reform-based or inquiry-based practices (e.g., Cohen & Hill, 2000; Laguarda, 1998). Although the standardized regression coefficients in our model were small, the consistent pattern across sites suggests that there may be a relationship but our design was unable to detect it due to a number of factors such as measurement error in the instruments.

The small magnitude may not be surprising given the brief period of time (less than one academic year) that was captured by teachers' questionnaire responses. Use of particular instructional strategies in a single course during a single school year would not be expected to lead to effects as large as those associated with student background characteristics. Several years of exposure may be needed to achieve a reasonably large effect. This suggests the need for a longitudinal investigation. Tracking the same students over time is especially important, given the within-school variability in teaching practices that we observed. Some students may be exposed to reform-based practices in one year but not in the subsequent year, whereas others may be exposed several years in a row. This suggests the need for a "dose-response" approach to studying relationships between instructional practices and student achievement. We are currently undertaking such a study, in which we are following students for a three-year period.

The direction of relationships was fairly consistent across sites, but their magnitudes displayed some variation. There are several potential sources of this variation. First, our models differed slightly across sites because we relied on locally available data to construct covariates. Second, various aspects of SI program implementation, such as the amount and quality of professional development activities, undoubtedly affected the kinds of teaching practices that were used. Even if two teachers report using reform practices with similar frequency, their approaches to using those practices in the classroom may differ substantially and may reflect specific features of the local reform program. Third, the achievement measures used in each site varied on a number of dimensions, including psychometric quality (e.g., reliability), content, and degree of alignment with the local curriculum. We discuss this last point below.

Addressing challenges in large-scale program evaluation

This study was designed to overcome some of the limitations of other large-scale evaluations, and specifically to address the three challenges outlined at the beginning of the article: variations in implementation and instructional practice among and within sites, lack of suitable outcome

measures, and lack of methods for combining information across multiple evaluation sites. After discussing our own attempts to address these challenges we discuss improvements that could be made to produce better information in future evaluations.

Measuring instructional practice

By gathering data from individual teachers and linking those data to the students who were exposed to instruction provided by those teachers, we attempted to investigate directly the relationship between achievement and exposure to reform-based teaching. Thus this study provides information that has been missing from other evaluations of the SI programs.

However, a weakness of our approach stems from the use of questionnaires to measure instructional practices. Like any such measure, our items are subject to inaccurate responses, particularly those that reflect social desirability. Perhaps more importantly, our questions addressed only the frequency with which teachers used particular practices and did not address the ways in which they were used or the overall quality of the instruction. Clearly, some approaches to using cooperative groups are more effective and more consistent with the intent of the reform than others, but we cannot detect these differences using our questionnaires. Multiple classroom observations, interviews, and inspection of classroom materials would undoubtedly provide a better measure of instructional practice. This type of data, however, is considerably more expensive to collect, and is usually only done on a small scale. Our questionnaire items are similar to those that have been used in numerous evaluations of this type, and to those that have been administered as part of some national longitudinal surveys (see, for example, Cohen & Hill, 2000; Swanson & Stevenson, 2002; Wenglinsky, 2002). Researchers have pointed out some of the limitations of such data collection, and there is clearly a need for new types of cost-effective, valid measures of instructional practices (Burstein et al., 1995). We are currently involved in a project that is intended to examine paper-and-pencil alternatives to traditional methods of measuring classroom practices, and we hope that the products of this project will be useful in future evaluations.

It would also be valuable to gather information on what led teachers to utilize particular practices and what role SI-sponsored professional development played in instructional decisions. Program developers would find this information particularly useful. Some teachers may have adopted certain strategies as a result of participation in the professional development activities that are provided by the SI funds, but other teachers may have found their practice shaped more strongly by different influences. It would be difficult to capture this information with a paper-and-pencil questionnaire, but the inclusion of more contextual information about the reforms in place at each site might have provided some valuable insights into the causes of the large variability in teaching practices within schools. Our initial intent was not to determine the reasons for teachers' use of practices, but information on this would be helpful to those who are designing and implementing professional development programs.

Measuring student outcomes

This study relied on achievement data that were already being collected at each site, but we supplemented these with additional assessments in many sites to ensure multiple measures of student outcomes. In particular, when only multiple-choice data were available, we tried to administer an open-response or performance-based test that was considered a better measure of some of the skills and knowledge the reforms were intended to promote. Although this approach allowed us to examine differences in the sensitivity of both types of measures to the effects of reform-based instruction, we lacked sufficient information to determine conclusively whether test format or other test characteristics are likely to affect evaluation findings.

Most evaluations rely on locally available student achievement data, in large part because it is expensive and often not feasible to administer additional measures. Many principals and teachers believe that their students spend far too much time taking the tests that are required by the district and/or state, and are therefore reluctant to volunteer for additional testing. Locally developed and administered tests may also be preferred because they are presumed to be more closely aligned with local curriculum standards than would a measure chosen and administered by outside evaluators. In many of our sites, how-

ever, test development lagged far behind the reform implementation, leaving local personnel to rely on tests that they did not necessarily believe were ideal measures of student learning. It is likely that most large-scale evaluations will have to continue to use tests that are not fully adequate for their purposes, and to find ways to combine information from a diverse set of outcome measures. Kirby et al. (in press) discuss this problem in the context of large-scale evaluations of federally funded programs such as Title I.

Although the overall differences we observed between multiple-choice and open-response tests were not significant, the general pattern suggests that format effects should be investigated further. In particular, it raises questions concerning whether the two types of tests measure different constructs. Most advocates of systemic reform believe that traditional, multiple-choice tests do not adequately reflect the range of competencies that the reforms are expected to develop, and that tests requiring students to construct their answers and to engage in complex problem-solving are more appropriate. The issue of format effects deserves further investigation, particularly given the resources that many states and districts are devoting to open-ended testing.

Combining information across multiple sites

Our planned meta-analytic approach was designed to provide a method for combining data and findings from multiple sites that differed on a number of dimensions, including the specific student background data and outcome measures available. One limitation of this approach is that it pools results from assessments that are designed to measure achievement in math or science but that differ in a variety of ways, such as the specific topics included and the degree of alignment with the curriculum. This is a problem that is frequently encountered in large-scale, multi-site evaluations, including most federal program evaluations (Kirby et al., in press). In the absence of a uniform national test, it will continue to challenge evaluators (Berends, Bodilly, & Kirby, 2002; Laguarda, Breckenridge, & Hightower, 1995). Furthermore, most published meta-analyses in education and in other fields must contend with nonequivalent outcome measures (see, e.g., Hedges, Laine, & Greenwald, 1994). There is a trade-off between comparability and relevance that can never be fully resolved. In this case we opted to use the measures that were adopted in each site and reflect criteria known and endorsed by local teachers and administrators. As a consequence, the outcomes in one site were not necessarily comparable to the outcomes in another. To address this concern, we conducted the analyses separately by site and presented these results as well as the pooled results. Providing both analyses does not solve the problem but it permits users of the results to see how conclusions might be affected by the choice of measures.

Although this approach is promising, it suffers from some of the drawbacks of traditional meta-analyses—most significantly, cross-study variations in how constructs are measured and how models are specified. Although we were able to control these variations in ways that are not possible with traditional meta-analyses by gathering and using student-level covariates, limitations in data systems (e.g., missing covariates) and lack of a common outcome measure prevented us from specifying identical models across sites. Many states are currently engaged in efforts to build better data systems, and these will certainly enhance future evaluation efforts. The resources needed to create a data system are substantial, but the investment will undoubtedly result in more useful information that can guide school reform efforts. However, evaluations that include multiple states will continue to face the problem of nonequivalent achievement measures.

Yet, another challenge may actually loom larger than any of the three we tried to address. That is the importance of linking student and teacher records. The substantial variability among responses of teachers within the same schools suggests that school-level analyses would have masked important differences in the curriculum and instruction to which students were exposed, and that assigning schools to categories such as high or low implementers is likely to produce misleading information about what is occurring in those schools. However, it is not always possible to obtain the data that are necessary to link students and teachers. Thus, many evaluators will probably be forced to use data that mask important within-school differences.

Improving future program evaluation

Countless education reform evaluations are currently underway, and many of these focus on

programs that try to change what happens in the classroom. The federal government recently funded a number of evaluations of comprehensive school reform models, for example, in an effort to determine which models show the most promise for improving student achievement. Most of these models include guidelines for curriculum and instruction, and many espouse instructional approaches similar to those we have studied in the context of NSF's Systemic Initiatives programs. The three components that we identified at the beginning of this article—(a) a cross-site analysis that allowed us to replicate the study in a variety of contexts; (b) multiple outcome measures with controls for prior achievement; and (c) measurement of implementation at the classroom rather than school level with direct links between teachers and students—are likely to be useful in other types of large-scale evaluation.

In addition to issues we have already discussed, there are some steps that program developers, evaluators, and policymakers could take to improve the prospects of gathering useful evaluative evidence of program efficacy. The first involves the difficulty inherent in linking student outcomes to the specific classroom environment to which the student was exposed. We were able to do this by collecting rosters, but this was an enormously time-consuming activity, and would not be feasible in a very large evaluation such as those that include representative samples of schools from multiple states. Building teacher-student links into large-scale data systems would dramatically expand the kinds of questions evaluators could ask, and would result in more refined information on what approaches are effective. It also permits the use of value-added methods for examining program effects on individual students (Sanders, Saxton, & Horn, 1997). Unfortunately, for many evaluations, only school-level data are available. In particular, the federal government has emphasized the use of school-level data in many of its funded evaluations due to human subjects and privacy concerns under the Family Education Rights and Privacy Act (Kirby et al., in press). Even when individual student-level data are available, linking those data to teachers is frequently impossible, and efforts to develop linked databases often meet with strong political resistance. It has been difficult to overcome this resistance in the past, but at least one state (Tennessee) has negotiated agreements with teachers to create a linked data system for research purposes. Tennessee's success in doing "value-added" analyses of the relationship between instruction and achievement has captured the attention of many policymakers (Sanders, Saxton, & Horn, 1997). It may be the case that the political will now exists in other states to create linked, longitudinal data systems.

Effective evaluation of educational programs also requires an appropriate outcome measure, as we discussed earlier. Currently the majority of state tests rely heavily or exclusively on multiple-choice items, and the use of the more expensive open-response format is likely to decline further as states increase the numbers of grades and subjects tested (and therefore the cost of testing). In addition, in states with high-stakes testing programs, gains on state tests are often not replicated on other tests (e.g., Koretz & Barron, 1998). In a few of our sites we observed unusual relationships among state test scores and scores on the tests we administered, suggesting that at minimum the two sets of tests are capturing different aspects of student achievement. In short, the method used for measuring achievement matters, but most evaluations do not have sufficient funding to permit the use of multiple outcome measures. Evaluators need to look carefully at the measures that they are using, and explore alternative methods for refining the information (e.g., the use of subsets of items that may be especially closely aligned with a particular curriculum reform). Where possible, those responsible for developing and implementing programs should build in assessments that are considered appropriate measures of the skills and knowledge the program is intended to promote.

Third, large-scale, quantitative evaluations such as this one should be supplemented with detailed case studies that provide richer information on curriculum, instruction, professional development, and other important aspects of the program. Although few evaluation budgets are large enough to support the collection of this type of detailed data across a large number of sites, a case-study approach may be fruitfully applied to a small sample of program sites in order to aid interpretation of the quantitative findings and help users of the results to understand the context in which the program operates (Stecher & Borko, 2002).

Finally, assessment and evaluation should be built into reform programs from the outset. As with most educational research, our inability to investigate effects using an experimental design limits the inferences that can be made from our results. Perhaps the primary problem is that without random assignment of students and teachers to treatments, we cannot be certain that the relationships we observed can be attributed solely to classroom practices. There may be other differences in student characteristics across classrooms that contribute to differences in performance and that influence what teachers do. For example, teachers may tend to engage in more reform-based practices with higher achieving students, or may simply be more highly skilled than those who do not engage in these practices. Controlling for prior achievement and examining relationships with teacher background, as we have done here, is helpful but does not eliminate the problem completely.

The new federal No Child Left Behind legislation places a heavy emphasis on scientifically-based programs, and much of the discussion surrounding education research recently has focused on the need for randomized experiments. Certainly educational policymakers should be encouraged to conduct randomized trials of new programs before implementing them on a large-scale. If such experiments are not feasible for practical, political, or ethical reasons, building evaluation into reform implementation may improve the prospects for carrying out well-designed studies of program effectiveness. For example, as new reforms are implemented in small numbers of schools, wherever feasible the schools should be chosen in a way that minimizes pre-existing differences. In addition, ongoing student assessment that is aligned with the program should be an integral part of the program package, so that the data necessary to evaluate the program's effects are collected from the very beginning. Maintaining longitudinal databases on students and program implementation would further strengthen our ability to do rigorous evaluation. This would enable evaluators to examine changes in implementation and growth in student achievement over time, providing a much stronger test of the program's effects than what is typically obtained through a cross-sectional comparison.

In short, our evaluation builds on earlier work (Cohen & Hill, 2000; Wenglinsky, 2002) to examine relationships between teacher-reported reform-based instruction and student achievement in mathematics and science, but our conclusions, as well as those of many of the other studies of this topic, are affected by the assessment and design issues discussed above. Changes to data systems and large-scale assessment programs as well as stronger links between program implementation and evaluation are needed to provide a means of conducting evaluations that can provide clear evidence concerning the relationship between classroom practices and student achievement.

Notes

[1] By contrast, there are reforms that focus primarily on structure (e.g., size of school or classroom, student promotion policies, length of school day or year) with little attention to curriculum or pedagogy.

[2] The No Child Left Behind legislation requires states to administer math and reading tests in grades 3-8, but science is only required to be tested in a few grades.

[3] Classrooms are, of course, nested within schools, leading to a three-level hierarchical structure. We focus on the classroom level rather than the school level because we were primarily interested in variation among teachers. It would have been possible to include schools as fixed effects (random effects would not be appropriate, given our sampling strategy), but this would lead to some confounding between the school and classroom levels because some schools have only one teacher. Also, adjusting for school-level differences might have removed some of the teacher effects that we were trying to capture.

[4] As we discuss later, the lower reliability of the traditional practices scale will tend to attenuate relationships with achievement.

[5] In each model we tested whether a nonlinear term provided a better fit than a linear term. In all but two models that used the traditional practices scale, the linear term provided an adequate fit.

[6] The full regression models, which present the coefficients for the student background characteristics in addition to the instructional practice coefficients, are omitted to preserve space. They are available from the authors.

[7] We used our model to predict the score for the "average" student (a student with all student background predictors set to the mean) with a teacher scoring 3 on each reform practices item (monthly use of reform practices). We then found the percentile of this predicted score among the test scores from the site, and repeated the process for the average student with a teacher scoring 4 on each item (weekly use). The percentile is based on our sample and is not a percentile from a national norming group.

Appendix A

Items on Instructional Practices Scales

Items on Reform Practices Scale for Math

About how often do you typically do each of the following in your *mathematics* instruction in this class?

- Arrange seating to facilitate student discussion
- 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
- Allow students to work at their own pace
- Read and comment on the reflections students have written in their notebooks or journals

About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?

- Participate in student-led discussions
- Work in cooperative learning groups
- Make formal presentations to the class
- Work on solving a real-world problems
- Share ideas or solve problems with each other in small groups
- Engage in hands-on mathematical activities
- Design or implement their *own* investigations
- Work on extended mathematics investigations (a week or more in duration)
- Participate in field work
- Record, represent, and/or analyze data
- Write a description of a plan, procedure, or problem-solving process
- Write reflections in a notebook or journal
- Work on portfolios
- Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
- Engage in performance tasks for assessment purposes

Appendix B

Items on Traditional Practices Scale for Math

About how often do you typically do each of the following in your *mathematics* instruction in this class?

Lecture/introduce content through formal presentations

About how often do students in this class typically take part in each of the following activities as part of their *mathematics* instruction?

Read from a mathematics textbook in class

Practice computational skills

Memorize mathematics facts, rules, or formulas

Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

Appendix C

Items on Reform Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?

- Arrange seating to facilitate student discussion
- Use open-ended questions
- Require students to supply evidence to support their claims
- Encourage students to explain concepts to one another
- Encourage students to consider alternative explanations
- Allow students to work at their own pace
- Read and comment on the reflections students have written in their notebooks or journals

About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?

- Participate in student-led discussions
- Work in cooperative learning groups
- Make formal presentations to the class
- Work on solving a real-world problem
- Share ideas or solve problems with each other in small groups
- Engage in hands-on science activities
- Design or implement their *own* investigations
- Design objects within constraints (e.g., egg drop, toothpick bridge, aluminum boats)
- Work on extended science investigations or projects (a week or more in duration)
- Participate in field work
- Record, represent, and/or analyze data
- Write reflections in a notebook or journal
- Work on portfolios
- Take tests requiring open-ended responses (e.g., descriptions, justifications of solutions)
- Engage in performance tasks for assessment purposes

Appendix D

Items on Traditional Practices Scale for Science

About how often do you typically do each of the following in your *science* instruction in this class?

Lecture/introduce content through formal presentations

About how often do students in this class typically take part in each of the following activities as part of their *science* instruction?

Read from a science textbook in class

Answer textbook/worksheet questions

Learn science vocabulary

Take short-answer tests (e.g., multiple-choice, true/false, fill-in-the-blank)

References

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy: Project 2061.* New York: Oxford University Press.

Berends, M., Bodilly, S. J., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform. New American Schools after a decade* (MR-1498-EDU). Santa Monica, CA: RAND.

Berman, P., & McLaughlin, M. W. (1978, May). *Federal programs supporting educational changes: Vol. VIII: Implementing and sustaining innovations* (R-1589/8-HEW). Santa Monica, CA: RAND.

Briars, D. (2001). *Mathematics performance in the Pittsburgh Public Schools.* Paper presented at a conference of the Mathematics Assessment Resource Service, San Diego, CA.

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal, 34,* 297–331.

Burstein, L. McDonnell, L. M., Van Winkle, J., Ormseth, T. H., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators.* MR-658-NSF. Santa Monica, CA: RAND.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12,* 331–338.

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

Consortium for Policy Research in Education (1995). *Reforming science, mathematics, and technology education: NSF's State Systemic Initiatives* (CPRE Policy Brief). New Brunswick, NJ: Author.

Corcoran, T. B., Shields, P. M., & Zucker, A. A. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's and professional development for teachers.* Menlo Park, CA: SRI International.

Education Week (2001, January). *Quality counts 2001:* Bethesda, MD: Author.

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology, 84,* 331–339.

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective: Results from a national sample of teachers. *American Educational Research Journal, 38,* 915–945.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90,* 560–569.

Goertz, M. E., & Duffy, M. C. (2001). *Assessment and accountability systems in the 50 states: 1999–2000.* Philadelphia, PA: Consortium for Policy Research in Education.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London, UK: Arnold.

Greenwood, C. R., Carta, J. J., & Hall, R. V. (1988). The use of peer tutoring strategies in classroom management and educational instruction. *School Psychology Review, 17,* 258–275.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.

Hill, P. T. (1995). *Reinventing public education.* Santa Monica, CA: RAND.

Kirby, S., McCaffrey, D., Sloan-McCombs, J., Naftel, S., Barney, H., & Lockwood, J. R. (in press). Using school-level test scores to evaluate federal programs: Proceed with caution. *Peabody Journal of Education.*

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Robyn, A., & Burroughs, D. (2000). *Teaching practices and student achievement: Report of first-year results from the Mosaic Study of Systemic Initiatives in Mathematics and Science* (MR-1233-EDU). Santa Monica, CA: RAND.

Knapp, M. S. (1997). Between systemic reforms and the mathematics and science classroom: The dynamics of innovation, implementation, and professional learning. *Review of Educational Research, 67,* 227–266.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Laguarda, K. G. (1998). *Assessing the SSIs' impacts on student achievement: An imperfect science.* Menlo Park, CA: SRI International.

Laguarda, K. G., Breckenridge, J. S., & Hightower, A. (1995). *Assessment programs in the Statewide Systemic Initiatives (SSI) states: Using student achievement data to evaluate the SSI.* Washington, DC: Policy Studies Associates.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

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

McCaffrey, D. F., Bell R. M., & Botts, C. H. (2001, August). Generalizations of bias reduced linearization. *Proceedings of the Annual Meeting of the American Statistical Association.*

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.

National Science Foundation. (2001a). *Academic excellence for all urban students.* Washington, DC: Author.

National Science Foundation. (2001b). *Urban Systemic Program in science, mathematics, and technology education (USP): A foundation for K-12 science and mathematics educational system reform* (Program solicitation NSF 01-15). Washington, DC: Author.

Neder, J., Kutner, M. H., Nachtsheim, C. J., & Wassmerman, W. (1996). *Applied linear statistical models* (4th Ed.). Chicago, IL: Irwin.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: J. Wiley & Sons.

Ruiz-Primo, A., Shavelson, R., Hamilton, L. S., & Klein, S. P. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal for Research in Science Teaching, 39,* 369–393.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-added Assessment System: A quantitative, outcomes-based approach to educational assessment. In Millman, J. (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Sage.

Schafer, J. L., (1997). *Imputation of missing covariates under a general linear mixed model.* Technical report retrieved May 6, 2003 from http://www.stat.psu.edu/~jls/.

Schafer, J. L., (1998). PAN Software for S-Plus retrieved May 6, 2003 from http://www.stat.psu.edu/~jls/misoftwa.html#splus.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13–25.

Shields, P. M., Corcoran, T. B., & Zucker, A. A. (1994). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: First-year report.* Menlo Park, CA: SRI International.

Shields, P. M., Marsh, J. A., & Adelman, N. E. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's Impacts on Classroom Practice.* Menlo Park, CA: SRI International.

Smerdon, B. A., Burkam, D. T., & Lee, V. E. (1999). Access to constructivist and didactic teaching: Who gets it? Where is it practiced? *Teachers College Record, 101,* 5–34.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–268). Bristol, PA: The Falmer Press.

Stecher, B. M., & Borko, H. (2002). Integrating findings from surveys and case studies: Examples from a study of standards-based educational reform. *Journal of Education Policy, 17,* 547–570.

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,* 1–27.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia.* Cambridge, MA: Harvard University Press.

Verschaffel, L., & De Corte, E. (1997). Teaching realistic mathematical modeling in the elementary school: A teaching experiment with fifth-graders. *Journal for Research in Mathematics Education, 28,* 577–601.

Von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *Journal of Educational Research, 95,* 151–160.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Macmillan.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., & Bond, S. L. (1998). *Local systemic change through teacher enhancement: Year three cross-site report.* Chapel Hill, NC: Horizon Research, Inc.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10,* Retrieved May 6, 2003 from http://epaa.asu.edu/epaa/v10n12/.

Williams, L. (1998). *The Urban Systemic Initiatives (USI) program of the National Science Foundation: Summary update.* Washington, DC: National Science Foundation.

Authors

LAURA S. HAMILTON is a Senior Behavioral Scientist with RAND, 1700 Main St. PO Box 2138, Santa Monica, CA 90407-2138; laurah@rand.org. Her areas of specialization are test-based accountability, educational measurement and evaluation.

DANIEL F. McCAFFREY is a Statistician with RAND, 201 North Craig St., Suite 202, Pittsburgh, PA; daniel_mccaffrey@rand.org. His areas of specialization are variance estimation, hierarchical and mixed models, weighting mixed data and imputation,

performance assessment, validity of test scores and gains, teaching practices, student achievement and school reform.

BRIAN M. STECHER is a Senior Social Scientist with RAND, 1700 Main St. PO Box 2138, Santa Monica, CA 90407-2138; brian_stecher@rand.org. His areas of specialization are educational measurement and evaluation.

STEPHEN P. KLEIN is a Senior Research Scientist with RAND, 1700 Main St. PO Box 2138, Santa Monica, CA 90407-2138; stephen_klein@rand.org. His areas of specialization are measurement, evaluation, and statistical analysis.

ABBY ROBYN is a Social Research Analyst with RAND, 1700 Main St. PO Box 2138, Santa Monica, CA 90407-2138; abby_robyn@rand.org. Her areas of specialization are program implementation, assessment.

DELIA BUGLIARI is a Statistical Programmer with RAND, 1700 Main St. PO Box 2138, Santa Monica, CA 90407-2138; delia@rand.org. Her areas of specialization are education and economic survey data.

Manuscript Received February 5, 2002
Revision Received September 11, 2002
Accepted, April 1, 2003