WORKING PAPER

Using Structured Classroom Vignettes to Measure Instructional Practices in Mathematics

BRIAN M. STECHER, VI-NHUAN LE,
LAURA S. HAMILTON, GERY RYAN,
ABBY ROBYN, J.R. LOCKWOOD

WR-336-NSF

February 2006

RAND EDUCATION

Abstract

Large-scale educational studies frequently require accurate descriptions of classroom practices to judge implementation and impact. However, such descriptions are difficult to obtain in a timely, efficient manner. To address this problem, we developed a vignette-based measure of one aspect of mathematics instructional practice, reform-oriented instruction. Teachers read contextualized descriptions of teaching practices that varied in terms of reform-oriented instruction and rated the degree to which the options corresponded to their own likely behaviors. Responses from 80 fourth-grade teachers yielded fairly consistent responses across two parallel vignettes and moderate correlations with other scales of reform-oriented instruction derived from classroom observations, surveys and logs. The results suggested that the vignettes are measuring important aspects of reform-oriented instruction that are not captured by other measurement methods. Based on this work, it appears that vignettes can be a useful tool for research on instructional practice, but cognitive interviews with participating teachers provided insight into possible improvements to the items.

Introduction

There are a number of reasons researchers are interested in developing methods to obtain better descriptions of instructional practice on a large scale. First, research using value-added assessment methods suggests that some teachers are more effective than other teachers (Ferguson & Ladd, 1996; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997), although this research has not yet identified the characteristics or practices that are associated with effectiveness. Second, many large-scale educational reforms, such as the National Science Foundation (NSF) Systemic Initiatives, involve changes in instruction, and it is difficult to evaluate the progress of the reforms (e.g., the degree of implementation) without measures of instructional practice. Third, information about changes in practice is important in validating changes in test scores (Koretz, 2003). Fourth, with increased demands for "highly qualified" teachers, measures of instructional practice can inform improvements in teacher education and professional development (King & Newman, 2000).

Despite the potential usefulness of objective measures of instructional practice, only recently have there been large-scale efforts to gather data about what actually transpires in classrooms (Mayer, 1999). Kennedy (1999) describes a continuum for characterizing indicators of classroom processes in terms of their "distance" from actual classroom events. Classroom observations (and by extension, video records of classrooms) are the most proximal indicators of instruction, providing rich descriptions of activities that occur in the classroom. However, repeated classroom observations conducted over an extended period of time—which some might consider the "gold standard" for measuring classroom practice—are difficult to conduct, time-consuming, and expensive. They are rarely feasible in large-scale studies. Less frequent observation is more practical, but sampling only one or two days of instruction introduces unknown sampling error into the measurement of practice and might not provide an accurate measure of the typical practices teachers use throughout the entire year.

Surveys, which are quite distant from actual classroom events, are the most common way of collecting information about instructional practice on a large scale. Written surveys provide a relatively cost-effective method for gathering information about practice from large numbers of teachers and can capture information that covers a longer time frame than is typically feasible using observations. A number of researchers have used surveys to measure specific aspects of classroom practice and to explore relationships between instructional practices and student achievement (Cohen & Hill, 2000; Gamoran, Porter, Smithson, & White, 1997; Wenglinsky, 2002; 2003).

However, surveys have limitations as measures of classroom practice (Rowan, Correnti, & Miller, 2002). One serious shortcoming is that teachers and researchers may interpret survey terminology in different ways, leading to invalid inferences about practice (Burstein, et al., 1995; Stigler & Perry, 2000). For example, Antil, Jenkins, Wayne, and Vasdasy (1998) found this to be true when they interviewed a group of teachers whose survey responses indicated frequent use of cooperative groups. After probing for more details about what teachers actually did, they found that teachers' descriptions did not match the various definitions of cooperative grouping put forth by researchers. Related findings were reported by Mayer (1999), who compared teachers' reports of the time they spent encouraging students to engage in oral discussion with the reports of classroom observers. The observers found that the nature of the student oral discussion varied in important ways not revealed by the survey responses. Mayer concluded that survey data can distinguish between teachers who frequently use certain types of practices from teachers who do not, but it cannot provide information about how those practices are substantively implemented. Desimone and LeFloch (2004) described methods for using cognitive interviews to improve the quality of survey measures of classroom practices and reduce these types of misunderstandings, although they noted that these methods are not yet widely used in survey development.

Surveys are also limited with respect to scope. Most large-scale surveys are designed to be applicable to a wide range of settings, and focus on long-term patterns of behavior. Because of this, they often ignore important variations in teaching strategies related to grade level or content (Mullens & Gayler, 1999). In this context, surveys provide at best general descriptions of classroom events.

A third method of data collection is logs, in which teachers report on specific details of their curriculum and instruction during a specified time period (Ball et al, 1998; Rowan, Harrison, & Haynes, 2004; Rowan, Camburn & Correnti, 2004). Logs are situated in the middle of Kennedy's continuum—closer to actual instruction than surveys, but more distant than direct observation. They offer a relatively cost-effective method for collecting information about the frequency of specific teaching behaviors. Logs suffer from some of the same problems as surveys, including possible uncertainty about the meaning of terminology. However, researchers can overcome some problems when logs are used for an extended period through better communication with teachers. For example, Rowan, Harrison and Haynes (2004) used logs to obtain reliable descriptions of students' exposure to specific strands of the mathematics curriculum by providing teachers with an extensive glossary defining key terms in the logs and a toll-free telephone support line teachers could call to obtain answers to questions about how to report their activities.

Although not a new technique, per se, vignettes—contextualized descriptions of classroom situations—offer another strategy for prompting descriptions of instructional practice. It is difficult to classify vignettes in terms of Kennedy's continuum—because they rely on contextualized descriptions of behaviors they might seem "closer" to classroom events, but the hypothetical nature of the responses might make them seem more "distant." Vignettes would appear to have certain advantages for gathering data on classroom practices. They may make the data collection process more realistic for teachers by providing a classroom context in which to situate their responses. In addition, they are standardized, so responses from teachers can be aggregated and compared (Kennedy, 1999; Ma, 1999; Ruiz-Primo & Li, 2002). However, using "real life" prompts does not guarantee that teachers' responses to vignettes will reflect their behavior in a classroom setting.

The research on classroom vignettes is quite limited. The psychological literature shows that intentions can predict behavior (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975; Sheeran, Orbell, & Trafimow, 1999), which suggests that responses to vignettes might reflect actual teaching practices to some degree. Ruiz-Primo and Li (2002) provide partial evidence that teachers' responses to vignettes bear some congruence to their instruction. Ruiz-Primo and Li compared teacher comments on hypothetical journal entries with teachers' actual feedback on their students' science notebooks from the previous year. They found that quality of feedback on the journal entries was related to quality of feedback provided on the notebooks. However, they were not able to control for other factors that may have influenced teachers' vignette-based responses so these results are not definitive. In addition, this study focused on teachers' written feedback rather than on their verbal interactions with students and other classroom behaviors, so it is not clear how well these results generalize to more typical aspects of classroom practice.

Thus, the initial evidence suggests that classroom vignettes are a promising approach for measuring instructional practices, but more research about the validity of vignette-based measures is warranted.

This study used a particular form of vignette; teachers were presented with a written description of a classroom situation and a set of possible actions the teacher might take. Respondents were asked to indicate the likelihood that they would engage in each action had they found themselves in a similar situation. This written, closed-ended form was adopted because it could be used easily in a large-scale research project. (Vignettes can also be presented orally or using video, and respondents can be allowed to give open-ended responses orally or in writing.)

Purpose of this Study

The purpose of this study is to evaluate the validity of structured, closed-ended vignettes as indicators of teaching practice. Responses to vignettes are compared with information gathered using alternative measures of instruction, including surveys, logs, and classroom observations. The extent to which the responses are consistent, particularly the responses between vignettes and classroom observations, provides some evidence in support of the validity of vignette-based measures of instruction.

However, it is important to keep in mind that surveys and logs may not be optimum sources for validating vignettes (Burstein et al., 1995). Teachers can provide equally inaccurate responses on vignettes as on surveys or logs (or other self-reported measures) so "the instruments cannot be used as checks against each other" (Mayer, 1999, p. 33). Furthermore, the research assessing the validity and reliability of surveys and logs is sparse, so it is unclear whether surveys and logs are the best criterion measures. Nonetheless, it makes sense to examine the relationships between vignettes and surveys and logs because a high degree of correspondence would suggest that the measures are capturing stable teacher responses that are more likely to be reflective of actual practice.

Classroom observations probably serve as the best external benchmark for validating teachers' responses to classroom vignettes because observations are far more proximal to instruction than the other methods. However, even observations have limitations as a validation criterion. It is not usually possible to compare a teacher's response to a vignette with his or her actions in a similar situation because it is impossible to predict when, or if, the teacher will face such a situation in his or her own classroom. The very feature that distinguishes vignettes—the ability to specify context richly—lessens the likelihood of observing a comparable event when visiting the classroom.

In this study, we examine the relationships between vignette-based measures, and surveys, logs, and classroom observations to better understand how vignettes function as measures of teaching practices. Although each of the measures is imperfect, in combination they provide a reasonable set of criteria for judging the validity of inferences drawn from the vignettes. The study focuses primarily on "reform-oriented" instructional practices (described below), i.e., a subset of instructional activities that is consistent with a particular philosophy or approach to teaching. Teaching is a complex endeavor, and, by focusing on one aspect of instruction, we do not mean to suggest that teaching is unidimensional. Quite the contrary; it is difficult to describe teaching effectively because it is a multidimensional activity. Experts disagree both on how to characterize those dimensions and on their relative importance. This study is restricted to one aspect of instructional practice—reform-oriented instruction—that has been found to be a meaningful construct for looking at mathematics instruction (Weiss, et al., 2001). We are testing the validity of vignettes to describe this dimension of instructional practice.

Context of this Study

The research reported here was part of a multiyear, multi-site study of instruction and student achievement. This larger study sought to extend prior research that found exposure to reform-oriented instruction was weakly related to student achievement in mathematics and science (Author, 2003). For both studies, reform-oriented instruction was defined as practices consistent with those highlighted in the *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989) and those associated with the *National Science Education Standards* (National Research Council, 1996). Such practices place greater emphasis on developing conceptual understanding, solving novel or complex problems, and communicating mathematical and scientific ideas. Expert Advisory Committees helped to develop operational definitions of reform-oriented mathematics and science instruction.

Reform-oriented instruction is often contrasted with "traditional" instruction, which emphasizes factual knowledge, mastery of algorithms, and solving structured problems. However, reform and traditional instruction should not be thought of as opposite ends of a single dimension but as separate dimensions; research shows that teachers use practices associated with both approaches (Cohen and Hill, 2000; Weiss et al, 2001). For example, in semi-structured interviews, 44% of a sample of teachers in this study described their practice as incorporating aspects of both traditional and reform practice. In fact, evidence suggests that the dimensions can be orthogonal in many contexts (Klein et al., 2000), that is, the use of reform-oriented practices is not associated, either positively or negatively, with the use of traditional practices.

The present study extended the previous work, in part, by using more-extensive indicators of classroom practice, including vignette-based measures. The research examined teaching practices in elementary and middle school mathematics and science in three large metropolitan school districts over a three-year period. In addition to information about instruction, the study gathered data on teacher background (including education, certification, and experience), participation in professional development, subject matter content (including textbook usage, supplementary materials, topic coverage), classroom characteristics (composition of student achievement, attendance, class length), and student behavior. Five cohorts of students (three in mathematics and two in science) were followed longitudinally from third through fifth grade, from sixth through eighth grade, or from seventh through ninth grade. This paper focuses on the elementary-school mathematics cohort, which was the group for which the largest number of classroom observations was conducted. The results presented here reflect responses received during the second year of the three-year study, when the largest number of teachers participated in the data collection.

Methods

Sample

The elementary school mathematics cohort of the study was sampled from a large urban/suburban district of approximately 70,000 students that was participating in the NSF-sponsored Local Systemic Change (LSC) program. A major part of the district's LSC effort was devoted to training to increase teachers' use of classroom practices that are consistent with mathematics reform principles, i.e., affording greater importance to developing conceptual understanding, solving problems, and communicating mathematical ideas (National Council of Teachers of Mathematics, 1989; 2000).

The district contained about 50 elementary schools, and the student population was 58% white, 32% Hispanic, 4% African-American, 4% Native American, and 2% other. Slightly over one-half of the students were classified as economically disadvantaged and were eligible to receive free or reduced-price lunches. To achieve adequate statistical power to detect a non-zero relationship between instructional practices and student achievement, we sampled 20 elementary schools. To minimize the variability in our final estimates we selected a sample that was uniformly distributed with respect to aggregate demographic and SES variables. . As a result, the school sample was representative of the district as a whole with respect to these student characteristics.

All 85 fourth-grade teachers in the participating schools were included in the study. To encourage participation, teachers received a modest honorarium for completing the instructional practice measures.

Data Collection

Four different data collection methods were used in the study: a written survey, which included a separate set of vignette-based questions, classroom logs, and direct observations. All teachers were asked to complete the surveys, vignette-based measures and logs. In addition, about one-half of the teachers were observed by members of the research team. After the formal data collection was completed, we gathered supplemental information about teachers' reactions to the vignettes through interviews. These cognitive interviews were conducted with a subset of teachers who participated in the larger research project, and these were designed to gain insights into the functioning of the vignette-based survey items. Table 1 lists the data collection instruments and the topics covered by each.

Table 1: Data Collection Instruments

Instrument	Topics Covered	Participating Teachers
Survey (excluding vignettes)	Professional development Curriculum materials Content coverage Use of skill-based groups Teaching strategies* Student learning activities* Teacher background and experience Classroom composition	85
Logs	Focus of lesson Curriculum materials Attendance Time devoted to administrative activities Groupwork Student learning activities* Teaching strategies*	85
Vignette-based items	Introducing the lesson* Responding to student error* Reconciling different approaches* Choosing learning objectives*	85
Classroom observations	Groupwork Student learning activities* Teaching strategies*	39
Cognitive interviews	Interpretation of vignette-based items	30

* These questions included alternatives that were reform-oriented and alternatives that were not.

With the exception of the vignette-based items, questions on the surveys and logs were similar to question used extensively in research on mathematics instruction (see for example, Cohen & Hill, 2000; Hamilton et al., 2003; Weiss, et al., 2000; Wenglinsky, 2002). Consequently, these instruments will be described very briefly in subsequent paragraphs. More attention will be paid to the development of the vignette-based items and the procedures used to conduct classroom observations.

Survey. Each teacher was asked to complete a written survey during the spring of the year. The survey included questions about the teacher's educational background and experience, participation in professional development, the mathematics textbooks used and the mathematical topics taught in the class, and the use of a variety of teaching practices. Teachers indicated how much class time was spent on various mathematical topics (e.g., multiplication/division of whole numbers, patterns/functions/algebra). They indicated the frequency with which they engaged in particular instructional activities (e.g., lecture or introduce content through formal presentations, encourage students to explore alternative methods for solutions). They also indicated the frequency with which students took part in specific learning activities (e.g., practice computational skills, work on extended mathematical investigations). The surveys also included the vignettes, which are described below.

Logs. For a two-day period during the spring, teachers were asked to complete a daily log describing specific activities that occurred during their mathematics lesson each day.[3] While the surveys focused on long-term patterns of behavior, the logs focused on specific activities that occurred each day. Teachers reported on the amount of time they spent on selected instructional activities (e.g., monitor students as they work, ask questions of individuals to test for understanding), the amount of time students were engaged in selected learning activities (e.g., use manipulatives to solve problems, complete worksheets or problem sets from text), and the frequency with which particular events occurred during the lesson (e.g., students engaged in a debate or discussion of ways to solve a problem, teacher or student connected today's math topic to another subject, such as social studies). Teachers completed the logs at their convenience, either immediately after class or after school.

Vignette-Based Items. The purpose of the vignette-based items was to provide a Behaviorally anchored indication of teachers' propensity to use reform-oriented instructional practices. By describing teaching options in terms of specific, situated behaviors, we wanted to avoid the problems of language and terminology that have been documented with surveys and logs. Rather than using reform terminology, such as "cooperative groups," or "mathematical communication," the vignette-based items attempted to describe a specific instructional situation and present alternative teacher actions in neutral, behavioral terms.

The first step in developing the vignette-based items was to distill a list of reform-oriented instructional practices that could serve as a basis for vignette development. We convened an expert panel of mathematicians and mathematics educators to help in this process. Using the NCTM Standards and other reform documents, the panel developed a taxonomy containing 23 elements of reform-oriented mathematics curriculum and instruction organized into three major categories: the nature of mathematics, students' mathematical thinking, and mathematics instruction (See Appendix A.). The expert panel helped us frame the taxonomy in behavioral terms, drawing clear contrasts between the presence or absence of each element. They also provided benchmark descriptions of behaviors that were indicative of more and less reform teaching, as well as behaviors that were not associated with reform teaching. The non-reform behaviors included a variety of practices associated with other approaches that were not the focus of this study.

In addition, the expert panel identified a subset of the elements that were most amenable to assessment in the form of written vignettes and responses. For example, while the use of multiple assessment methods was identified as an important element of reform-based mathematics, it was a practice that occurred over time and was not easily measured in the context of a specific classroom event. In contrast, “questioning”—i.e., eliciting information from students about why mathematical procedures are followed rather than telling students the algorithm for completing them—is an element that can be modeled in a specific teacher-student exchange.

Four classroom events were selected as likely situations in which reform-oriented practice would be manifest, and they were blended into a larger scenario that provided context for the whole set. Figure 1 shows the general framework of each scenario. We selected two mathematical topics to be the focus of the classroom vignettes. Area/perimeter and two-digit multiplication were chosen because they are common to almost all fourth-grade mathematics curricula, and they were covered extensively in the fourth-grade textbook used in the district participating in the study.

Figure 1. Scenario framework

Instructions. General instructions for vignette-based item.

Specific instructions for responding.

<u>SCENARIO TITLE</u>

Description of classroom context (e.g., grade level, student characteristics, topic area, prior knowledge)

1. First Problem: Introducing the Lesson

    Response options and answer grid

2. Second Problem: Responding to Student Error

    Response options and answer grid

3. Third Problem: Reconciling Different Approaches

    Response options and answer grid

4. Fourth Problem: Choosing Learning Objectives

    Response options and answer grid

The vignette-based section of the survey began with a general description of the scenarios and specific instructions about how to respond. Each scenario began with a brief context section that described the length and purpose of the unit and topics already covered in prior lessons. For example, in the multiplication scenario, teachers were told that they were beginning a one-week unit on two-digit multiplication, and that the students were familiar with place values and with multiplication facts up to 10 times 10. Teachers were told to assume that other features of the classroom that had not been described were similar to their current school and their current students.

Following the context description, each scenario contained four instructional problems that provided teachers with hypothetical situations at different stages within the unit. The first problem, "introducing the lesson," focused on the manner in which the teacher would begin the unit. Specifically, teachers were asked to indicate what kinds of activities they would use to introduce the concept. This was followed by several response options that were designed to include a range of teacher behaviors from less reform-oriented (e.g., finding out what prior knowledge students held related to the unit) to more reform-oriented (e.g., posing a novel problem that might lead students to discover key elements of the unit). We also included options that reflected traditional practices (e.g., demonstrating how to do the first procedure in the unit), which the expert panel believed were frequently taken by teachers. Figure 2 gives the "introducing the lesson" problem for two-digit multiplication, and provides selected examples from the 8 possible options presented to teachers.[1]

The second problem, "responding to student error," involved teachers' responses to a student mistake that occurred in the middle of the unit. For the two-digit multiplication vignette, we asked teachers to react to a student who inaccurately reasons that 23 times 41 is 400. Teachers were provided with 7 options that described various reactions, including low- and high-reform behaviors and common traditional practices (see Figure 3).

Figure 2. “Introducing the Lesson” problem from the two-digit multiplication scenario

You are ready to start the unit by introducing multiplication of two-digit numbers. How likely are you to do each of the following as an initial activity for the unit?

	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Demonstrate the standard algorithm for multiplying 2 two-digit numbers	1	2	3	4
Pose a problem that leads to a discussion of two-digit multiplication, such as how much you would earn in a year if your allowance is $15 per month	1	2	3	4

Note.
List of response options continues.

Figure 3. “Responding to Student Error” problem from the two-digit multiplication scenario

You ask your students to figure out how many miles Jerry can drive on 17 gallons of gas if Jerry’s car goes 23 miles per gallon of gas. You ask Leah if she will share her solution. She says “I multiplied 20 times 20 and got 400. This works because 23 is three more than 20 and 17 is three less than 20. So, Jerry can drive 400 miles.”
You know, however, that the answer is 391 miles. How likely are you to do each of the following?

	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Ask Leah if subtracting 3 miles per gallon is the same as adding 3 gallons of gas	1	2	3	4
Explain that 23 x 17 is actually 391. Then pose another similar problem	1	2	3	4

Note.
List of response options continues.

The third problem, “reconciling different approaches,” occurred near the end of the unit, and involved teachers’ reactions to groups of students who described two different approaches to solving a problem, both of which were correct, but differed in their efficiency. For the two-digit multiplication scenario, one group of students found the answer by applying the distributive property, while the other applied the standard algorithm. Teachers responded to a total of 5 options (see Figure 4).

The final problem, "choosing learning objectives," addresses the priority teachers would have given to different student objectives had they designed the unit (see Figure 5). Teachers were provided with 7 objectives including traditional goals as well as low-reform and high-reform goals. Respondents indicated the likelihood that they would emphasize each objective, so it was possible to endorse combinations of non-reform, low-reform and high-reform objectives.

For each problem, teachers were asked to rate the likelihood of engaging in each option using a four-point scale from "very unlikely" to "very likely." In the case of questions of emphasis, the scale indicated responses from "no emphasis" to "great emphasis."

The two scenarios were designed to be "parallel" in the sense that we presented teachers with similar instructional problems situated in different mathematical contexts, and with as comparable a set of response options as possible. However, the different topic of each scenario meant that some options had to be modified to fit the specific mathematical context. In addition, many scenario-specific changes were made on the basis of pilot testing, which also reduced the comparability. Thus, while the two scenarios are highly similar, they cannot be considered strictly parallel. Appendix B contains the complete area/perimeter scenario. The vignette-based items were administered as part of the survey, following the questions about professional development, curriculum and teaching practices.

Figure 4. “Reconciling Different Approaches” problem from the two-digit multiplication scenario

You ask students to work in pairs or groups to solve the following problem.

> Each school bus can hold 41 people. How many people can go on a field trip if the district has 23 buses available? Find the answer and demonstrate why it is correct.

You ask each group if they will share their work. The first group says the answer is 943, and they use the standard algorithm to show how they obtained this result.

$$\begin{array}{lr} & 41 \\ \text{x} & 23 \\ \hline & 123 \\ & 82 \\ \hline & 943 \end{array}$$

The second group says the answer is 943, and they explain that they broke the numbers into tens and ones and used the distributive property.

> 41 x 23 = (40 +1) x (20 + 3) = (40 x 20) + (40 x 3) + (1 x 20) + (1 x 3) = 800 + 120 + 20 + 3 = 943

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

	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Tell them they are both right and move on to the next problem	1	2	3	4
Have a classroom discussion about the differences between the two approaches	1	2	3	4

Note.
List of response options continues.

Figure 5. "Choosing Learning Objectives" problem from the two-digit multiplication scenario

If you were to design a unit on two-digit multiplication for your current class, how much emphasis would you place on each of the following learning objectives?

	No emphasis	Some emphasis	Moderate emphasis	Great emphasis
Students will be able to define the terms 'factor' and 'product'	1	2	3	4
Students will understand place value relates to two-digit multiplication	1	2	3	4

Note.
List of response options continues.

Classroom Observations. Observations were conducted in a subset of participating classrooms to provide an additional, independent perspective on instructional practice. To create an observation protocol, we identified 16 elements from our taxonomy of standards-based instruction that were thought to be readily observable key indicators of reform practices. These elements encompassed a variety of teacher and student activities, including the nature of student groupwork, the kind of tasks students worked on, the types of questions teachers asked, and the extent to which problem solving and conceptual understanding were promoted. The observational protocol also asked for judgments about the level of student engagement, the extent to which students appeared to understand the material presented in the lesson, and a holistic rating of the overall reform orientation of the lesson. To improve the consistency of the ratings, the protocol contained benchmark descriptions of high, medium, and low implementation for each element as well as examples of the kinds of behaviors that would characterize level. Observers first assigned a broad rating of high, medium or low, and then refined that rating into one of three sublevels, creating a 9-point scale. Raters also provided a written justification for each of their ratings.

Budget and time constraints limited the number and extent of observations we could conduct. Three members of the research team, who were familiar with the project goals and measures, served as observers. They were trained on how to use the observation protocol and independently rated three videotaped mathematics lessons (lasting about 30-minutes each). They reconvened, compared and discussed their ratings, and reached consensus on the use of the protocols.

We selected a convenience sample of 39 participating teachers in 13 schools based primarily on scheduling constraints. (A comparison of observed and non-observed teachers showed no significant differences with respect to teacher background characteristics or measured teaching practices.) Each classroom was observed by a single researcher for one full class period during a two-week observation window in the spring. The research took open-ended notes during the class and completed the rating form after the class period was over. The notes were used as a source of information and evidence for the written justifications. In one instance, two researchers observed the same classroom, and 80% of their independent ratings agreed within one point on the nine-point scale.[2] Based on the observational ratings, we developed two scales reflecting the teacher's emphasis on mathematical understanding and their overall use of reform-oriented practices.

Cognitive Interviews. Finally, we re-contacted participating teachers after the other data collection was complete and solicited volunteers for follow-up interviews to evaluate the validity of the vignette-based items using cognitive interviewing strategies. These techniques, which include think-aloud protocols and verbal probing, have commonly been used to explore whether there is a shared understanding of the constructs being measured by written survey items. In this case, the respondents were asked to read the vignette-based items and share their thought as they answered them. Following that, respondents were queried about their understanding of specific questions, their reasons for choosing or rejecting specific options, and the influence of certain contextual elements on their choices. These interviews were conducted with thirty mathematics and science teachers across each of the sites that were included in the larger study described earlier. The interview results shed some light on the strengths and weaknesses of the vignette approach, and they are used primarily to inform the discussion at the end of the paper. They are not part of the quantitative analyses.

Results

Response Rates

We received completed surveys, including the vignette-based items, from 80 of the 85 eligible mathematics teachers (94%) and completed logs from 78 of the 85 teachers (92%). We observed mathematics lessons in 39 of the classrooms (46%).

Scale Creation

For the purpose of validating the vignettes, we created scales from items related to reform-oriented instruction chosen from the four data sources. The set of scales, most of which focused on reform-oriented instruction, is listed in Table 2, along with illustrative items. These scales were selected to provide overlapping evidence about the extent of reform-oriented teaching that occurred in each class. Each of the underlying items provided teachers with a range of responses from less reform-oriented to more reform-oriented.

Survey, Log, and Observation Scales

The scales listed in Table 2 were created using a combination of empirical analyses (e.g., factor analysis), prior results with similar items in other studies, and expert judgments based on the underlying response options. We focused primarily on scales we thought to be key indicators of high-reform instruction but we also included one scale that measured traditional instruction because these practices were believed to be commonly occurring and because we had prior expectations about the relationship between traditional and reform practice. (The complete set of items that comprise each scale is presented in Appendix D.)

We developed two scales based on teachers' responses to survey items. The first, called Mathematical Processes, measured teachers' self-reported emphases on proof and justifications, problem solving, mathematical communication, connections, and mathematical representations.

Table 2

Summary of Teacher Practices Scales

Scale	Description	Illustrative Item
Survey		
Mathematical Processes	Extent to which teachers emphasized NCTM-endorsed mathematical processes	How much emphasis does the teacher place on proof and justification/verification (e.g. , using logical argument to demonstrate
Reform Practices	Frequency with which the teacher engaged in reformed instructional practices	How often does the teacher encourage students to explore alternative methods for solutions?
Log		
Discussion	Amount of time the class engaged in dialogue about mathematical thinking and	How much time did students spend explaining their thinking about mathematical problems?
Seatwork	Extent to which students engaged in seatwork and other low-reform activities	How much time did students spend reading from textbook or other materials?
Vignette		
Reform-High	Standardized sum of teachersÕ answers to the high-reform response options across th e two scenarios	Have a classroom discussion about the differences between the two approaches
Reform-Full	Euclidean distance of the teacher from the ideal high reform teacher	Have a classroom discussion about the differences between the two approaches; Tell them they are both right and move on to the next problem
Observation		
Mathematical Understanding	Extent to which the teacher facilitated deeper mathematical understanding	To what extent did the teacher provide incremental help (ŅscaffoldingÓ)so students could build understanding of complex relationships or learn new
Overall Reform	Extent to which the lesson embodied selected elements of reformed teaching	To what extent did the teacher facilitate discussion about ideas or con cepts among students?

These are aspects of mathematics emphasized by the NCTM *Standards*. The score on the Mathematical Processes scale represented the number of reform-oriented cognitive processes that had received moderate or great emphasis by the teacher. The values on this scale ranged from 0 to 5. The second scale, Reform Practices, measured the average frequency with which teachers engaged in nine specific reform-oriented instructional practices, including student learning activities, student assignments, and teacher questioning techniques. The potential values of this scale ranged from 1 to 5 (although the range of actual values was 2.6 to 4.4). The items on this scale are similar to those used in other research on mathematics reform, including some national longitudinal surveys (Cohen & Hill, 2000; Hamilton et al., 2003; Swanson & Stevenson, 2002; Weiss, et al., 2001; Wenglinsky, 2002).

Two scales were derived from the pair of daily logs completed by each teacher. The Discussion scale was based on the amount of time teachers reported that the class was engaged in a dialogue about mathematical thinking and understanding. The Seatwork scale described the amount of time students spent reading from the textbook or completing worksheets or problem sets. The activities on this latter scale are associated with more-traditional instruction, and we did not anticipate that they would correlate with the reform-oriented scales. Both scales ranged from 1 to 5.

Finally, two scales were based on the classroom observations we conducted in 39 of the classrooms. Mathematical Understanding was a composite based on five of the rating dimensions, which measured the extent to which the teacher facilitated deep and rich learning. Overall Reform was a broader composite of the observers' ratings including the 13 reform-oriented items on the observation protocol. The scores ranged from 0 to 8 for the observation scales.

Higher scores on each scale indicate more frequent use of the practices measured by that scale. For example, a score of 5 on the Seatwork scale indicates that students spent much of their time on activities such as reading from the text. In contrast, a score of 5 on a scale such as Discussion indicates that a relatively large proportion of class time was devoted to talking about mathematical thinking. Other scales are interpreted analogously.

Table 3 presents the mean, standard deviation, minimum, maximum, and internal consistency for each teacher practice scale. Results showed there was moderate variation in teachers' reported use of particular instructional practices. The scales had moderate to high levels of internal consistency, with coefficient alpha values ranging from 0.60 to 0.87. Because all classrooms (but one) were observed by a single researcher, we do not have a measure of inter-rater reliability for the observational scales.

Vignette Scales

The first task was to score the vignette-based questions and create scales to use in subsequent analyses. We hoped to build an overall judgment of reform orientation out of the separate judgments derived from each problem. That is, our goal was to characterize teachers along a single dimension of reform from highly reform-oriented to not very reform-oriented.

We used a judgmental process to assign a value from 1 (non-reform) to 4 (high reform) to each vignette response option. Although the vignette options were initially constructed to reflect particular levels of reform practice, these designations had to be revisited because many modifications occurred during the development process and because some options were added due to their perceived likelihood of occurring rather than their location on a hypothetical continuum of reform orientation. Traditional options were rated as non-reform. Members of the differences. The panel of expert mathematicians rated a comparable set of scenarios in

Table 3

Descriptive Statistics of Teacher Practices Scales

Scale	Mean	SD	Min	Max	Alpha
Survey					
Mathematical Processes	3.91	1.17	.00	5.00	.64
Reform Practices	3.71	.47	2.56	4.44	.73
Log					
Discussion	2.59	.67	1.25	4.38	.80
Seatwork	2.46	.80	1.00	4.25	.60
Vignette					
Reform-High	.00	1.00	-3.11	2.00	.80
Reform-Full	1.47	.54	.34	2.90	.86
Observation					
Mathematical Understanding	3.10	2.08	.20	7.80	.87
Overall Reform	2.60	1.63	.31	5.85	.87

research team independently rated each response option, then the group convened to reconcile a previous year, and we used the decision guidelines they established. In three instances, we could not agree on the value of an option, and these options were omitted from the ratings. In each case, the disagreement occurred because some researchers felt that the option, while not reform-oriented, *per se*, could be thought of as part of a larger reform strategy, but other researchers disagreed. Final ratings for each of the response options used in the area/perimeter vignette are provided in Appendix C.

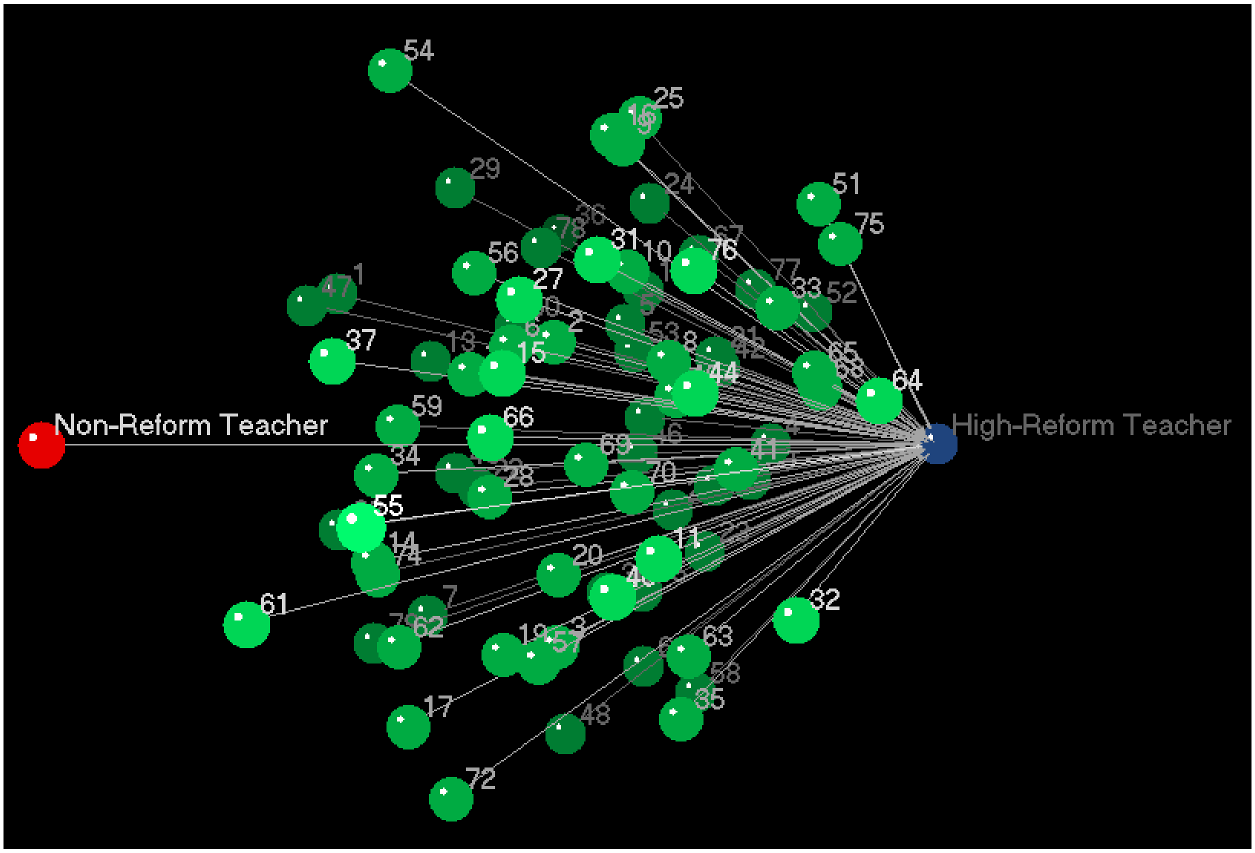
For the purposes of analysis, we considered options that had been rated a 3 or 4 to be indicative of high-reform teaching, and options that had been rated a 1 or 2 to be indicative of non-reform teaching. Using this criterion, of the 51 possible response options (across both scenarios and all eight instructional problems), 27 were high-reform options and the remaining 24 were non-reform options.

We used two different methods to derive vignette-based scores for each teacher from their responses to these 51 items. The first method, which used only the 27 high-reform options, was informed by examination of the correlations among problems and scenarios. This measure, which we refer to as Reform-High, is the standardized sum of teachers' answers to the high-reform response options across the two scenarios. Higher scores on this measure represent greater intentions to engage in reform-oriented practices.

The second method for creating vignette-based scores used all 51 separate response options (both high- and non-reform) from the 80 teachers. From the 80 (teacher) by 51 (response options) matrix, we created an 80 by 80 teacher similarity matrix, and used multidimensional scaling to plot the teachers in three-dimensional space. We included in this plot two additional points corresponding to a simulated ”consistently” high-reform teacher (whose self-reported likelihood of engaging in each option corresponded exactly to our judgments of reform orientation), and a simulated "consistently" non-reform teacher (whose self-reported likelihood was just the opposite). A graphical examination of the teachers' responses showed that teachers were generally more similar to our idealized high-reform teacher than they were to our idealized non-reform teacher, although there was also variation in responses (see Figure 6).

To examine the internal consistency of the teachers responses, we used the 80 (teacher) by 51 (response option) matrix to construct and plot a 51 by 51 response option similarity matrix in 3-dimensional space. The similarity between any two response-options is calculated by taking the Pearson correlation between the two response options across all 80 teachers. In this process, response options that received similar ratings from all teachers are clustered close together and response options that were rated differently among the teachers are clustered further apart. If teachers disagree about how they rated the items, we would expect to see reform and non-reform

Figure 6. Distribution of teachers' responses relative to ideal high- and non-reform teachers



items distributed randomly in the plot. If teachers were in complete agreement in how they rated the items, we would expect to see to see the items ordered along a single dimension. Although the teachers are far from complete agreement, the predicted order is apparent. The plot shows a gradient, with the most reform items on top and the non-reform items at the bottom.

Using the results of the multidimensional scaling of the teacher similarity matrix, we then created a scale of reform-oriented instruction by calculating the Euclidean distance between each teacher and the simulated consistently high-reform teacher (see Figure 6). We scaled this measure, which we refer to as Reform-Full so that teachers who are closer to the consistently high-reform teacher receive higher scores. That is, larger values of Reform-Full are associated with teachers whose responses are more like our consistently high-reform teacher. It is worth noting that the Reform-Full scale is the three-dimensional "distance" from the consistently non-reform teacher. Therefore, Reform-Full reflects variation on two additional, unnamed dimensions. Since the Reform-Full scale was derived from the full set of vignette options, which included non-reform behaviors, scores on Reform-Full may reflect some combination of reform-oriented and other practices.

Internal Validity of Vignette-Based Responses

One way to judge the validity of the vignette results as indicators of reform-oriented teaching is to examine the consistency of teacher responses to similar items from the two parallel scenarios. If reform-oriented teaching were a stable characteristic of practice, we would expect to find similar results in the two different mathematical contexts. That is, we would expect items that were designed to be comparable to elicit similar responses across the two scenarios. Each scenario had 27 separate response options (across the four instructional problems). After various revisions and deletions, 22 of these items remained reasonably parallel. We paired the 22 "parallel" response options, and computed weighted kappa statistics to assess similarity of ratings across scenarios. Kappa statistics ranged from low ($\underline{K}$ = -.04) to high ($\underline{K}$ = .69), with a median value of .25. Using the guidelines put forth by Landis and Koch (1977), we considered kappa values below .20 as "slight," from .21 to .40 as "fair," from .41 to .60 as "moderate," and higher than .60 as "substantial."

The distribution of the kappa statistics, shown in Table 4, indicated that teacher responses to some, but not all, item pairs were very similar. Kappa values tended to be higher for items drawn from the instructional problems "responding to student error" and "reconciling differing

Table 4

Distribution of Weighted Kappa Statistics for Paired Response Options (N = 22 pairs)

Kappa Value	Interpretation	Introducing the Lesson (n = 7)	Responding to Student Error (n = 7)	Reconciling Differing Approaches (n = 5)	Choosing Learning Objectives (n = 3)
.20 or below	Slight	3	2	1	2
.21 to .40	Fair	3	2	1	1
.41 to .60	Moderate	1	2	0	0
Above .60	Substantial	0	1	3	0

approaches” than for items relating to ”introducing the lesson” and “choosing learning objectives.” This suggested that items relating to teacher-student interactions elicited more consistency than did items relating to the structure or goals of the lesson. We also explored whether options that had been judged to be indicative of high-reform instruction tended to have higher (or lower) kappa values, but no clear pattern emerged.

The next step was to create aggregate measures of practice at the problem level by combining responses to the items that comprise each problem. Table 5 presents the correlations among the problem scores within and between the two scenarios. If each problem is measuring a stable and distinct aspect of instruction, we would expect the correlations between similar problems set within different mathematical contexts to be larger than the correlations between dissimilar problems set within the same mathematical context. Although there are exceptions, the patterns of correlations in Table 5 generally correspond to our expectations. For instance, the correlation between the “responding to student error” problems in the area/perimeter and the two-digit multiplication scenarios (r = .42) was higher than the correlations between the “responding to student error” and the other problems within the same scenario (r = .16-.25 for the area/perimeter scenario and r = .17-.41 for the two-digit multiplication scenario). This finding suggests that the problems within the vignettes measure stable aspects of reform-oriented teaching, and provide evidence for the validity of our proposed interpretation of the measures.

Table 5

Correlations Among Problem-Level Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area/Perimeter								
(1) Introducing the Lesson	1.0							
(2) Responding to Student Error	.25*	1.0						
(3) Reconciling Differing Approaches	.44**	.16	1.0					
(4) Choosing Learning Objectives	.22	.18	.22	1.0				
Two-Digit Multiplication								
(5) Introducing the Lesson	.41**	.13	.35**	.47**	1.0			
(6) Responding to Student Error	.35**	.42**	.34**	.25*	.17	1.0		
(7) Reconciling Differing Approaches	.29*	.06	.58**	.15	.38**	.20	1.0	
(8) Choosing Learning Objectives	.37**	.22	.29*	.43**	.48**	.41**	.26*	1.0

Notes.
Correlations between parallel scores are shaded.
* indicates statistical significance at the .05 level.
** indicates statistical significance at the .01 level.

Given the consistency of responses to items within problems, we combined the pairs of scores across the two scenarios to create four problem-level composite scores. For example, we added the introducing the lesson score from the area/perimeter scenario to the introducing the lesson score from the two-digit multiplication scenario to create a composite we labeled "introducing the lesson composite." We created the three other problem-level composites scores in an analogous manner. Our conception of reform-oriented teaching includes all four of these behaviors, and we would expect them to be related. Table 6 shows the correlations among these four problem-based composites, which range from 0.22 to 0.53. When we corrected the correlations for attenuation due to unreliability, the correlations are significantly higher, ranging from .34 to .87 with a median value of .59 (Allen & Yen, 1979). As a result of these moderate-to-high correlations, we felt justified in combining the four composites into a single measure of practice.

Table 6

Correlations Among Composite Problem-Level Scores

	(1)	(2)	(3)	(4)
(1) Introducing the Lesson Composite	1.0			
(2) Responding to Student Error Composite	.34**	1.0		
(3) Reconciling Differing Approaches Composite	.47 **	.52 **	1.0	
(4) Choosing Learning Objectives Composite	.53 **	.27 *	.22 *	1.0

Notes.
  * indicates statistical significance at the .05 level.
** indicates statistical significance at the .01 level.

Cross Method Comparisons

A second way to look at the validity of the vignettes is to examine the patterns of correlations between the vignette-based scales and the survey, log, and observation scales. We made *a priori* judgments about the reform-orientation of each of the survey, log, and observation scales based on theory and on past research. These judgments form the basis for predictions about the pattern of correlations that should exist among the measures. The degree to which the vignette-based measures show the expected relationships with the other measures is further evidence of the validity of the vignette-based scales as indicators of reform-oriented instruction.

Table 7 shows the correlations among the vignette-based scores and the survey, log, and observation measures. The top rows of the table show the predicted relationships among the different indicators of reformed teaching. With the exception of the Seatwork scale, all the survey, log and observation scales should be measuring aspects of reform-oriented instruction. Therefore, all the correlations among the scales should be positive. Seatwork is measuring an aspect of instruction that has been associated with more traditional practices, so the correlations of the reform measures with Seatwork should be zero. They are shown as zero in column 6 of the table.

Table 7

Predicted and Actual Correlations among Vignettes, Survey, Log, and Observation Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vignette-Based (N=80)								
(1) Reform-High	1.0	+	+	+	+	0	+	+
(2) Reform-Full	.15	1.0	+	+	+	0	+	+
Survey (N=79)								
(3) Mathematical Processes	.47 **	.08	1.0	+	+	0	+	+
(4) Reform Practices	.51 **	.15	.46 **	1.0	+	0	+	+
Log (N=79)								
(5) Discussion	.32*	-.01	.45 **	.32 **	1.0	0	+	+
(6) Seatwork	.06	-.02	-.03	.04	.13	1.0	+	+
Observation (N=38)								
(7) Mathematical Understanding	.09	.55 **	-.05	.08	.20	.00	1.0	+
(8) Overall Reform	.10	.55 **	.12	.15	.26	-.06	.91	1.0

Notes.
* indicates statistical significance at the .05 level.
** indicates statistical significance at the .01 level.

The remainder of Table 7 shows the actual correlations among the measures of reform practice. For the most part, the measures behave as predicted. Seatwork is uncorrelated with the other measures of reform practices, which is consistent with the premise that traditional and reform practices are orthogonal dimensions. The survey and log measures of reform are positively related. Similarly, the Reform-High scale behaves as predicted in comparison to the survey and log scales. Reading down column 1, the correlations between Reform-High and Mathematical Processes, Reform Practices, and Discussion are all positive. Similarly, the correlation between Reform-High and Seatwork is not statistically different from zero, as expected. These relationships provide further evidence of the validity of the Reform-High scale as an indicator of reform-oriented practice.

In comparison, the Reform-Full scale (column 2) is positively related to the two observational scales of reform-oriented practices and has a null relationship with the Seatwork scale, all as predicted. However, Reform-Full is not positively correlated with the other log and survey measures of reform-oriented practice. This pattern of relationships suggests that the Reform-Full scale and the observational scales may be measuring some aspect of practice that is not well represented in the Reform-High scale and the survey and log scales. We know that Reform-Full reflects more than a single dimension, and it may be that these additional elements relate to classroom interactions (like discourse) that are also present in the observation-based scales.

Discussion

Summary of Results

As interest in improving instructional practices grows, it becomes increasingly important to find effective ways of measuring classroom instruction. Some researchers have used familiar measures such as surveys and logs to good effect (Porter, et al, 1993; Ball, et al., 1998), while others have identified limitations in these measures (Burstein, et al., 1995; Mayer, 1999). More recently, researchers have developed methods to improve their quality (Desimone & LeFloch, 2004). The goal of this study was to evaluate whether closed-ended, vignette-based measures could provide a new tool for measuring important aspects of teaching practice. The study focused on the construct "reform-oriented instruction," and caution should be used when drawing inferences to other aspects of instruction as well.

Teachers' responses to our vignette-based prompts appear to measure relatively stable aspects of teaching. That is, responses to common problem situations remained fairly consistent across both scenarios, which were set in different mathematical contexts. Responses to the "interactive" elements of practice (responding to errors, reconciling differences) were more consistent than responses to the "structural" elements (introducing topics and setting objectives).

We also found that the Reform-High measure, derived from only the "high-reform" options, showed the expected positive correlations with the reform-oriented survey and log measures and no relationship with Seatwork. The Reform-Full measure, which was derived from all the vignette responses and reflected more than just the reform dimension, showed the expected positive correlations with the observational scales and no relationship with Seatwork. However, the Reform-High measure (and the reform-oriented survey and log measures that related to it) and the Reform-Full measure (and the observational scales that related to it) were not positively related to one another.

Explaining the Observed Relationships

Interactive versus structural classroom problems. One interpretation of these results is that the "interactive" aspects of teaching are more firmly established in teachers' practice. The bulk of teachers' time is spent interacting with students, so it is reasonable to think that teachers' interactive behaviors (i.e., working with students) would be more routinized than their behaviors in other aspects of teaching, and their responses to hypothetical questions would be more consistent. Another possibility is that the "interactive" domains we selected were narrower in some sense than the "structural" domains we selected. To the extent that these inferences are correct, it suggests that vignette-based items might be more effective if they target "common" teacher behaviors with "narrow" ranges of options. However, we are unable to say precisely what constitutes "common enough" or "narrow enough."

Reform-High versus Reform-Full. The fact that Reform-Full and Reform-High have different relationships to the survey, log, and observational scales was surprising to us, and it suggests that they are measuring different constructs. Although the two vignette-based measures were derived from the same set of teacher responses, they were constructed differently, and, as noted previously, they may reflect different aspects of instructional practice. Reform-High was constructed using only those response options deemed to be indicative of high-reform instruction. It is a linear combination of these elements, much like the survey and log based measures are linear combinations of responses to multiple reform items. Reform-Full, in contrast, is a multidimensional scale, and it reflects more than just "reform-orientation." Although we do not know what elements of practice characterize the other dimensions that are present in Reform-Full, it is possible that they were also apparent to observers and are present in observers' ratings of practice.

Another possible explanation is that the measures are sensitive to different levels of reform-orientation. Reform-High was constructed from only the high-reform options in the vignettes, and it is most sensitive to differences at the high end of the reform scale, i.e., to differences between teachers who endorse many reform-oriented behaviors and those who endorse some. The survey and log scales were similar to Reform-High in that they focused primarily on high-reform activities. In contrast, the observational scales and Reform-Full both included some elements that captured non-reform practices, and they may have been more sensitive at the lower end of the scale. Although this is quite speculative, such differences in the construction of the scales may explain the different patterns of correlations with other indicators of reform practices.

Observer ratings versus teacher reports. We also do not have a simple explanation for the lack of strong correspondence between observer ratings and teacher reports (on logs and surveys), but there are a number of possibilities. First, there is some reason to believe that the differences may have stemmed from our sampling strategy. We sampled teachers from districts that were selected to participate in the larger study because they were providing training to promote the use of reform-oriented practices. We felt this was essential to make sure there was adequate variation in the use of these practices. (Others have reported little or no use in the absence of specific training). However, one consequence of the training may have been heightened sensitivity to the language of reform and the practices of reform that biased survey and vignette responses. Respondents may have recognized reform behaviors (described in neutral terms) and responded in ways they assumed were socially desirable.

Additionally, the issue of time sampling is relevant to results from logs, surveys and observations. In this study, we were limited by practical constraints to a single, annual, data-collection cycle. In theory, collecting all the information about teaching practices at roughly the same time has some advantages. For example, it reduces variation in responses due to differences in curriculum content. Variations due to time may play a significant role in responses to logs and observations because both were designed to capture information about events in a narrow period of time. Ideally, both logs and observations would reflect the same lesson, but that degree of correspondence rarely occurred in the study. However, the log and observational data usually reflected the same unit of instruction. The survey items were retrospective, covering the whole school year, and the vignette items were hypothetical, so perhaps it was less important that they be collected at the same time as the other measures.

Researcher versus teacher. Differences among the measures in this study may also reflect differences in the perspectives of researchers and teachers. Research has also shown that teachers and observers ascribe different meanings to practices associated with reform instruction. Some teachers may perceive themselves to be engaging in reform-oriented activities, while observers judge their teaching to be substantively more indicative of non-reform pedagogy (Cohen, 1990; Ingle & Cory, 1999; Mayer, 1999). Indeed, we have some evidence that observers and teachers interpreted the same events differently. Brief interviews with teachers immediately following the observations confirmed that teachers and observers sometimes had different interpretations of reform teaching. One teacher, for example, believed that she engaged in reform-oriented pedagogy because her test items did not use a multiple-choice format and were therefore "open-ended." The observer, however, found the test items to be very structured, requiring only a brief response as the correct answer.

The observational scales may also have been influenced by differences in teachers' and researchers' frames of reference. On average, teachers reported using reform activities about once or twice a week, but our observers viewed only a single lesson. Because we conducted classroom observations on a single day, we necessarily sampled a very small portion of teachers' practices (Shavelson, Webb, & Burstein, 1986). It is impossible to determine whether the practices we observed capture teachers' typical approach, or are instead an unrepresentative sample of their instruction.

Proximity. As we noted at the outset, the four measures used in this study differed in terms of their proximity to instruction. It may be the case that proximity explains some of the similarities and differences we observed. If proximity was having an influence, we would expect responses from observations and logs which were the closest to actual events to be the most consistent, while survey and vignette-based scales would be less so. However, that is not what we found, and the proximity dimension does not seem to have a large effect on the results.

Instead, we suggest that "task format" may play a significant role, and we have identified three format dimensions that may have influenced responses. The first dimension is response format. Survey items were closed-ended items with four or five ordered choices, as were the vignette-based items. The log items used in this study were similar to these, although the logs also contained open-response items. The observation scales required ratings on a scale with three defined anchor-points and six levels of interpolation. To the extent that response format affects responses, we would expect the observational results to differ from the other three.

The second dimension is the cognitive demand of the response. The surveys called for long-term memory about the frequency of past events or the extent of past practice. The logs called for short-term memory about the frequency of recent events or the extent of recent practice. The observation ratings involved judgments against a descriptive criterion based on short-term memory and written notes. The vignettes called for projections of potential behavior in hypothetical situations. They simultaneously evoke the complex, multidimensional nature of teacher (by contextualizing the setting and encouraging teachers to think about all the elements that determine their classroom behaviors) and request a single-dimensional response. This contrast challenges respondents in ways we do not fully understand. All four measures were different in terms of cognitive demand, and it is reasonable to think that these differences might influence responses. Unfortunately, we know of no research that would predict how these differences might have influenced responses.

Third, the measures differed in terms of the respondents themselves. The surveys, logs, and vignettes were completed by teachers, whereas the observations were completed by external observers. As a result, the observational scales may differ from the others due to the influence of the respondent.

Each of these three “format” dimensions may account for difference among the responses in this study, and, further research would be needed to try to untangle the potential interrelationships. While educators are often encouraged to use multiple measures to provide more reliable information, researchers should be aware that multiple formats and multiple respondents can reduce consistency of responses by adding additional sources of variation.

Lessons Learned about Vignette Development

Our experience sensitized us to the difficulties of developing the vignettes in a systematic manner. Expert advisors helped us identify the instructional problems, so we had some confidence in the underlying framework. However, we had no experience creating similar problem situations in different topic settings that admitted “parallel” teacher options. For example, it took many false starts before we identified similar "student errors" relating to two-digit multiplication and perimeter and area, which could then be followed by equivalent sets of teacher responses reflecting a range of reform-oriented practice. Much more research needs to be done to develop a methodical and efficient approach to the creation of this type of vignette.

Vignettes also appear to be more difficult to develop than surveys or logs. Not only was it difficult to generate appropriate hypothetical situations and teacher actions, but because they were tightly linked to curriculum, vignettes written for one grade level were not necessarily appropriate for another. As a result, we had to develop new vignettes each year, while the other survey questions could be reused with only small changes. The need to link the vignettes to the curriculum also required the development of different vignettes for different districts, even at the same grade level. On the basis of this study, it is clear that closed-ended vignettes are far more costly to develop than are surveys covering similar domains.

Level of detail. We attempted to specify aspects of the teaching situation that were likely to affect teachers' instructional decisions (e.g., prior learning history of the students), but our vignettes were brief and left much unstated. For example, we rarely provided information about the affective and psychological characteristics of students or the physical characteristics of the setting. Including only a few details may provide an inadequate frame for teachers to use to project their behavior. Indeed, in cognitive interviews after the completion of the main data collection, teachers indicated that they usually adapted their instruction to the abilities of their students, and the lack of information about ability levels made it difficult for them to respond to some of the options. Teacher also said they wanted to know how much time was left within the lesson because they said this would affect their choice of actions.

A somewhat paradoxical problem is that providing more details can actually make the vignettes more distant from instruction, i.e., the tradeoff between detail and verisimilitude is complex. On the one hand, it would seem that more detail would make the vignettes more realistic; on the other hand, more details provide more instances in which the situation may vary from the teacher's own classroom. Thus, the effort to make the vignette proximal may actually make it more distant from the teacher's familiar classroom setting.

Longer and more detailed vignettes are also less desirable because of the additional reading burden they place on respondents, and offer additional opportunities to misread or misinterpret the prompt. Hill (2005) pointed out that even something as apparently well understood as mathematical terminology can be interpreted differently by researchers and teachers. In our own cognitive interviews, we found evidence that teachers interpreted options in ways that we did not intend, despite the fact that we attempted to frame the descriptions in neutral, behavioral terms. For instance, teachers assumed that response options that depicted teachers providing an explanation would ostensibly be followed by student discourse despite the fact that the latter activity was not described. More glaringly, teachers sometimes reported that they did not like the way we phrased an option and so they rephrased it in their own terms and responded to the version they had created. That we observed these types of unintended interpretations from brief descriptions of behaviors leads us to believe that more extended descriptions may increase the opportunity for miscommunication.

Quantity versus quality. Another unresolved issue in the use of vignettes (and other commonly used measures of practice) is the confounding of quantity and quality. The scales we constructed for this study all were based on quantity (more frequent use of a reform-oriented practice) with no consideration of quality (how effectively the practice was used). For example, the vignette-based measures assigned higher scores when behaviors were rated as more likely to occur. As researchers, we understand that quantity is not the same as quality, but we find it much easier to quantify than to make qualitative judgments. Experience suggests that quality can be incorporated into observational ratings more easily than into any of the other methods, although training observers to rate quality is complicated. Vignettes may hold some promise for incorporating quality into measures of instructional practice but that is a distant possibility.

Future Research Directions

Overall, we are cautiously optimistic about the use of closed-ended vignettes to provide valid information about instructional practice. Interviews with teachers indicated they found the tasks engaging and realistic. Despite the difficulties we encountered in constructing the tasks for the first time, we think it is possible to develop systematic methods for task generation. One key to making this approach work is to put more effort into building the task generation frameworks. We recommend focusing on common instructional practices, simplifying the situations as much as possible, and including information about student abilities. We also suggest involving teachers earlier and more extensively in the process. For example, in developing closed-ended vignettes, it may make sense to begin with open-ended vignettes and generate the response option framework based on actual responses from teachers (and other knowledgeable educators).

As noted earlier, it is unclear whether providing more detailed descriptions of teaching situations make vignettes more or less realistic for teachers. Thus, further investigation of length, level of detail, and scope of description in vignettes is warranted. Additionally, the generalizability of vignettes needs to be explored. Vignette-based questions take more time to read and respond to than typical survey items, so the range of teaching situations that can be portrayed in a handful of vignettes is limited. This may limit the generalizability of the results compared to the universe of actual teaching situations. In our study, we reduced the universe of generalization by restricting the range of practices to reform-oriented mathematics instruction. Such a restriction improved the quality of inferences we could draw regarding the prevalence of reformed teaching in a district that is actively promoting reform-oriented practices. It remains to be seen whether vignettes can be used to measure instruction more generally, as teachers use a combination of approaches—reform, traditional, hands-on, abstract, student-centered, didactic, and so forth—and effective teaching may be a function of the balance among these. Additional research needs to be done before we are able to say which constructs, if any, are best measured using vignettes.

Before making any final decision about the validity of descriptions derived from vignettes, it would also be important to study the relationships of vignette-based measures to student outcomes, particularly achievement. Our purpose in developing these vignette-based questions was to measure aspects of reformed teaching that were not captured by surveys and logs but that helped to explain student achievement. We are currently studying how the vignette-based responses relate to changes in student achievement, and these data will provide another indication of the usefulness of vignettes as measures of instructional practices.

Notes

[1] In the interest of space, the instructional problems are not presented in their entirety.

[2] The weighted kappa statistic for these two sets of ratings was 0.66. For the purpose of analysis, the two ratings for this teacher were averaged.

[3] We previously collected daily logs from teachers for a five day period, but found it did not provide substantially different information than was obtained in two days.

References

Allen, M. J, & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.

Antil, L.R., Jenkins, J.R., Wayne, S.K., & Vasdasy, P.F. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. *American Educational Research Journal, 35*, 419-454.

Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Ball, D.L., Camburn, E., Cohen, D., & Rowan, B. (1998). *Instructional improvement and disadvantaged students*. Ann Arbor, MI: University of Michigan.

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

Cohen, D.K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 14*, 327-345.

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

Desimone, L. M., & LeFloch, K. C. (2004). Are we asking the right questions? Cognitive interviews to improve surveys in educational research. *Educational Evaluation and Policy Analysis*, 26(1), 1-22.

Ferguson, R.F., & Ladd, H.F. (1996). How and why money matters: An analysis of Alabama schools. In H. Ladd (Ed.), *Holding Schools Accountable* (pp. 265-98). Washington, DC: Brookings Institute.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley.

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

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

Hill, H. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy,* 19(3), 447-475.

Ingle, M., & Cory, S. (1999). Classroom implementation of the national science education standards: A snapshot instrument to provide feedback and reflection for teachers. *Science Educator, 8*, 49-54.

Kennedy, M.M. (1999). Approximations to indicators of student outcomes. Educational *Evaluation and Policy Analysis, 21* (4), 345-363.

King, B., & Newmann, F. (2000). Will teacher learning advance school goals? *Phi Delta Kappan, 81* (8), 576–580.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000). *Teaching practices and student achievement*. MR-1233-EDU. Santa Monica: RAND.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22* (2), 18-26

Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Ma, Liping. (1999). Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States. New Jersey: Lawrence Erlbaum Associates.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*, 29-45.

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

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

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

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

Porter, A.C., Kirst, M.W., Osthoff, E.J., Smithson, J.L., & Schneider, S.A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms. Final report to the National Science Foundation*. Madison, WI: Wisconsin Center for Educational Research.

Rowan, B. Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *The Elementary School Journal,* 105(1), 75-101.

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

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal*, 105(1), 103-127.

Ruiz-Primo, M. A., & Li, M. (2002). *Vignettes as an alternative teacher evaluation instrument: An exploratory study.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans (April 1-5).

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.* Knoxville, TN, University of Tennessee.

Shavelson, R.J. Webb, N.M., & Burstein, L. (1986). Measurement of teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York Macmillan.

Sheeran, P., S.Orbell & D.Trafimow. (1999). Does the temporal stability of behavioral intentions moderate intention-behavior and past behavior-future behavior relations? *Personality and Social Psychology Bulletin, 25* (6), 721-730.

Stigler, J.W. & Perry, M. (2000). *Developing classroom process data for the improvement of teaching.* In Raju, N.S., Pellegrino, J.W., Bertenthal, M.W., Mitchell, K.J., Jones, L.R. (Eds.), *Grading the Nation's Report Card: Research from the Evaluation of NAEP* (pp. 229-264). Washington, DC: National Academy Press.

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

Weiss, I. R., Banilower, E. R., McMahon, K. C., and Smith, P. S. (2001). *Report of the 2000 national survey of science and mathematics education*. 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* (12). Retrieved December 12, 2003 from http://epaa.asu.edu/epaa/v10n12/.

Wenglinsky, H. (2003). Using Large-Scale Research to Gauge the Impact of Instructional Practices on Student Reading Comprehension: An Exploratory Study. *Education Policy Analysis Archives, 11* (19). Retrieved December 12, 2003 from http://epaa.asu.edu/epaa/v11n19/.

Wright, S., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11* (1), 57-67.

Appendix A. Elements of Standards-Based Mathematics

I. Nature of mathematics (What aspects of mathematics should be emphasized?)
A. Content/topics
Depth. Explore fewer topics in greater depth rather than covering more topics quickly or superficially.
Complexity. Explore more complex problems rather than only simple problems that emphasize only one concept or skill.
Breadth. Emphasize significant reoccurring themes in mathematics rather than only addressing specific skills.
Relevance. Select topics that help students connect mathematics to their own experience and the larger mathematics community rather than understanding mathematics as isolated skills or procedures.
Connections within mathematics. Give more emphasis to the inter-relatedness of mathematical concepts than to presenting each skill in isolation.
Integration across subjects: Apply mathematics in authentic ways in the context of other subjects rather than in isolation from other content.
B. Procedures/processes
Reasoning/problem solving. Place greater emphasis on reasoning and problem solving than on operations and computation.
II. Student mathematical thinking (What behaviors effective student mathematicians should be fostered?)
Active thinking. Focus lessons on reasoning processes rather than only on obtaining the right answer.
Meta-cognition. Help students monitor and evaluate their own problem solving and evolve more sophisticated mathematical thinking rather than leaving thinking procedures unexamined.
Curiosity. Encourage students to take risks in exploring problem-solving strategies rather than only trying the most familiar approach.
Solution process. Encourage students to identify more than one reasonable problem solving strategy: to double-check results, to learn which is more efficient, and to reveal the underlying mathematics rather than obtaining answers in only one way.
III. Mathematics teaching (How should mathematics be taught?)
A. Classroom Discourse
Task selection. Present rich activities (reflecting the content of the mathematical standards) which challenge students, provoke discussion and develop mathematical power rather than simple tasks that demand only simple thinking.
Community of Practice. Create an environment where students and the teacher engage in a dialogue about the mathematical content and processes rather than one in which the teacher is the primary source of knowledge and insight.
Questioning. Emphasize “why” things are done rather than “how” to do them to draw out students’ mathematical thinking as a basis for discussion.
Critical listening and explaining. Have students critique/explain their reasoning and the reasoning of others rather than not engaging in such analysis.
Representation/Communication. Encourage student to represent and communicate about mathematics in many ways rather than only recording results by writing numbers (e.g., orally or pictorially). (This helps to illuminate their mathematical thinking and to relate their ideas to larger mathematical concepts.)
Progression. Provide a clear and coherent progression of ideas rather than disconnected pieces.
Scaffolding. Present information and ask questions incrementally based on students’ responses rather than explaining complete procedures. (This type of support encourages students to continue developing and refining their mathematical thinking.)

Larger mathematical context. Foster conversations that address both the immediate problem in the local context as well as the larger problem and the historical mathematical context rather than only exploring problems narrowly. (This relationship between the immediate subject and the larger mathematical ideas plays out over time.)
B. General Strategies
Assessment. Use assessments in a variety of formats that reflect the mathematical content and processes taught as well the desired student mathematical thinking rather than using one or two limited assessment formats.
Equity. Help all students achieve similar understanding by using alternative instructional strategies and representations when warranted and by eliciting/presenting multiple ways of solving problems when appropriate rather than offering only one way for students to learn or assuming students learn in a particular way.
Grouping. Have students work in a variety of groups rather than doing the bulk of their assignments alone. (This approach helps students learn to communicate mathematical ideas and integrate other students' knowledge and skills into their own thinking.)
Materials. Have students use a variety of materials to solve mathematical problems rather than just pencil and paper and chalkboard.

Appendix B. Area/Perimeter Scenario

Teaching Scenarios

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

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

Please do the following:

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

SCENARIO I. AREA AND PERIMETER (4 QUESTIONS)

Imagine you are teaching mathematics to a fourth-grade class. The students are familiar with the operations addition and multiplication, and they know the meaning of perimeter. You are about to begin a week-long unit on area.

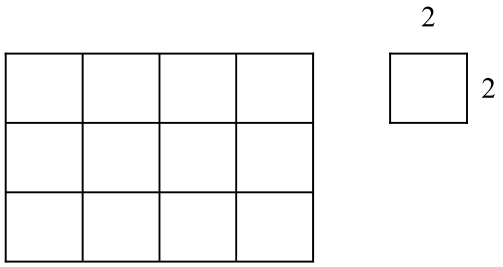
1. You are ready to start the unit by introducing the concept of area. How likely are you to do each of the following as an **initial activity** for the unit?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask students what they know about area	1	2	3	4
b. Have students work in groups to measure area, e.g., finding the number of square tiles that are needed to cover the desks in the classroom	1	2	3	4
c. Demonstrate how to compute the area of a rectangle by using the formula A = L x W (i.e., the product of the length and width)	1	2	3	4
d. Review students' knowledge of perimeter	1	2	3	4

e. Have students use geoboards to construct figures and determine their areas	1	2	3	4
f. Pose a problem that leads to a discussion of area, such as which of two shapes with equal perimeters students think is bigger	1	2	3	4
g. Lead a classroom discussion about the methods for determining the area of rectangular-shaped objects (e.g., parks of different sizes)	1	2	3	4
h. Define area as the number of square units that cover a surface.	1	2	3	4

2. You are at the midpoint of your unit on area, and most students appear to understand how to compute the area of a rectangular-shaped figure. You ask your students to compute the area of the following shape, which is made up of 2 x 2 squares.



When most students appear to have completed the task, you ask Chris if he will share his solution. He says “The area is 24. I multiplied 4 wide times 3 high and got 12, and then I multiplied by 2 to get 24.”

You know, however, that the area is 48 square units.

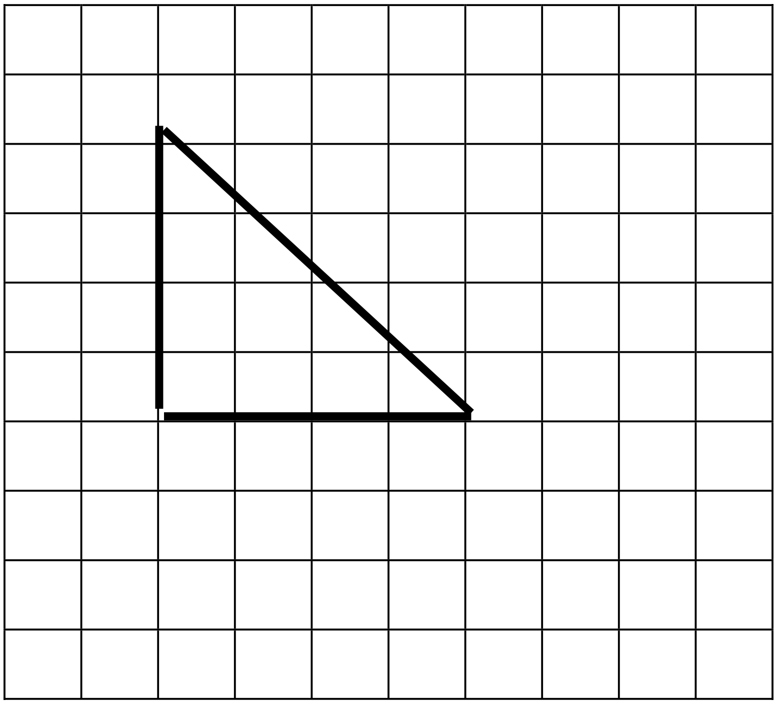
After praising Chris for remembering the formula for area, what do you do next? How likely are you to do each of the following?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask Chris what are the actual length and width of the rectangle	1	2	3	4
b. Leave the problem unresolved for a moment, and ask Chris to calculate the area of a single 2 by 2 square	1	2	3	4
c. Suggest that Chris use graph paper or a geoboard to solve the problem	1	2	3	4
d. Remind Chris that each square in the rectangle is 2 by 2 so the area of each square is four square units	1	2	3	4
e. Explain that figure is actually 8 units wide and 6 units high, so its area is 48 square units. Then pose another similar problem	1	2	3	4
f. Ask the class if anyone can explain how s/he obtained a different answer	1	2	3	4
g. Ask Chris to demonstrate how he solved the problem, using either concrete objects or a picture	1	2	3	4

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

   What is the area of the figure drawn below?



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

> The first group responds that the area is 8 square units. They explain that they counted the squares inside the triangle and had 6 whole squares and 4 half squares. Together that makes 8 square units.
>
> The second group says they got the same answer but did it differently. They computed the area of the 4x4 square containing the triangle and then took one-half of it. The area of the square is 4 times 4 or 16 square units, so the area of the triangle must be 8 square units.

After you praise both groups for using effective strategies, how likely are you to do each of the following in response to the two explanations?

(Circle One Response in Each Row)

	Very unlikely	**Somewhat unlikely**	**Somewhat likely**	**Very likely**
a. Ask the class if they can think of another way to solve the problem	1	2	3	4
b. Suggest that the class check the result by counting the area of the triangle that results from the other half of the square	1	2	3	4
c. Tell them that the second group's method is easier because it is sometimes difficult to count the parts of the squares	1	2	3	4
d. Tell them they are both right and move on to the next problem	1	2	3	4
e. Have a classroom discussion about the differences between the two approaches	1	2	3	4

4. If you were to design a unit on area for your **current** class, how much emphasis would you place on each of the following learning objectives?

(Circle One Response in Each Row)

	No emphasis	**Slight emphasis**	**Moderate emphasis**	**Great emphasis**
a. Students will be able to write the formula for calculating the area of a rectangular figure	1	2	3	4
b. Students will be able to use geoboards and graph paper to represent area	1	2	3	4
c. Students will be able to compute the area of squares and rectangles given the lengths of their sides	1	2	3	4
d. Students will be able to extend what they have learned to calculating the area of irregular shapes	1	2	3	4
e. Students will be able to solve applied problems that use area, such as figuring out the amount of tile needed to cover a rectangular floor	1	2	3	4
f. Students will understand the relationship between area and perimeter	1	2	3	4
g. Students will be able to construct shapes of a given area	1	2	3	4

Appendix C. Reform Ratings of Response Options Presented in the Area/Perimeter Scenario

Item No.	Rating
1a	3
1b	3
1c	1
1d	2
1e	3
1f	4
1g	3
1h	1
2a	3
2b	2
2c	2
2d	2
2e	1
2f	3
2g	4
3a	3
3b	2
3c	2
3d	1
3e	4
4a	N/A
4b	2
4c	N/A
4d	4
4e	3
4f	4
4g	2

Appendix D. Scale Items

Survey

Mathematical Processes

How much emphasis do you place on each of the following (no emphasis, slight emphasis, moderate emphasis, great emphasis)?

Proof and justification/verification (e.g., using logical argument to demonstrate correctness of mathematical relationship)

Problem solving (e.g., finding solutions that require more than merely applying rules in a familiar situations)

Communication (e.g., expressing mathematical ideas orally and in writing)

Connections (e.g., linking one mathematical concept with another; applying math ideas in contexts outside of math)

Representations (e.g., using tables, graphs and other ways of illustrating mathematical relationships)

Reform Practices

On average throughout the year, approximately how often do you employ the following teaching strategies during your mathematics lessons (never, a few times a year, once or twice a month, once or twice a week, almost every day)?

Use open-ended questions

Require students to explain their reasoning when giving an answer

Encourage students to communicate mathematically

Encourage students to explore alternative methods for solutions

Help students see connections between mathematics and other disciplines

On average throughout the year, approximately how often do your students take part in the following activities as part of their mathematics lessons (Never, a few times a year, once or twice a month, once or twice a week, almost every day)?

Share ideas or solve problems with each other in small groups

Engage in hands-on mathematics activities

Work on extended mathematics investigations (a week or more in duration)

Record, represent, or analyze data

Log

Discussion

How much time did students spend on each of these activities during today's mathematics lesson (none, 1-5 minutes, 6-10 minutes, 11-20 minutes, 21 minutes or more)?

Explain their thinking about mathematical problems

Lead discussion of a mathematics topic

How much time did you spend on each of these activities during today's mathematics lesson (none, 1-5 minutes, 6-10 minutes, 11-20 minutes, 21 minutes or more)?

Ask open-ended questions and discuss solutions to them

Ask questions of individuals to test for understanding

Seatwork

How much time did students spend on each of these activities during today's mathematics lesson (none, 1-5 minutes, 6-10 minutes, 11-20 minutes, 21 minutes or more)?

Read from textbook or other materials

Complete worksheets or problem sets from text

Observation

Mathematical Understanding (9-point scale, with descriptive anchors for high--6,7,8; medium—3,4,5; and low—0,1,2)

To what extent did the lesson/teacher focus on/guide toward conceptual understanding?

To what extent did the lesson/teacher focus on/guide reasoning and problem solving?

To what extent did the teacher engage in scaffolding to help students make large conceptual or procedural jumps (e.g., asking questions incrementally based on students response)?

To what extend did the students work on complex problems that involve multiple steps, multiple solutions or assess more than one mathematical skill or concept?

To what extent did the lesson have a sequence of activities that build conceptual understanding?

Overall Reform (9-point scale, with descriptive anchors for high--6,7,8; medium—3,4,5; and low—0,1,2)

To what extent did the lesson/teacher focus on/guide toward conceptual understanding?

To what extent did the lesson/teacher focus on/guide reasoning and problem solving?

To what extent did the teacher connect the lesson to other topics in mathematics?

To what extent did the teacher connect the lesson to other subject areas or disciplines?

To what extent did the teacher engage in scaffolding to help students make large conceptual or procedural jumps (e.g., asking questions incrementally based on students response)?

To what extent did the teacher encourage students to come up with more than one way of solving the problem?

To what extent did the teacher facilitate discussion about ideas or concepts among students?

To what extent did the teacher stress the relevance of mathematics to the students' own lives?

Did lessons involve a variety of pictures, graphs, diagrams, or other representations to illustrate an idea or concept?

To what extent did the students work cooperatively in groups?

To what extent did the students participate in hands-on activities using manipulatives that was related to mathematical ideas or concepts?

To what extend did the students work on complex problems that involve multiple steps, multiple solutions or assess more than one mathematical skill or concept?

To what extent did the lesson have a sequence of activities that build conceptual understanding?

Author Note

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