Artifact Packages for Measuring Instructional Practice: A Pilot Study

CSE Report 615

Brian M. Stecher and Alicia Alonzo
RAND

Hilda Borko, Shannon Moncure, and Sherie McClam
University of Colorado, Boulder

December 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analysis of Current Assessment and Accountability Systems, Strand 1: The Impact of State Accountability on Classroom Practice
Brian M. Stecher, Project Director, RAND

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences or the U.S. Department of Education.

ARTIFACT PACKAGES FOR MEASURING INSTRUCTIONAL PRACTICE: A PILOT STUDY

Brian M. Stecher and Alicia Alonzo
RAND

Hilda Borko, Shannon Moncure, and Sherie McClam
University of Colorado, Boulder

Abstract

A number of educational researchers are currently developing alternatives to survey and case study methods for measuring instructional practice. These alternative strategies involve gathering and analyzing artifact data related to teachers' use of instructional materials and strategies, classroom learning activities, and students' work, and other important features of practice. "The Impact of Accountability Systems on Classroom Practice" is one such effort. The goals of this 5-year project, funded through the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), are to develop artifact collection and scoring procedures designed to measure classroom practice in mathematics and science; validate these procedures through classroom observations, discourse analysis, and teacher interviews; and then use the procedures, in conjunction with other CRESST projects, to conduct comparative studies of the impact of different approaches to school reform on school and classroom practices. The first phase of the project was a set of pilot studies, conducted in a small number of middle school science and mathematics classrooms, to provide initial information about the reliability, validity, and feasibility of artifact collections as measures of classroom practice. This report presents the results of these pilot studies.

Project Goals and Rationale

Information about instructional practice is important for at least three reasons. First, teachers play a key role in determining the success of reform efforts. Their actions mediate the impact of changes in components of educational reforms such as accountability systems, curriculum programs, and instructional approaches on student achievement. As Fullan & Miles (1992) noted, "Local implementation by everyday teachers, principals, parents, and students is the only way that change happens" (p. 752). Spillane (1999) made a similar argument: "While policy makers and reformers at all levels of the system are crucial if these reforms are to be enacted locally, teachers are the key agents when it comes to changing classroom practice.

They are the final policy brokers" (p. 144). Thus, first, information on teachers' instructional practice is key to understanding why programs of reform succeed or fail. Second, information about instructional practice provides evidence that is relevant to judgments about the validity of test scores and score gains. Better measures of instructional practice can help us to understand what happens under the broad heading of "teaching to the test" and can reveal specific classroom activities that may affect inferences from test scores to the broader domain they are supposed to represent (Borko & Elliott, 1999; Koretz, Stecher, Klein, & McCaffrey, 1994; Stecher, Barron, Chun, & Ross, 2000; Wolf & McIver, 1999). Third, higher state standards demand more, not only of students, but of teachers as well. Many of these standards call for core changes in instructional practices that teachers may not be prepared to incorporate in their classrooms (Firestone, Mayrowetz, & Fairman, 1998). To help teachers develop the capacity to prepare students to meet the higher standards, it is important to have reliable measures of instructional practices that can inform improvements in teacher education and professional development programs.

Some educational scholars have suggested that reform efforts of the 1970s and 1980s were unsuccessful, at least in part, because they remained aloof from curriculum and teaching practices, focusing instead on factors such as resource allocation and outcome goals (Mayer, 1999). Not surprisingly, by the 1990s policymakers were beginning to call for more and better measures of instructional practices in schools—measures that would enable researchers and policymakers to capture instruction reliably and efficiently across a large number of classrooms over time, without causing an unreasonable burden on teachers, and in a way that could be linked to evidence of student achievement (Brewer & Stasz, 1996; Burstein et al., 1995; Mayer). Our project addresses this challenge.

Survey and Case Study Methods for Measuring Instructional Practice

Two types of instruments most commonly used to measure classroom practice are surveys and case studies. In this section we consider strengths and limitations of each. In the next section, we explore alternative methods for measuring classroom practice.

Surveys—strengths and limitations. Teacher surveys are the most common method for gathering data on classroom practice. As Patton (1990) noted, one advantage of surveys is "that it is possible to measure the reactions of a great many

people to a limited set of questions, thus facilitating comparison and statistical aggregation of the data. This gives a broad, generalizable set of findings presented succinctly and parsimoniously" (p. 14). In the educational arena, teacher surveys are a cost-effective way to include large numbers of classrooms in studies (Mayer, 1999). Thus, they are particularly useful in large systems for exploring differences in practice among teachers and schools and identifying broad patterns of change.

The information that surveys can provide about classroom practices is limited, however. Based on a program of research to validate indicators of the curriculum that students experience in American high schools, Burstein et al., 1995, noted:

> Some aspects of curricular practice simply cannot be measured without actually going into the classroom and observing the interactions between teachers and students. These interactions include discourse practices that evidence the extent of students' participation and their role in the learning process, the specific uses of small-group work, the relative emphasis placed on different topics within a lesson, and the coherence of teachers' presentations. (p. 7)

Surveys also cannot provide sufficient information about the complex interactions among factors that may be the underlying determinants of change within a system.

Surveys may be especially limited in their ability to capture the instructional features of standards-based reform. During times of educational reform, the language used to describe emerging practices may not be well understood, and educators may not agree about the meaning of key terms such as *student-centered lessons*, *authentic tasks*, *active learning*, *problem solving*, and *reasoning*. For these reasons, surveys that rely on such terminology to ask whether teachers are engaged in practices that are consistent with standards-based reforms are not likely to yield valid results. As one example, interviews conducted as a follow-up to surveys in a study by Antil, Jenkins, Wayne, and Vasdasy (1998) revealed that, at best, only about one-fourth of the interviewed teachers were using cooperative grouping practices that met formal definitions proposed by other researchers even though their survey responses indicated that all were using cooperative groups.

With respect to validity, Burstein et al. (1995) further cautioned that we must use care in drawing conclusions from national teacher surveys. They explained:

> None of the national survey data collected from teachers have been validated to determine whether they measure what is actually occurring in classrooms.... Little effort

> has been made to validate these measures by comparing the information they generate with that obtained through alternative measures and data collection procedures. (p. 8)

Case studies—strengths and limitations. In-depth case studies of specific classrooms and schools constitute another approach to determining instructional practices. Indeed, as Mayer (1999) claimed, "Much of what the country currently knows about the instructional process comes from in-depth studies done in only a handful of classrooms" (p. 30). In an era of educational reform, case studies play an important role in developing "systemic understanding of patterns of practice in classrooms where teachers are trying to enact reform" (Spillane & Zeuli, 1999, p. 20). They are particularly well suited to exploring, in depth and in a small number of schools and classrooms, the complexity of factors that interact to determine the differential success of the reform efforts.

One central limitation of case studies is that the generalizability of findings and conclusions to classrooms other than those investigated is unknown. As Knapp (1997) pointed out:

> Case studies ... give little indication of system-wide trends and tendencies, and even though intelligent guesses can be made in some instances, there is a clear need for large-sample research that can locate case study patterns in a larger, system-wide context. (p. 257)

Case studies are also very time- and labor intensive. While generalizability issues preclude relying on in-depth studies of a small number of classrooms to assess the impact of large-scale reform efforts, in-depth studies based on larger, representative samples of classrooms involved in a reform are often cost prohibitive. Thus, as Mayer (1999) warned, "As reform efforts increasingly focus on classroom processes, demand for impact analysis increases, and the generalizability limitations of in-depth studies become more and more problematic" (p. 30). For these reasons, case studies are typically not feasible tools for policy research on the impact of large-scale educational reform efforts.

Alternative Approaches for Studying Instructional Practice

Given the increasing demand for viable ways of obtaining information about the status of instructional practice in K-12 schools, coupled with clear limitations of both surveys and case studies, it is not surprising that numerous scholars around the country are exploring alternative approaches for collecting teacher practice data.

Three such approaches are vignettes of practice, teacher logs, and instructional artifacts. Kennedy (1999) described vignettes and teacher logs as "situated descriptions of teaching"—tools that attempt "to obtain, from teachers, as situated a description as possible of the teachers' own teaching practices" (p. 349) without directly observing their classrooms. More specifically, the aim of both vignettes and logs "is to move past broad generalities, vagaries, or espoused principles of practice toward teachers' actual practices, but without the expense of observing them firsthand" (p. 349). Instructional artifacts—the focus of our work—share these characteristics.

Vignettes of practice. In items constructed around vignettes of practice, teachers are asked how they would respond to a specific hypothetical teaching situation (e.g., Kennedy, 1999; Ma, 1999; Stecher et al., 2002). For example, Ma presented Chinese and American teachers with a set of four vignettes posing problems related to the teaching of specific mathematics topics (e.g., subtraction with regrouping, dividing by a fraction). Vignettes in the Teacher Education and Learning to Teach study focused on what teachers would do in specific situations related to the teaching of writing and mathematics. With respect to writing, one vignette presented an example of a student's written story. A series of questions explored how teachers would respond to the student and what grade they would assign to the piece (Kennedy).

When carefully constructed around fundamental ideas in a subject area and central issues related to teaching and learning of that subject, vignettes can reveal teachers' beliefs about the nature of knowledge, how students learn, appropriate pedagogical practices, etc. In mathematics, for example, vignette-based items can be constructed to reveal differences in the relative importance that teachers place on procedural and conceptual knowledge, their preference for different ways of representing mathematical ideas, and the relative value they assign to learning activities such as teacher explanation, hands-on student investigations, open-ended problems, and guided practice. By using standardized situations and questions, vignettes have the advantage of enabling researchers to aggregate findings across teachers, and thereby describe patterns of variation across teachers or patterns of change in practice over time.

The psychological literature provides evidence that intention is a strong predictor of behavior (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975), and therefore efforts to measure teachers' instructional intentions might serve as good predictors

of instructional practice. However, for a variety of reasons, teachers' responses to vignette-based questions may provide a weak measure of their intentions to act in a real situation. Because the vignettes are hypothetical, they differ in innumerable ways from real teaching situations. For example, vignettes used in research must be reasonably short, so they necessarily leave much unstated, including information about the affective and psychological domains, the physical characteristics of the setting, and the learning history of the particular students described. Researchers cannot be sure that teachers would respond to actual situations in the way they respond to these "lean" descriptions. In addition, it is difficult to draw inferences about general patterns of instructional practice from responses to vignettes because the range of teaching situations that can be portrayed in a handful of vignettes is small compared to the universe of actual teaching situations. It is also difficult to sample situations in any systematic way (and thus improve the generalizability of results) because we lack a comprehensive description of the universe of mathematics or science instructional situations from which to sample (Kennedy, 1999).

One way to improve the quality of inferences from vignette-based items is to reduce the universe of generalization by restricting either the range of practices or the subject matter under investigation. Such restrictions would make sense if one were developing instruments to measure the prevalence of specific practices being fostered by a particular reform program or focusing on a particular site whose curriculum was known. The Mosaic II project (Stecher et al., 2002) uses both strategies to improve its measures of instructional practice. The researchers developed a conceptual framework that specified dimensions of standards-based science and mathematics education that were of interest, and they selected a subset of those dimensions to be the focus of their scenario-based items. They also worked closely with the curriculum materials from specific sites and selected content around which to develop scenarios that were familiar to teachers in those sites. Given this approach, the Mosaic II researchers expect that responses to their vignette-based questions will yield better information about a specific class of behaviors in a specific location than they would about general science teaching behaviors in a randomly selected district. Results suggest that the vignettes measure "a stable aspect of teaching," but its relationship to other measures such as surveys and logs is inconsistent (Li et al., 2003).

Teacher logs. In recent years, several studies have pioneered the use of teacher logs to generate data on teachers' curriculum and instruction for a specified period

of time (e.g., Burstein et al., 1995; Porter, Floden, Freeman, Schmidt, & Schwille, 1988; Smithson & Porter, 1994). Logs typically are self-administered instruments, asking teachers to report on concrete features of their instructional practices, such as topics covered and pedagogical strategies used. They are designed to be efficient and brief, thus enabling researchers to collect information on many facets of instruction without undue burden on teachers. One potential limitation of logs is that teachers may be tempted to make their records reflect what they intended to do rather than what they actually did. Like surveys and other self-report instruments, they are also susceptible to the possibility that teachers may have different understandings of the meanings of some of the terms included in the logs. To the extent that teachers use the same educational terms (e.g., student-centered, problem solving) to refer to different concepts or practices, teacher logs will not provide reliable or valid information about the nature of teaching and learning experiences in their classrooms. Further, logs appear to be more effective as tools for collecting information about topics and tasks, rather than for capturing the character of intellectual work done by teachers and students (Kennedy, 1999).

Ball, Rowan, and colleagues (Ball, Camburn, Correnti, Phelps, & Wallace, 1999; Rowan, Camburn, & Correnti, 2002) are currently developing and pilot testing logs that attempt to address these concerns. In their pilot work, they are considering key questions related to reliability and validity, such as how to divide the day into discrete blocks of instructional time and what language to use to discuss topics of instruction, instructional treatment of the topics, and the organization of individual students for instruction, as well as practical issues, such as how to design a log that teachers will be willing to fill out on a daily basis, and what kinds of intrinsic and extrinsic incentives might motivate teachers to complete the log. They are also working to achieve an appropriate balance between specificity and generality.

Instructional artifacts. Several researchers have incorporated a variety of instructional artifacts into their data collection packages (e.g., Aschbacher, 1999; Burstein et al., 1995; Clare, 2000; Clare & Aschbacher, 2001; Clare, Valdes, Pascal, & Steinberg, 2001; Matsumura, Garnier, Pascal, & Valdes, 2002; McDonnell & Choisser, 1997). Researchers typically ask teachers to collect and annotate a set of materials such as classroom exercises, homework, quizzes, projects, exams, and samples of student work. These materials may be defined by the researchers (for example, Burstein and colleagues asked teachers to provide copies of all assignments in a specified time period) or chosen by the teachers based on a set of criteria specified

by the researchers (for example, teachers in Matsumura's research selected "typical" and "challenging" language arts assignments along with samples of "medium-" and "high-quality" student work).

We discuss Matsumura[1] and Aschbacher's work in some detail because it is most closely related to our own in structure, although they focused on instruction in language arts, whereas we focus on science and mathematics. The purpose of their project was to develop a measure of students' learning environments—consisting of collections of teacher assignments and associated student work samples—that could be used as an alternative to classroom observations in studies to determine the influence of school reform efforts.

In their initial study (Clare, 2000), third- and seventh-grade teachers were asked to select four language arts assignments—three "typical" and one "cognitively challenging," along with two "medium-" and two "high-quality" examples of student work for each assignment. Researchers observed two "typical" language arts lessons taught by each teacher and wrote detailed field notes describing the classroom, lesson activities, and interactions between the teacher and students. The three data sources were independently rated by the researchers on a number of dimensions such as cognitive challenge, clarity of learning goals, and clarity of grading criteria. The researchers then conducted several tests to determine the consistency of ratings of classroom assignments and the relationships between ratings of assignments, student work, and classroom observations. They concluded that these instructional artifacts appear to provide a reliable and valid measure of the quality of classroom assignments in language arts.

A subsequent study with a larger sample provided some additional support for the conclusion that classroom assignment ratings are reliable estimates of the quality of language arts assignments. In this larger study, each teacher selected three assignments and provided four samples of student work (two medium and two high quality) for each assignment. Each assignment was assessed by five raters on a number of dimensions, including cognitive challenge of the task, clarity of learning goals, clarity of grading criteria, and overall quality. This design yielded a consistent (or stable) estimate of quality of classroom practice at the secondary school level but not the elementary school level. The researchers suggested that the lack of consistency at the elementary school level might be due to the fact that for

[1] Matsumura's previous work was published under the name "Clare."

elementary teachers, in contrast to secondary teachers, within-teacher variation in assignment quality was greater than between-teacher variation. Also, at the secondary level, higher quality teacher assignments were associated with higher quality student work and higher reading and language achievement outcomes (Matsumura et al., 2002).

Matsumura and colleagues also identified some limitations to the use of artifacts for measuring instructional practice. For example, artifacts may embody some features of instructional practice more accurately than others. In another follow-up to the original study, Clare et al. (2001) noted that scales measuring the cognitive challenge of the task, clarity of the learning goals and task, and overall quality of the assignment were significantly associated with the quality of observed lessons and student work. However, scales measuring the quality of grading criteria and the alignment of goals and grading criteria were not associated with the quality of student work. They suggested that the fact that teachers in some of the schools were using rubrics designed by outside sources might explain these findings. In addition, the researchers found that using only two assignments produced an unacceptable level of stability. They concluded that at least three assignments are needed to determine a stable estimate of quality. At the same time, they cautioned that the amount of time and effort required for collecting assignments should be a consideration in determining how many assignments to request in any given study. Additional research is needed to further explore the conflicting patterns of results they obtained regarding the relationship between teachers' grading criteria and other indicators of instructional quality, and to determine whether the findings and conclusions from these studies extend to subject areas other than language arts.

Our Approach to Measuring Instructional Practice

Our research also investigates the feasibility, reliability, and validity of using instructional artifact packages to measure instructional practices. We selected instructional artifacts rather than teacher logs because of our interest in characterizing the intellectual work represented by instructional activities and student work samples, and the potential strength of artifacts for representing what teachers and students actually do (rather than believe they should do) in the classroom. Our goal is to develop instruments that are widely applicable to reform-oriented instructional programs and can be used in any school trying to enact instructional reforms, rather than instruments that are meaningful only to a small

subset of programs and schools. To that end, we have identified characteristics of instruction that are broadly endorsed in the reform literature. These characteristics have informed our development of guidelines for collecting instructional artifacts and rubrics for scoring the artifact collections.

In addition, we are attempting to address some of the unresolved issues identified by Matsumura and her colleagues (Clare et al., 2001; Clare & Aschbacher, 2001). We are focusing on two additional subjects—middle school mathematics and science—thus enabling us to examine whether the findings from their work extend to other subject areas and grade levels. In contrast to their work, we asked teachers to collect *all* assignments that they used in a given time period. In doing so, we are attempting to avoid the problem they encountered that assignments of one type were more effective at judging language arts practice than assignments of another type (Clare & Aschbacher).

This report presents results of the first phase of our project—pilot studies of the science and mathematics artifact collection and scoring procedures. These pilot studies were designed to test the feasibility of the artifact approach prior to use on a large scale and to provide data that could be analyzed for preliminary answers to the research questions that guide our overall study:

- Do raters agree on the scores that they assign to characterize various dimensions of instructional practice, based on the artifact packages?

- Is agreement among raters higher for some dimensions of classroom practice than for others?

- Is agreement among raters higher for some classrooms than for others?

- Do the scores assigned by raters based only on the artifact packages agree with scores assigned by raters who also observed in the classrooms and based their ratings on artifact packages *and* observational data ("gold standard" ratings)?

- Is agreement among notebook-only and gold standard ratings higher for some dimensions of classroom practice than for others?

- Is agreement among these ratings higher for some classrooms than for others?

Answers to the first set of questions provide information about the reliability of ratings assigned to the artifact packages. The results indicate how consistently readers use the scoring guides and what features of the scoring guides or the notebooks are most problematic. Answers to the second set of questions are relevant to the validity of scores derived from the notebooks. This information shows whether impressions gained from the notebooks alone are similar to impressions gained from notebooks combined with direct observation of classes.

The science artifact collection and scoring procedures were developed and tested first; the mathematics pilot study occurred a few months later and built on the lessons learned during the science pilot. To reflect this process, we first present methods and results for the science pilot study, and then present parallel information for the mathematics pilot study. The final section of the paper highlights patterns across the two pilot studies and addresses some of the modifications we are making in the artifact collection and scoring procedures for the validation studies.

Science Pilot Study

Methods

Overview. Six middle school science teachers from two states (California and Colorado) participated in the science pilot study in Spring 2002. Each teacher gathered artifacts of classroom practice for approximately one week of instruction according to directions we provided. Members of the research team observed each classroom for 2 to 3 days during the time in which the teacher collected artifacts. In addition, in the Colorado classrooms, instruction was audiotaped during the days class was observed. Researchers who observed the lessons rated instructional practices in each classroom along a number of dimensions that characterize features of reform-based science instruction, using scoring rubrics developed specifically for the pilot study. These ratings constituted the "gold standard" for the purposes of the study. Researchers who did not observe the lessons rated instructional practices on the basis of artifacts only, using the same rating form. Ratings based on the artifact collections were compared across raters; and these ratings were then compared to the "gold standard" ratings.

Participants. We solicited recommendations for teachers who were interested in participating in a research project and represented a range of teaching experience, from regional training institute staff, district administrators and staff development personnel, and building principals. Based on this information, we attempted to

select teachers to represent both traditional and reform-oriented approaches to teaching science. As shown in Table 1, the six teachers also represented a range of middle school grade levels and school contexts. Each teacher received $250 for participating in the pilot study.

Data collection: The "Scoop Notebook." Development of the science artifact collection procedures started with consideration of the kinds of instructional artifacts we might expect to see in a middle school science classroom—physical evidence that would provide a comprehensive picture of the learning environment and resources available to students in that classroom, and the types of work they produced. We generated a list that included instructional materials (e.g., equipment, textbooks); assignments; quizzes and tests; student work; feedback or comments on student work; student projects (e.g., written reports, posters, models); wall displays; teacher descriptions of lessons; and teacher reflections on lessons. We thought about this set of artifacts using an analogy to the way in which scientists approach the study of unfamiliar territory (e.g., the Earth's crust, the ocean floor). Just as scientists may scoop up a sample of materials from the place they are studying and take the sample to their laboratory for analysis, we planned to "scoop" materials from classrooms for *ex situ* examination. Further, like the scientists who do not actually spend time beneath the Earth's crust or on the ocean floor, we hoped to structure the collection of artifacts to obtain information similar to that which could be obtained through classroom observations, without the time and expense of such methods. Because of the usefulness of the analogy, we called our artifact collection package the "Scoop Notebook."

Table 1

Science Pilot Teachers

Teacher	Grade level	State	Setting	Informant description of classroom practice
Lebett	6	CO	Rural	Traditional
Onker	8	CO	Rural	Reform
Clement	7/8	CO	Urban charter school	Reform
Mason	7	CA	Suburban	Reform
Glebe	7	CA	Suburban	Traditional
Hammer	8	CA	Suburban	Unknown

Note: The artifacts package for Clement was not available for analysis of the pilot.

We designed the Scoop Notebook to incorporate a variety of methods for capturing aspects of classroom practice: photocopies, photographs, and teachers' responses to reflective questions. We also believed that discourse is a crucial aspect of classroom practice. However, we struggled with the question of whether it would be possible to capture classroom conversations efficiently and inexpensively. We decided to collect audiotapes of lessons in half of the classrooms, to explore the feasibility of obtaining classroom discourse data as part of the artifacts collection process, as well as to determine what additional information discourse analysis provided.

Using the Aschbacher/Clare materials (Aschbacher, 1999) as a model, we drafted the instructions for the Scoop Notebook for science. We decided to ask teachers to collect artifacts from one of their classrooms for a period of 5 to 7 consecutive days of instruction. For teachers whose instruction varies from day to day, this seemed to be a sufficient length of time to capture a range of teaching practices. We specified that the teacher should begin the "scoop" on a day that was a logical starting point from an instructional perspective (e.g., the beginning of a unit or series of lessons on a single topic), not necessarily the first day of the week. Teachers with block scheduling or other scheduling anomalies were instructed to "scoop" for an amount of instructional time approximately equivalent to 5 to 7 days on a normal schedule. We asked teachers to select a class comprised of students who were fairly typical of their students and to pick a series of lessons that were fairly typical of instruction in their classroom. For the purposes of the pilot, we wanted to avoid unusual groups of students and highly distinctive units that were taught in a manner unlike the other parts of the curriculum. Our goal was to sample a range of practices that would be found in the instructional program provided to most students.

When we described the Scoop Notebook to teachers, we framed the discussion in terms of the question: "What is it like to learn science in your classroom?" Because we were interested in all types of materials used to foster student learning, we asked teachers to "scoop" materials that they generated, as well as materials drawn from a textbook or other curricular resources. We considered three possible ways in which to structure the scoop task: specifying the artifacts to be collected, allowing teachers to select those artifacts most representative of their practice, or asking for all possible artifacts for a given period of time. Given our intention to design an instrument with applicability to a broad range of teaching practices, we did not want to attempt to

specify "most representative artifacts" for all possible teachers and classrooms. We therefore decided to use a combination of the other two options: having teachers collect as diverse a set of artifacts as possible for a given period of time and then indicate which of these artifacts they felt best represented their practice.

We packaged the Scoop Notebook as a three-ring binder, consisting of the following components.

- daily reminder and final checklists
- project overview
- directions for collecting a "Classroom Scoop"
- folders for assembling artifacts
- Post-It notes for labeling artifacts
- calendar of "scooped" class sessions
- daily interview questions
- photograph log
- disposable camera
- consent forms

Directions in the Notebook asked teachers to collect three categories of artifacts: materials generated prior to class (e.g., handouts, scoring rubrics), materials generated during class (e.g., writing on the board or overheads, student work), and materials generated after class (e.g., student homework, projects). The teachers were encouraged to include any other instructional artifacts not specifically mentioned in the directions. For each instance of student-generated work, they were asked to collect two examples of "high quality" and two examples of "average quality" work. Because we were interested in teachers' judgments about the quality of student work, we requested that their selections be based on the quality of the work rather than the ability of the students, and that they make an independent selection for each instance rather than tracking the same four students throughout the artifacts collection process. In addition, the teachers were given disposable cameras and

asked to take pictures of the classroom layout and equipment, transitory evidence of instruction (e.g., work written on the board during class), and materials that could not be included in the notebook (e.g., posters and three-dimensional projects prepared by students). Teachers also kept a photograph log in which they identified each picture taken with the camera; completed a daily entry in the calendar, giving a brief description of the day's lesson; and responded to questions, either orally or in writing, reflecting on the day's lesson. After collecting and assembling all of these materials, teachers were to select the five artifacts which best represented their typical practice and mark them with stars.

At the conclusion of the artifacts collection process, a researcher met with each participant for an "exit interview" to obtain feedback on their experience with the Scoop Notebook. At the beginning of the interview, we reminded teachers that we were interested in determining how well the scooped artifacts would represent their practice to someone who had never been in their classroom. Conversation captured on the tapes was heard directly by the classroom observer and was considered in assigning his/her gold standard rating.

Scoring guide. Our efforts to identify and define the dimensions on which to rate the artifact packages were guided by the vision of a science classroom portrayed in the *National Science Education Standards* (National Research Council [NRC], 1996): "Schools that implement the *Standards* will have students learning science by actively engaging in inquiries that are interesting and important to them. Students thereby will establish a knowledge base for understanding science" (p. 13). We generated an initial list of dimensions based on a thorough review of the *National Science Education Standards* (NRC). We paid particular attention to the standards for teaching, assessment, and science content because of their specific relevance to our focus on features of reform that are evident in classroom practice and affect students' opportunity to learn science. We revised this list based on the elements of standards-based science instruction defined by the Mosaic II project (Stecher et al., 2002). We drew upon the work of Mosaic II because of its similar focus on instructional practice. Their list of elements was generated from the NRC *Standards* and then reviewed and modified by a panel of experts. Our focus on instructional artifacts led us to eliminate some of their dimensions and modify others. For example, although equity and engagement—both elements identified by the Mosaic II panel—are important aspects of standards-based practice, it did not seem likely

that they could be captured on the basis of classroom artifacts. This process led to a list of nine dimensions:

- **Collaborative Grouping.** The extent to which the series of lessons uses student groups of varying size and composition to promote collaborative learning of science.

- **Materials.** The extent to which the lesson uses sufficient quantities of appropriate instructional materials to provide access to information, enhance observation, support investigations, and help students develop scientific understanding.

- **Assessment.** The extent to which the series of lessons includes a variety of approaches to gather information about student understanding, guide instructional planning, and inform student learning.

- **Scientific Discourse.** The extent to which the teacher and students "talk science" (Lemke, 1990)—that is, explicitly engage in discussions that promote scientific habits of mind and ways of knowing.

- **Structure of Instruction.** The extent to which instruction is organized to be conceptually coherent such that activities build on one another in a logical way.

- **Hands-On.** The extent to which students are interacting with physical materials or models to learn science.

- **Minds-On.** The extent to which students participate in activities that engage them in wrestling with scientific issues and developing their own understanding of scientific ideas.

- **Cognitive Depth.** The extent to which the lessons promote students' understanding of important concepts and the relationships among them, and their ability to use these ideas to explain a wide range of phenomena.

- **Inquiry.** The extent to which students are actively engaged in formulating and answering scientific questions.

In addition to these nine dimensions, raters were asked to consider the whole collection of artifacts holistically, judging "how well the series of lessons reflect a model of instruction consistent with the *National Science Education Standards*" (NRC, 1996).

The scoring process. The scoring process for the Scoop Notebooks was a simplified version of the process used in the classroom observation component of the Mosaic II project. In that study, raters first classified the observed behavior into one of three levels (high, medium, or low), and then further refined the classification into one of three sub-categories (high, medium, or low) within each major category and provided a written justification for their choice. The purpose of this two-stage process was to derive finer distinctions based on the evidence contained in the justifications. We dropped the second level of classification at the suggestion of staff from the Mosaic II project. In addition, we added an "indeterminate" option for each dimension, to allow for the possibility that a rater would judge that the Scoop Notebook contained insufficient information to enable rating that dimension. We also added a rating of "nonexistent," where appropriate, to indicate the total absence of an element of standards-based science teaching practice. For example, a rating of "low" on the dimension "Hands-On" indicated rare use of physical materials or models, whereas a rating of "nonexistent" indicated that no physical materials or models were apparent in artifacts included in the notebook. We wrote descriptions of high, medium, and low practice for each dimension and added specific examples of practice to further anchor each level. Raters were asked to assign a value to the notebook on each dimension and to justify their scores in terms of the evidence in the notebook (see Appendix A). Finally, to assess how well the artifacts each teacher selected as most representative of his/her practice matched the entire Scoop, raters considered "the extent to which the starred artifacts portray the same impression of instruction as the full notebook."

A team of six researchers (five of whom had participated in the data collection process) convened to score the Scoop Notebooks. Prior to conducting the scoring, we engaged in extensive discussions of the scoring rubrics to ensure that all raters had similar understandings of the dimensions and scoring levels, and we revised the rubrics on the basis of these discussions. During these "calibrating discussions" we agreed to add (+) and (-) ratings to the medium level. Thus, we effectively had a six-point scale: 0 (non-existent), 1 (low), 2 (medium -), 3 (medium), 4 (medium +), and 5 (high).

Results

Five of the six Scoop Notebooks were complete and available for analysis. The six-person research team met for 2 1/2 days to finalize the scoring rubric, calibrate ratings along each dimension, rate each of the notebooks, and discuss ratings and "lessons learned" from the data collection and scoring activities. Each notebook was rated by at least two researchers who were not familiar with the teacher or the classroom; their ratings were based solely on the materials in the Scoop Notebook. In addition, a "gold standard" rating was assigned by the researcher who observed in the teacher's classroom, based on all available information—classroom observation field notes, artifacts, and the teacher's responses to the exit interview questions. In presenting the results, we first focus on patterns in the notebook-only ratings; these patterns address our first set of research questions, which are reliability questions. We next focus on comparisons of notebook-only and gold standard ratings, which address our validity questions.

Ratings and rater agreement for the Scoop Notebooks. Some dimensions were more difficult to rate than others. Judgments of how well the starred artifacts represented the teacher's typical practice were especially problematic. For two of the notebooks (Lebett & Onker), one rater assigned a score of 1, while the other assigned a score of 5. Additionally, several teachers actually forgot to attach stars to any work and only did so, in a hurried manner, when prompted by the researchers during the exit interview. Thus, having teachers select these artifacts did not seem to be a useful strategy. For these reasons, we do not include ratings of the starred artifacts in further analyses of the pilot data, and we did not include this component of the Scoop Notebook in the mathematics pilot study.

Focusing on the remaining dimensions, raters had the most difficulty assigning a score for "Scientific Discourse" based solely on the artifact packages. In one-third of the cases, the rater assigned a score of "indeterminate" on this dimension. None of the other dimensions posed such a problem for raters; they were able to assign scores on all notebooks on all the other dimensions. These difficulties provide some support for our initial conjecture that it may not be possible to determine quality of scientific discourse in a classroom through a collection of artifacts such as the Scoop Notebook. This is discussed further in the section comparing the science and mathematics pilot studies.

Table 2 shows the average ratings and degree of agreement among raters for each dimension of each notebook. With one exception, all notebooks were scored by at least two raters on all 10 dimensions. Two raters were unable to score two of the notebooks on the Scientific Discourse dimension. These two notebooks were scored by at least two raters on only 9 dimensions. Thus, there were 48 opportunities to compare ratings. Two measures of rater agreement are provided—the percentage of ratings that were identical and the percentage of ratings that fell into two adjacent score levels (e.g., 2 or 3). (The number of raters per notebook ranged from two to four; when there were only two raters, the percentage of agreement was either 0% or 100%.) Twenty of these 48 sets of ratings were in exact agreement (42%), and 35 of the 48 fell within adjacent score levels (73%). These results are substantially higher

Table 2

Average Science Ratings and Percent Agreement by Dimension and Notebook

	Lebett			Onker			Mason			Glebe			Hammer		
	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
Collaborative Grouping	1	100%	100%	3	0%	0%	4	0%	0%	2	100%	100%	2.75	75%	100%
Materials	1	100%	100%	2	0%	0%	2.5	0%	100%	3	0%	0%	4	50%	75%
Assessment	1	0%	0%	2	0%	0%	1.5	0%	100%	1	100%	100%	2.25	50%	75%
Scientific Discourse	1	-	-	3	-	-	1.5	0%	100%	1	100%	100%	1.5	0%	50%
Structure of Instruction	2	100%	100%	4	0%	0%	3	100%	100%	3	0%	0%	3	100%	100%
Hands-On	0.5	0%	100%	2	0%	0%	3	100%	100%	2	100%	100%	4.5	75%	75%
Minds-On	1	100%	100%	2.5	0%	100%	2.5	0%	100%	1.5	0%	100%	1.25	75%	100%
Cognitive Depth	1	100%	100%	3	100%	100%	2.5	0%	100%	1	100%	100%	1.5	50%	100%
Inquiry	0	100%	100%	2	100%	100%	1.5	0%	100%	1	100%	100%	1	100%	100%
Overall	1	100%	100%	3	100%	100%	2.5	0%	100%	1.5	0%	100%	1.5	50%	100%

Note: Agreement "within one" is the percent of ratings that fell into two adjacent score levels, e.g., 2 or 3. For example, a notebook which received scores of 2, 2, 2, and 3 has 100% agreement "within one," while a notebook with scores of 2, 3, 3, and 4 has 75% agreement "within one." When considering the Scientific Discourse dimension for Hammer's notebook, the two missing scores were counted as "not matching" in calculations of both types of agreement.

than one would expect by chance alone. Assuming two readers assigned ratings at random with a 20% probability of assigning each level, then the probability of obtaining exact agreement by chance alone would be 20% (5 identical pairs out of 25 possible pairs of ratings). The probability of assigning adjacent or identical ratings by chance alone would be 52% (13 of 25 possible combinations). These chance probability estimates are a slight oversimplification in the case of the Hammer notebook, which had four raters, but they provide a helpful baseline comparison.

Table 3 summarizes the rating information for each of the 10 dimensions across all the notebooks. The values shown in Table 3 are the averages of the ratings and agreement percentages in Table 2, with each notebook weighted equally.

In addition to the difficulties already discussed with respect to Scientific Discourse, Table 3 shows that the degree of rater agreement varied by dimension. On four of the nine other dimensions (Cognitive Depth, Minds-On, Inquiry, and Overall), scores for all notebooks fell consistently within adjacent levels. A middle level of agreement was obtained for Collaborative Grouping, Structure of Instruction, and Hands-On. For these dimensions, raters assigned adjacent ratings 60% or more of the time. Agreement was lower for the remaining two dimensions,

Table 3

Rater Agreement by Dimension (Average Across All Science Notebooks)

Dimension	Avg	EA	w/n 1
Collaborative Grouping	2.55	55%	60%
Materials	2.50	30%	55%
Assessment	1.55	30%	55%
Scientific Discourse*	1.60	33%	67%
Structure of Instruction	3.00	60%	60%
Hands-On	2.40	55%	75%
Minds-On	1.75	35%	100%
Cognitive Depth	1.80	70%	100%
Inquiry	1.10	80%	100%
Overall	1.90	50%	100%

*Based on only three notebooks.

Materials and Assessment, for which raters assigned adjacent ratings only 55% of the time.

Examination of Table 3 also provides one possible explanation for the differences in rater agreement across dimensions. Raters were more consistent in their judgments when the notebook contained little or no evidence of reform practice on a particular dimension than when the notebook contained greater evidence of reform-oriented practices. In general, the dimensions with the highest levels of agreement were those with the lowest average scores, and the dimensions with the lowest levels of agreement were those with the highest average scores. For all five of the dimensions with 100% agreement within adjacent levels, average scores were below 2.00. For four of five of the dimensions with average or low agreement within adjacent levels, average scores were above 2.00. The dimension with the lowest average rating across notebooks was Inquiry (average rating: 1.10). For that dimension, there was exact agreement between raters for four notebooks (80%). Between-rater agreement was within adjacent levels for all five notebooks (100%). At the other extreme, Structure of Instruction received the highest average rating (3.00). There was exact agreement for three notebooks (60%); that percent remained unchanged (60%) for agreement within adjacent levels. Assessment was the only dimension not fitting this pattern. Although the average rating was 1.55, low agreement was obtained. Issues with rating the Assessment dimension are discussed below.

The discussion that took place after the scoring task suggested one possible explanation for the inverse relationship between rater agreement and the presence of reform-oriented practices. During that discussion, raters indicated that it was relatively easy to judge when a notebook contained little or no evidence of a dimension or showed only low levels of that dimension. Judgments between low and moderate levels or between moderate and high levels were sometimes more difficult. This increased difficulty may have resulted in greater disagreement.

Discrepancies in ratings for the Materials dimension may have been associated with differences in raters' interpretations of this dimension. Specifically, during the pre-scoring discussion, we realized that there were differences in the extent to which raters were including in their judgment how easily the topic of the lesson lent itself to the use of materials. For example, some raters assigned a lower score to a notebook because the teacher did not use materials the rater knew to be appropriate for the lessons, although there was no evidence the teacher had access to these

materials. Because this consideration was based largely upon individual raters' knowledge of available materials, we decided that the scoring rubric should be revised to indicate that the materials dimension should be judged independently of the topic of the lesson. However, it is possible that this consideration was too difficult to remove from the scoring process. During the post-scoring discussion, we realized that there were also differences in whether raters took the instructional goals of the lesson into account when evaluating the use of materials—some raters gave a lower rating for materials when the lesson was not focused on scientific inquiry or understanding, even though the materials were used effectively in service of the lesson; others did not.

There were also problems with the definition of the Assessment dimension. The pre-scoring discussion revealed differences among raters with respect to the question of what constitutes assessment. Virtually any classroom interaction can be used to inform teachers' judgments about student learning, and it became clear in the discussion that some raters were taking a very broad view of assessment, while others were viewing assessment as a more formalized event. This realization led to an extensive revision of the definition to more clearly specify the range of activities we would count as assessment. The post-scoring discussion revealed that although raters agreed on the characteristics of quality assessments, we did not agree on how they should be weighted. For example, we agreed that both quantity (frequency of assessment events) and quality (extent to which a given assessment addresses higher-level thinking and deep conceptual understanding) are important aspects of assessment. However, we did not agree on ratings for situations in which a teacher gave frequent assessments of lower-level thinking skills, or those in which a teacher gave only summative assessments but focused on deep conceptual understanding.

Table 4 shows the average ratings across dimensions for each notebook and the average exact agreement and agreement within adjacent levels. Similar to the pattern for dimensions, the degree of rater agreement varied by notebook, and notebooks with lower average scores tended to have greater rater agreement than notebooks with higher average scores. This difference is perhaps best illustrated by comparing ratings for Lebett (average notebook rating 0.95, average exact agreement 78%) and Mason (average notebook rating 2.45, average exact agreement 20%). Again, it seems that it was easier to rate the absence of reform-oriented practice than to rate its quality when present.

Comparisons between average notebook-only scores and gold standard scores. Table 5 shows the average notebook-only ratings and gold standard ratings (based on classroom observations *and* the artifacts package) for each dimension for each classroom. Using these data we computed two measures of agreement, which are displayed in Table 6. Since the average notebook-only ratings are often decimals rather than whole numbers, we could not use the same measure of exact agreement used previously. Instead we used cutoff values of 0.5 and 1.0 for the difference between the average notebook-only rating and gold standard rating. Considering either measure, the average ratings based solely on the artifacts package were quite consistent with the gold standard ratings. Of the 50 scoring opportunities (five notebooks, each scored on 10 dimensions), the difference between the average

Table 4

Rater Agreement by Science Notebook (Average Across All Dimensions)

	Lebett			Onker			Mason			Glebe			Hammer		
All	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
Dimensions	0.95	78%	89%	2.65	33%	44%	2.45	20%	90%	1.7	60%	80%	2.32	63%	88%

Table 5

Average Artifacts and Gold Standard Ratings for Science Notebooks

	Lebett		Onker		Mason		Glebe		Hammer	
	Avg	GS	Avg	GS	Avg	GS	Avg	GS	Avg	GS
Collaborative Grouping	1	2	3	3	4	4	2	2	2.75	3
Materials	1	1	2	3	2.5	3	3	2	4	4
Assessment	1	2	2	2	1.5	3	1	1	2.25	1
Scientific Discourse	1	2	3	4	1.5	1	1	1	1.5	1
Structure of Instruction	2	2	4	4	3	5	3	2	3	3
Hands-On	0.5	1	2	3	3	5	2	3	4.5	5
Minds-On	1	2	2.5	3	2.5	2	1.5	1	1.25	1
Cognitive Depth	1	2	3	3	2.5	4	1	2	1.5	1
Inquiry	0	0	2	2	1.5	3	1	1	1	1
Overall	1	1	3	3	2.5	3	1.5	1	1.5	1

Table 6

Average Artifacts Ratings, Gold Standard Ratings, and Agreement by Dimension Across Science Notebooks

Dimension	Avg	GS	Within 0.5	Within 1.0
Collaborative Grouping	2.55	2.8	80%	100%
Materials	2.5	2.6	60%	100%
Assessment	1.55	1.8	40%	60%
Scientific Discourse	1.6	1.8	60%	100%
Structure of Instruction	3	3.2	60%	80%
Hands-On	2.4	3.4	40%	80%
Minds-On	1.75	1.8	80%	100%
Cognitive Depth	1.8	2.4	40%	80%
Inquiry	1.1	1.4	80%	80%
Overall	1.9	1.8	100%	100%

notebook-only rating and the gold standard rating was within 0.5 for 32 scores (64%), and within one point for 44 scores (88%). Notably, average notebook-only and gold standard scores on the Overall dimension were in exact agreement for two of the notebooks and within 0.5 point for the remaining three. There were only two 2-point discrepancies, both found in the Mason notebook.

As was the case for the comparison of notebook-only ratings, the lowest level of agreement between average notebook-only rating and gold standard rating occurred for the Assessment dimension. Again, these inconsistencies may be due to lack of agreement regarding how to weight the various characteristics of quality assessments.

Table 7 provides a comparison of average ratings and levels of agreement, by dimension, for the two states in our sample. Because some aspects of curriculum are determined by states, we thought it was appropriate to see whether there were differences between classrooms that might be indicative of differences in state-level policies. There are no clear patterns of differences between the two states. This finding is discussed further in the section comparing results of the science and mathematics pilot studies.

Table 8 shows the average notebook-only ratings, gold standard ratings, and percent of agreements within 0.5 point and within 1.0 point, for each notebook,

averaged across dimensions. As mentioned above, the greatest differences were found for Mason, where the match was exact for only one of the 10 dimensions and within 0.5 point for five dimensions. The remaining four dimensions had differences of 1.5 points or greater. At the other extreme, for Onker, there was exact agreement between average notebook-only ratings and gold standard ratings for six dimensions. Since Onker and Mason received the highest average ratings across

Table 7

Average Artifacts and Gold Standard Science Ratings by State

	Colorado		California	
	Avg	GS	Avg	GS
Collaborative Grouping	2	2.5	2.92	3
Materials	1.5	2	3.17	3
Assessment	1.5	2	1.58	1.67
Scientific Discourse	2	3	1.33	1
Structure of Instruction	3	3	3	3.33
Hands-On	1.25	2	3.17	4.33
Minds-On	1.75	2.5	1.75	1.33
Cognitive Depth	2	2.5	1.67	2.33
Inquiry	1	1	1.17	1.67
Overall	2	2	1.83	1.67

Table 8

Average Artifacts Ratings, Gold Standard Ratings, and Agreement by Notebook Across Science Dimensions

	Lebett	Onker	Mason	Glebe	Hammer
Artifacts Average	0.95	2.65	2.45	1.70	2.33
Gold Standard	1.5	3	3.3	1.64	2.1
Within 0.5	50%	70%	50%	60%	90%
Within 1.0	100%	100%	50%	100%	90%

dimensions, it is clear that extent of agreement between notebook-only and gold standard ratings cannot be attributed to the presence or absence of reform-oriented practices in the classrooms. There are several possible explanations for these differences. For example, it may be that variations in agreement were associated with characteristics of the particular classes (e.g., inconsistency across observations, instructional practices that varied widely with respect to reform-orientation) or Scoop Notebooks (e.g., extensiveness of information included). With our limited data set, we cannot test these conjectures.

These results and reflections on the science pilot study were used to refine and adapt procedures for the mathematics pilot.

Mathematics Pilot Study

Methods

Overview. Eight middle school mathematics teachers from two states (California and Colorado) participated in the mathematics pilot study in Fall 2003. As in the science pilot study, each teacher gathered artifacts of practice for approximately one week of instruction, according to the guidelines in the Scoop Notebook. Seven of the teachers were observed three times during the Scoop period; one teacher was observed only twice. As in the science pilot study, we audiotaped class sessions for the four Colorado teachers during the times in which they were observed. At the end of the data collection period, we conducted an exit interview in Colorado and exit survey in California in order to obtain the teachers' feedback concerning the Notebook and the study in general. As in the science pilot study, the researchers who observed the lessons provided gold standard ratings along a number of dimensions that characterize features of reform-based mathematics instruction. Researchers who did not observe the lessons rated the classrooms on the basis of artifacts only, using the same rating form. Artifact-only ratings were compared across raters, and these ratings were then compared to the gold standard ratings.

Participants. As in the science pilot study, we attempted to select teachers who represented both traditional and reform-oriented approaches to teaching mathematics, and we relied on recommendations from district personnel and school principals in making these selections. In Colorado we identified four middle school mathematics teachers, all of whom were using the same reform-based curriculum, *Mathematics in Context* (Encyclopedia Britannica, 1998). We therefore chose four

teachers in California who were using a variety of curricula, for the most part more traditional in nature.

Within the set of eight middle school math teachers, we sought variety in grade level of students taught, type of district (urban/suburban/rural) and amount of experience with the current curriculum. As shown in Table 9, across the eight research sites, classes from all three middle school grades (6-8) were represented, as were urban, suburban, and rural schools. Among the Colorado teachers, participants had been working with the *Mathematics in Context* curriculum for between 1 and 3 years, and were reported to range from "struggling with the curriculum" to "doing very well—an award-winning teacher." Among the California teachers, participants' experience with a particular curriculum ranged from no experience—picking and choosing from various sources—to more than 5 years experience; one was reported to be particularly innovative. Each teacher received a $250 honorarium for participating in the pilot study.

Data Collection: The Scoop Notebook. The mathematics Scoop Notebook was very similar to the science Notebook, with only a few alterations. Our revisions for the mathematics pilot study took into account both differences between the two subject areas and results of the science pilot study.

Teachers were not asked to affix stars to five items that best represented their typical practice. As indicated in the discussion of the science pilot study, we did not find the information to be helpful when scoring the science notebooks, and rater agreement regarding whether the starred artifacts represented the teachers' typical practice was poor.

Table 9
Math Pilot Teachers

Teacher	Grade level	State	Setting	Informant description of classroom practice	Type of curriculum
Watson	6	CO	Rural	Reform	Reform
Wainright	6	CO	Rural	Reform	Reform
Boatman	6	CO	Rural	Reform	Reform
Caputo	8	CO	Rural	Reform	Reform
Hill	6	CA	Urban	Traditional	Traditional
Lever	7	CA	Urban	Traditional	Traditional
Mandell	7	CA	Suburban	Reform	Traditional
Young	7	CA	Suburban	Reform	Traditional

We revised the Interview Questions section of the Scoop Notebook extensively in an attempt to encourage teachers to provide more detailed reflections about the scooped lessons. In addition to minor changes in wording for several of the questions, alterations included:

- Adding a set of two Pre-Scoop Reflection Questions: "What about your teaching situation is important for us to know, in order to understand the lessons that you will include in the Scoop?" "What are your overall plans for the set of lessons that will be included in the Scoop?"

- Providing examples of responses for most of the Pre-Scoop and Daily Reflection Questions to serve as models for the level of detail we expected teachers to provide. Instructions for the science pilot study included examples for only one of the questions ("Briefly describe the activities or components of this class session from beginning to end"). We hoped that by having access to more examples, teachers would be more consistent in the nature and extent of their responses.

- Adding a set of three Post-Scoop Reflection Questions: "How does this series of lessons fit in with your long-term goals for this group of students?" "How representative of your typical instruction was this series of lessons? What aspects were typical? What aspects were not typical?" "If you were preparing this notebook to help somebody understand your teaching, what else would you want the notebook to include? Why?" Similar questions were addressed in informal exit interviews conducted with the science pilot study teachers.

Scoring guide. We revised the Scoring Guide prior to use in the mathematics pilot study, taking into consideration both successes and problems encountered in the science pilot study and features of reform-oriented practice specific to mathematics classrooms. We again drew upon the work of the Mosaic II project (Stecher et al., 2002), as well as the *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) and the expertise of members of the research team. We identified the following dimensions to use when scoring the mathematics artifact collections.

- **Collaborative Grouping.** The extent to which the series of lessons uses student groups to promote the learning of mathematics. The extent to which work in groups is collaborative, addresses non-trivial tasks, and focuses on conceptual aspects of the tasks. Note: groups typically will be of varying sizes (e.g., whole class, various small groups, individual), although the structural aspect is less important than the nature of activities in groups.

- **Structure of Instruction.** The extent to which instruction is organized to be conceptually coherent such that activities build on one another in a logical manner leading toward deeper conceptual understanding and are enacted in ways that scaffold students' current understanding.

- **Multiple Representations.** The extent to which the series of lessons promotes the use of multiple representations (pictures, graphs, symbols, words) to illustrate ideas and concepts, as well as students' selection, application, and translation among mathematical representations to solve problems.

- **Hands-On.** The extent to which students participate in activities that are *hands-on*. The extent to which the series of lessons affords students the opportunity to use appropriate instructional materials (including tools such as calculators, compasses, protractors, Algebra Tiles, etc.), and that these tools enable them to represent abstract mathematical ideas.

- **Cognitive Depth.** The extent to which the series of lessons promotes command of the central concepts or *big ideas* of the discipline and generalizes from specific instances to larger concepts or relationships.

- **Mathematical Communication.** The extent to which the teacher and students *talk mathematics*. The extent to which students are expected to communicate their mathematical thinking clearly to their peers and teacher and use the language of mathematics to express their ideas. The extent to which the classroom social norms foster a sense of community so that students feel free to express their ideas honestly and openly.

- **Explanation and Justification.** The extent to which students are expected to explain and justify their reasoning and how they arrived at solutions to problems. The extent to which students' mathematical explanations and justifications incorporate conceptual, as well as computational and procedural, arguments.

- **Problem Solving.** The extent to which instructional activities enable students to identify, apply, and adapt a variety of strategies to solve problems. The extent to which problems that students solve are complex and allow for multiple solutions.

- **Assessment.** The extent to which the series of lessons includes a variety of formal and informal assessment strategies to support the learning of important mathematical ideas and furnishes useful information to both teachers and students.

- **Connections/Applications**. The extent to which the series of lessons helps students connect mathematics to their own experience, to the world around them, and to other disciplines.

- **Overall.** How well the series of lessons reflects a model of instruction consistent with the NCTM Standards (NCTM, 2000). This dimension takes into account both curriculum and instructional practices.

Scoring process. As an initial step in the scoring process, each researcher assigned the gold standard ratings for the teacher(s) he or she had observed. As in the science pilot study, these gold standard ratings took into account evidence from both the Scoop Notebook and our classroom observations. For cases in which two members of the research team had observed the same teacher, those researchers met and created a joint gold standard rating. In contrast to science, the researchers completed the gold standard ratings in mathematics prior to meeting as a group to rate the notebooks.

We next met to rate the Scoop Notebooks. All eight mathematics notebooks were complete and available for analysis. The seven-member research team (six of whom had also participated in the data collection) met for 2 1/2 days to rate the notebooks and debrief the data collection and scoring activities. As was the case for the science pilot study, the meeting began with a discussion during which we finalized the rubric and attempted to calibrate our understanding of the levels along each dimension. After the "calibrating discussion," each gold standard rating was reviewed by the researcher(s) who observed in that class and any necessary changes were made, based on the discussion and revision of the dimensions. (See Appendix B for the version of the rubric used to score the notebooks.)

The research team then used the revised dimensions to assign the notebook-only ratings. Two or three members of the research team, none of whom had any in-person experience with the classroom or teacher, rated each Scoop Notebook. We also attempted to assign raters so that each notebook was scored by at least one person who had participated in scoring the science notebooks and one person who had not. We used the same rating scales as we did in the science pilot study: 0 (non-existent), 1 (low), 2 (medium -), 3 (medium), 4 (medium +), and 5 (high). Again, we first present findings and patterns related to consistency among notebook-only ratings (reliability questions); we then examine the notebook-only versus gold standard comparisons (validity questions).

Results

Ratings and rater agreement for the Scoop Notebooks. Across the 88 scoring opportunities (8 notebooks, each scored on 11 dimensions), all raters for a notebook agreed exactly on 26 scores (30%) and within adjacent levels on 56 scores (64%). To judge whether this level of agreement is high or not, it is helpful to compare it to the amount of agreement one would get by chance alone. In this case we make the simplifying assumption that each notebook was scored by three raters, even though some were actually scored by only two raters. Assuming three readers assigned ratings at random with a 20% probability of assigning each level, then the probability of obtaining exact agreement by chance alone would be 4% (5 identical triplets out of 125 possible combinations). The probability of assigning adjacent or identical ratings by chance alone would be 21% (26 of 125 possible combinations). Thus the actual ratings were markedly above chance in both cases.

Table 10 shows the average ratings, percent of exact agreements among raters, and percent of agreements within adjacent levels, for each dimension, for each notebook. Percent of exact agreement for a dimension was computed as the percent of ratings that matched exactly. Thus, if two of the three ratings were identical, the percent of exact agreement was calculated as 67%. Similarly, percent of agreement within adjacent levels was computed as the percent of ratings that were assigned to two adjacent levels. For ratings of 1-2-3 or 4-4-1 the percent of adjacent agreement was 67%; the percent for ratings of 3-3-2 was 100%. Table 11 summarizes the rating information for each dimension, across all notebooks. The values in this table are averages of the ratings and agreement percentages in Table 10, with each notebook weighted equally (regardless of number of raters).

As was true in the science pilot study, data in these tables demonstrate that some dimensions were more difficult to rate than others. There were only two instances in which a researcher judged that a rating could not be determined for a particular dimension on the basis of information in the Scoop Notebook. Both of these instances were for the dimension of Mathematical Communication. These

Table 10

Average Mathematics Ratings and Percent Agreement by Dimension and Notebook

	Watson			Wainright			Boatman			Caputo		
	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
Collaborative Grouping	3.00	100%	100%	4.00	100%	100%	4.33	67%	100%	3.00	0%	67%
Structure of Instruction	4.00	100%	100%	3.00	0%	0%	4.00	0%	67%	4.00	100%	100%
Multiple Representations	5.00	100%	100%	2.67	0%	67%	4.67	67%	100%	3.33	67%	100%
Hands-On	2.33	0%	67%	2.67	0%	0%	4.33	67%	100%	4.00	0%	67%
Cognitive Depth	3.67	67%	100%	3.00	67%	67%	4.00	0%	67%	4.33	67%	100%
Mathematical Communication	3.00	100%	100%	3.00	67%	67%	3.67	67%	100%	3.50	0%	67%
Explanation and Justification	3.33	67%	100%	2.00	0%	67%	3.33	67%	100%	3.67	67%	100%
Problem Solving	3.00	100%	100%	3.67	67%	67%	3.33	0%	67%	4.00	0%	67%
Assessment	3.33	67%	100%	3.33	0%	67%	4.33	67%	100%	3.33	0%	67%
Connections/ Applications	2.00	67%	67%	3.00	0%	0%	2.67	0%	67%	4.00	100%	100%
Overall	3.67	67%	100%	3.00	0%	67%	4.33	67%	100%	4.00	100%	100%
	Hill			Lever			Mandell			Young		
	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
Collaborative Grouping	1.00	100%	100%	1.00	0%	0%	4.00	0%	67%	1.00	100%	100%
Structure of Instruction	2.00	0%	0%	3.00	0%	0%	2.67	67%	100%	1.50	0%	100%
Multiple Representations	1.00	100%	100%	5.00	100%	100%	3.00	100%	100%	3.00	100%	100%
Hands-On	0.50	0%	100%	3.50	0%	100%	3.00	67%	67%	3.00	100%	100%
Cognitive Depth	1.00	100%	100%	3.00	100%	100%	2.33	67%	67%	2.00	100%	100%
Mathematical Communication	1.00	100%	100%	2.00	0%	0%	2.00	0%	67%	2.50	0%	100%
Explanation and Justification	0.50	0%	100%	2.00	100%	100%	1.00	100%	100%	1.00	100%	100%
Problem Solving	0.50	0%	100%	2.00	100%	100%	3.33	67%	100%	2.50	0%	100%
Assessment	2.50	0%	100%	2.50	0%	100%	3.33	67%	67%	2.00	100%	100%
Connections/ Applications	0.50	0%	100%	4.00	0%	0%	4.00	0%	67%	2.00	0%	0%
Overall	1.50	0%	100%	2.50	0%	100%	2.67	67%	100%	2.00	100%	100%

Table 11

Rater Agreement by Dimension (Average Across All Mathematics Notebooks)

Dimension	Avg	EA	w/n 1
Collaborative Grouping	2.67	58%	79%
Structure of Instruction	3.02	33%	58%
Multiple Representations	3.46	79%	96%
Hands-On	2.92	29%	75%
Cognitive Depth	2.92	71%	88%
Mathematical Communication	2.58	42%	75%
Explanation and Justification	2.10	63%	96%
Problem Solving	2.79	42%	88%
Assessment	3.08	38%	88%
Connections/Applications	2.77	21%	50%
Overall	2.96	50%	100%

ratings provide some additional support for our conjecture that it may not be possible to determine the quality of communication in a classroom through a collection of artifacts.

More generally, the degree of rater agreement varied by dimension. On 6 of the 11 dimensions—Multiple Representations, Cognitive Depth, Explanation and Justification, Problem Solving, Assessment, and Overall—the percent of ratings within adjacent levels, averaged across schools, was over 85%. On three of these dimensions—Multiple Representations, Cognitive Depth, and Explanation and Justification—exact agreement among raters was also high (average across schools above 60%). A middle level of agreement was obtained for Collaborative Grouping, Hands-On, and Mathematical Communication. For these dimensions, raters assigned adjacent ratings between 70% and 80% of the time. Agreement was lower for the remaining two dimensions: Connections/Applications (exact agreement: 21%, within one point: 50%) and Structure of Instruction (exact agreement: 33%; within one point: 58%).

Our debriefing discussions following the rating sessions provide some insights concerning the ratings of Connections/Applications and Structure of Instruction. Students in the participating classrooms represented a range of geographic locations and backgrounds. Raters commented that it was difficult to judge the extent to

which instruction connected to "their own experience and the world around them" without information about these school and community contexts. These difficulties may help explain the inconsistencies in our ratings of Connections/Applications.

Discrepancies for Structure of Instruction may have been associated with differences in raters' interpretations of the dimension. Specifically, the dimension included aspects related to both design (the extent to which instruction was organized to be conceptually coherent) and enactment (the extent to which instruction scaffolded student understanding). Raters reported that it was difficult to assign scores when our judgments of these two aspects differed, and we apparently differed in relative weight we assigned to each of the aspects.

Table 12 shows the average ratings and percent of agreement, across dimensions, for the mathematics Scoop Notebooks. In contrast to the science pilot study, there does not appear to be a relationship between average ratings (high vs. low) and consistency in rater judgments. For example, high levels of agreement were obtained for notebooks with both relatively high average scores (e.g., Watson) and relatively low average scores (e.g., Young). One possible explanation for the differences in rater agreement is that some teaching practices, while representing similar levels of reform, may differ in the extent to which they can be captured by classroom artifacts. When practices that are difficult to represent in notebooks are present in a classroom, it may be more difficult for raters to agree on their ratings. Also, some notebooks may have contained less information than others, thus possibly causing raters to make greater inferences and making it less likely that their ratings would agree.

Table 12

Average Ratings for Notebooks in the Mathematics Pilot

	Watson			Wainright			Boatman			Caputo		
	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
All dimensions	3.30	76%	94%	3.03	27%	52%	3.91	42%	88%	3.74	45%	85%
	Hill			Lever			Mandell			Young		
	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1	Avg	EA	w/n 1
All dimensions	1.09	36%	91%	2.77	36%	64%	2.85	55%	82%	2.05	64%	91%

Comparisons between average notebook-only scores and gold standard scores. Table 13 shows the average notebook-only ratings and gold standard ratings for each dimension, for each classroom. As with the science notebooks, we report two levels of agreement using cutoff values of 0.5 and 1.0 for the difference between average notebook-only rating and gold standard rating. In general, agreement between the gold standard ratings and the average ratings based solely on the artifacts package were moderate. Of the 88 scoring opportunities (8 notebooks, each scored on 11 dimensions), the difference between average notebook-only ratings and gold standard ratings was within 0.5 point for 36 scores (41%) and within 1 point for 60 scores (68%).

Table 14 shows the average notebook-only ratings, gold standard ratings, and percent of agreements within 0.5 point and within 1.0 point, for each dimension, averaged across all classrooms. As was the case with comparisons of notebook-only ratings, the extent of agreement between average notebook-only ratings and gold standard ratings varied by dimension. For the Hands-On, Cognitive Depth, and Overall ratings, agreement between average notebook-only ratings and gold standard ratings was within 0.5 point for 5 of the 8 classrooms (63%). At the other

Table 13

Average Artifacts and Gold Standard Ratings for Mathematics Notebooks

	Watson		Wainright		Boatman		Caputo		Hill		Lever		Mandell		Young	
	Avg	GS	Avg	GS	Avg	GS	Avg	GS	Avg	GS	Avg	GS	Avg	GS	Avg	GS
Collaborative Grouping	3	5	4	4	4.33	3	3	5	1	1	1	0	4	2	1	3
Structure of Instruction	4	5	3	3	4	4	4	5	2	1	3	4	2.67	1	1.5	2
Multiple Representations	5	4	2.67	4	4.67	4	3.33	5	1	1	5	3	3	2	3	1
Hands-On	2.33	4	2.67	3	4.33	3	4	4	0.5	0	3.5	3	3	1	3	3
Cognitive Depth	3.67	4	3	3	4	3	4.33	4	1	1	3	2	2.33	1	2	2
Mathematical Communication	3	5	3	3	3.67	2	3.5	4	1	1	2	1	2	1	2.5	1
Explanation and Justification	3.33	5	2	3	3.33	2	3.67	4	0.5	1	2	1	1	1	1	2
Problem Solving	3	5	3.67	4	3.33	5	4	5	0.5	1	2	1	3.33	0	2.5	2
Assessment	3.33	4	3.33	4	4.33	3	3.33	4	2.5	3	2.5	3	3.33	1	2	3
Connections/ Applications	2	3	3	3	2.67	3	4	5	0.5	0	4	2	4	2	2	2
Overall	3.67	4	3	3	4.33	3	4	4	1.5	1	2.5	2	2.67	1	2	1

Table 14

Average Artifacts Ratings, Gold Standard Ratings, and Agreement by Dimension Across Mathematics Notebooks

Dimension	Avg	GS	Within 0.5	Within 1.0
Collaborative Grouping	2.67	2.88	25%	38%
Structure of Instruction	3.02	3.13	38%	88%
Multiple Representations	3.46	3.00	13%	50%
Hands-On	2.92	2.63	63%	63%
Cognitive Depth	2.92	2.50	63%	88%
Mathematical Communication	2.58	2.25	38%	63%
Explanation and Justification	2.10	2.38	38%	75%
Problem Solving	2.79	2.88	38%	63%
Assessment	3.08	3.13	25%	75%
Connections/Applications	2.77	2.50	50%	75%
Overall	2.96	2.38	63%	75%

extreme, agreement was within 0.5 points for only one classroom (13%) for Multiple Representations, and two classrooms (25%) for Collaborative Grouping and Assessment. For all dimensions except Collaborative Grouping and Multiple Representations, agreement between average notebook-only ratings and gold standard ratings was within 1 point for at least 60% of the classrooms. It was highest (seven of the eight classes, or 88%) for Structure of Instruction and Cognitive Depth.

Thus, considering both cutoff values, agreement was most problematic for Collaborative Grouping and Multiple Representations. Interestingly, these were not the dimensions that posed the most difficulty with respect to consistency in the notebook-only ratings. One possible explanation is that these two dimensions are difficult to capture through artifacts; thus, evidence that is readily available to a classroom observer may not be easily represented in the Scoop Notebook. On the other hand, based on similar reasoning, we expected lower agreement between the average notebook-only ratings and gold standard ratings for Mathematical Communication. That was not the case, however. One possible explanation is that the ratings for Mathematical Communication were the lowest of all dimensions. It may be that notebooks are just as revealing as observations when there is little or no mathematical communication occurring in class, despite the difficulty in

determining the quality of classroom communication through a collection of artifacts.

Table 15 shows average notebook-only ratings, gold standard ratings, and percent of agreements within 0.5 point and within 1.0 point, for each class, averaged across all dimensions. One feature that is striking is the difference in levels of agreement across classes. Agreement between average notebook-only ratings and gold standard ratings was highest for Hill (91% within 0.5 point, 100% within 1 point) and Wainright (73% within 0.5 point, 91% within 1 point). It was lowest for Mandell (9% within 0.5 point, 27% within 1 point) and Boatman (18% within 0.5 point, 36% within 1 point). As was the case for the science pilot study, these differences cannot be attributed to the presence or absence of reform-oriented practices in the classrooms. For example, the average scores for Hill and Wainright differed by almost two points, but there was high agreement among raters on both notebooks.

For both Mandell and Boatman (classrooms with low agreement), the average notebook-only ratings were higher than the average gold standard ratings. Our debriefing conversations indicated that while the artifacts demonstrated use of reform-oriented materials and activities in these two classes, instructional enactment (more apparent to the gold standard raters than the notebook-only raters) was decidedly more traditional in nature. It may be that discrepancies between curricular materials and instructional enactments in these two classrooms led the researchers who took into account both artifacts and observations to assign lower ratings on several dimensions than raters who based their judgments solely on artifacts.

Another feature that is striking are the differences in ratings between the Colorado classes (Watson, Wainright, Boatman, Caputo) and the California classes (Hill, Lever, Mandell, Young), as shown in Table 16. The higher average ratings in Colorado classes (2.93-3.82 for average notebook-only ratings; 3.2-4.4 for gold standard ratings), compared to California classes (1.1-2.95 for notebook-only; 1.0-2.2 for gold standard) is consistent with what we know about curriculum differences between the two states. California classes were using more traditional mathematics curricula, while the Colorado classes used a more reform-oriented curriculum. This pattern provides preliminary validity evidence, indicating that the Scoop Notebook captures the nature of the curriculum (reform-oriented vs. traditional).

Table 15

Average Artifacts Ratings, Gold Standard Ratings, and Agreement by Notebook Across Mathematics Dimensions

	Watson	Wainright	Boatman	Caputo	Hill	Lever	Mandell	Young
Artifacts Average	3.30	3.03	3.91	3.74	1.09	2.77	2.85	2.05
Gold Standard	4.3	3.3	3.2	4.4	1	2.2	1.1	1.9
Within 0.5	18%	73%	18%	45%	91%	27%	9%	45%
Within 1.0	55%	91%	36%	82%	100%	82%	27%	73%

Table 16

Average Artifacts and Gold Standard Mathematics Ratings by State

	Colorado		California	
	Avg	GS	Avg	GS
Collaborative Grouping	3.58	4.25	1.75	1.5
Structure of Instruction	3.75	4.25	2.29	2
Multiple Representations	3.92	4.25	3	1.75
Hands-On	3.33	3.5	2.5	1.75
Cognitive Depth	3.75	3.5	2.08	1.5
Mathematical Communication	3.29	3.5	1.88	1
Explanation and Justification	3.08	3.5	1.13	1.25
Problem Solving	3.5	4.75	2.08	1
Assessment	3.58	3.75	2.58	2.5
Connections/Applications	2.92	3.5	2.63	1.5
Overall	3.75	3.5	2.17	1.25

Conclusions

Patterns Across Pilot Studies

In both the science and mathematics pilot studies, consistency across notebook-only raters was substantially greater than chance on all dimensions, and quite high in absolute terms on many. Thus, in general, researchers were able to rate instructional practice, based on artifacts alone, with a reasonable amount of agreement. At the same time, the extent of agreement between raters varied across dimensions, indicating that it was more problematic to rate some aspects of classroom practice than others.

A similar pattern was evident in the comparisons between average notebook-only ratings and gold standard ratings. Again, in both science and mathematics, agreement was substantially greater than chance on all dimensions, and quite high on many. And agreement was more difficult to achieve for some dimensions than others.

These patterns of agreement indicate that the Scoop Notebook has the potential to be a viable tool for addressing the question, "What is it like to learn science/mathematics in your classroom?" Further, although it was necessary to take into account differences between the two subject areas when designing the notebooks and defining the dimensions of classroom practice, the final products appear to be as applicable in one subject as the other.

The Scoop Notebooks and scoring rubrics were designed to elicit information about reform-based practices. However, we see no reason why this approach could not be adapted to examine other aspects of instruction that are not specifically associated with mathematics and science reform. For example, in our discussions we often commented that we could probably revise the materials to capture some aspects of classroom management, attention to student misconceptions, and pacing if we wanted to do so.

At the same time, some dimensions and some notebooks posed more problems for raters than others. In the final section of the report, we offer possible explanations for these problems—focusing on issues related to both reliability and validity.

Possible explanations for inconsistencies. One factor that may account for some of the disagreement among raters, across all dimensions, is the lack of a shared image of reform-oriented instructional practices. Because there are few good examples of classrooms characterized by high levels of reform-oriented practices, raters' images of the top of the scales may differ. Without a clear, shared vision of reform-oriented practice to serve as an anchor, individual researchers' ratings are vulnerable to comparison with other classrooms with which they are familiar. Thus, it may be more difficult for raters to agree on the nature and extent of reform-oriented practices than traditional practices. Such difficulties would explain the association between lower scores and higher levels of agreement in the science pilot study (although we did not find a similar pattern in the mathematics pilot study). Also, to the extent that agreement was higher for the science notebooks than the

mathematics notebooks, this subject-matter difference may be due to the fact that the science notebooks were generally rated lower on the reform scale.

Another possible explanation for several patterns of inconsistency in ratings is that a teacher's instruction may not be consistent from day to day, or activity to activity, with respect to particular dimensions of practice. For example, we observed one mathematics teacher whose expectations for explanations and justifications were minimal on a day when students worked on worksheets consisting of number facts and computational problems. The next day, they worked in small groups, using pictures and Unifix cubes to solve an open-ended problem. On that day, her expectations for explanations and justifications were substantially greater.

The process of educational change, itself, may provide another source of inconsistency in the data. As teachers engage in the process of adopting more reform-oriented practices, some elements of their practice are likely to change before others. As one example, for some teachers, changes in beliefs occur before changes in practices, while for others, changes in practice precede changes in beliefs (Borko, Davinroy, Bliem, & Cumbo, 2000; Richardson, 1994). In these situations, researchers are likely to encounter mixed messages, both when they observe in classrooms and when they review the artifacts a teacher compiles to represent instructional practice. For a teacher in the process of transition, reflections may appear to be more reform-oriented than assignments, or instructional tasks may appear to be more reform-oriented on paper than they do when enacted in the classroom. In one mathematics classroom we observed, for example, the task posed to students was to design a survey, collect data, and create visual displays to report the data. On face value, this task had many characteristics of reform-oriented instruction; however, it was posed to students and enacted in a mechanistic way that minimized cognitive depth and opportunities for fostering deep conceptual understanding. Average rater agreement for classes with these types of inconsistencies was low (particularly agreement between average artifact ratings and gold standard ratings), suggesting that individual raters may have differed in how we resolved these apparent contradictions.

Inconsistencies and mixed messages may have been more common for some dimensions than others. As we noted in the Results section, our debriefing sessions revealed that definitions of some dimensions—for example, Assessment in science and Structure of Instruction in mathematics—incorporated two or more features of instructional practice. For these dimensions, raters were more likely to encounter

inconsistencies in the data and to differ in their resolutions of the inconsistencies. We have revised definitions of these dimensions for the validation study in an attempt to lessen these inconsistencies.

Another possible explanation for low rater agreement focuses on the extent to which collections of artifacts are able to capture certain instructional practices—particularly practices that relate to interactive aspects of teaching. In designing the Scoop Notebook, we anticipated that some instructional practices would be more difficult to capture with artifacts than others. We were most skeptical about our ability to rate classroom discourse accurately and with confidence, based only on the Scoop Notebooks. Our skepticism was at least partially confirmed. The only dimensions for which raters assigned a score of "indeterminate" were Scientific Discourse and Mathematical Communication. However, although raters reported difficulty in assigning scores on these two dimensions based only on the Scoop Notebooks, agreement across notebook-only ratings and between average notebook-only and gold standard ratings was not noticeably lower for these dimensions as compared to other scoring dimensions. We anticipate that our analyses of the audiotaped discourse from a subset of classrooms, which are currently in progress, will provide additional insights relevant to this explanation.

Teachers varied in the amount of information they included in their Scoop Notebooks. The number of class sessions and student assignments included in the notebooks was fairly similar across teachers, and readers appeared to have sufficient examples to make judgments on almost all dimensions. However, the nature and extent of comments about student work and reflections varied widely from teacher to teacher. We found that for notebooks with less reflection (e.g., Caputo in mathematics), there was less agreement across raters. One possible explanation for this pattern is that when data provided by the teachers were incomplete, raters were forced to make larger inferences, and the potential for disagreement was increased.

Finally, the use of a scoring guide to rate materials that can vary widely in content necessitates the application of generic rules to specific cases. Although the definitions in the scoring guides were accompanied by examples, these examples necessarily represent a very limited subset of the practices encompassed by the more generic definition. To assign a score, raters must compare the real case to the generic description and the hypothetical situation described in the example. And the real case rarely—if ever—matches the description in the scoring guide exactly. Raters may have differed in the way in which they made these comparisons, which

contributed to rater disagreement. The validation study, currently in progress, will provide additional data to explore these possible explanations.

References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting behavior*. Englewood Cliffs, NJ: Prentice-Hall.

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.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Tech. Rep. No. 513). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ball, D. B., Camburn E., Correnti, R., Phelps, G., & Wallace, R. (1999). *New tools for research on instruction and instructional policy: A web-based teacher log*. Seattle: University of Washington, Center for the Study of Teaching and Policy.

Borko, H., & Elliott, R. (1999). Hands-on pedagogy versus hands-off accountability: Tensions between competing commitments for exemplary math teachers in Kentucky. *Phi Delta Kappan*, *80*, 394-400.

Borko, H., Davinroy, K. H., Bliem, C. L., & Cumbo, K. B. (2000). Exploring and supporting teacher change: Two teachers' experiences in an intensive mathematics and literacy staff development project. *Elementary School Journal, 100*, 273-306.

Brewer, D. J., & Stasz, C. (1996). *Enhancing opportunity to learn measures in NCES data.* Santa Monica, CA: RAND.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE Tech. Rep. No. 532). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clare L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, *7*(1), 39-59.

Clare, L., Valdes, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Tech. Rep. No. 545). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Encyclopedia Britannica Educational Corporation (1998). *Mathematics in Context*. Chicago: Author.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change the effects of testing in Maine and Maryland. *Educational and Policy Analysis, 20*(2), 95-113.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Fullan, M. G., & Miles, M. B. (1992). Getting reform right: What works and what doesn't. *Phi Delta Kappan, 73*, 745-752.

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

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

Koretz, D., Stecher, B. M., Klein, S., & McCaffrey, D. (1994, Fall). The Vermont portfolio assessment program: Findings and implications, *Educational Measurement: Issues and Practices, 13*(3), 5–16.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex.

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

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Erlbaum.

Matsumura, L. D., Garnier, H. E., Pascal, J., & Valdes, R. (2002). *Measuring instructional quality in accountability systems: Classroom assignments and student achievement* (CSE Tech. Rep. No. 582). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Tech. Rep. No. 442). Los Angeles: University of

California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

Patton, M. Q. (1990). *Qualitative evaluation and research methods, (2nd edition).* Newbury Park, CA: Sage.

Porter, A., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 96-113). Hillsdale, NJ: Erlbaum.

Richardson, V. (1994). The consideration of teachers' beliefs. In V. Richardson (Ed.) *A theory of teacher change and the practice of staff development: A case in reading instruction* (pp. 90-108). New York: Teachers College Press.

Rowan, B., Camburn, E., & Correnti, R. (2002, April). *Using logs to measure "enacted curriculum" in large-scale surveys: Insights from the study of instructional improvement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from the efforts to describe the enacted curriculum—The Reform Up-Close Study.* Madison, WI: Consortium for Policy Research in Education.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies, 31*, 143-175.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis*, *21,* 1-27.

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). The effects of the Washington state education reform on schools and classrooms (CSE Tech. Rep. No. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Stecher, B., Hamilton, L., Ryan, G., Le, V.-N., Williams, V., Robyn, A., et al. (2002, April). *Measuring reform-oriented instructional practices in mathematics and science.*

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wolf, S. A., & McIver, M. (1999). When process becomes policy: The paradox of Kentucky state reform for exemplary teachers of writing. *Phi Delta Kappan*, *80,* 401-406.

APPENDIX A
SCORING GUIDES FOR SCIENCE PILOT

In all dimensions...

- *Go with the highest level you can be sure you observed in the data we collected.*
- *If you can determine some features of the dimension but not others, rate on what you do see with a note in the justification about what you could not tell.*

Collaborative Grouping. The extent to which the series of lessons uses student groups of varying size and composition to promote collaborative learning of science.	
High	Students are organized into groups of different sizes and composition over the series of lessons. The group activities appear to foster communication, collaborative problem solving, etc. to promote the learning of science. Over time, students have varying responsibilities within the groups. Example: The class is divided into groups, with each group focusing on a different planet. Students conduct research to design a travel brochure, describing the environment of their planet. Students are then reorganized into groups, with one student from each planet group in each of the new groups, to explore how the distance from the Sun affects characteristics of planetary environments such as the length of a day, the length of a year, temperature, weather, and surface composition. Example: Students are divided into small groups to brainstorm how animals in different habitats are adapted to the unique features of their environments. Each group is considering a different environment (desert, mountain, woodland, etc). The class reconvenes to consider what characteristics of animals are important to examine when thinking about how an animal is adapted to its environment. Armed with the class list, students work in pairs to examine a spider and hypothesize about where this animal might live.
Medium	Students are organized into groups of different sizes and composition over the series of lessons. However, the group activities do not take full advantage of the learning opportunities afforded by the groups, so that opportunities for communication, collaborative problem solving, etc are limited. Example: The teacher delivers a lecture on the solar system, students read about it in their textbooks, and then work in groups of 3-4 to complete a worksheet. Example: Students read about spiders in their textbook, and then they break into groups of 3-4 to study real spiders in terrariums. They return to their desks to complete a worksheet about their observations.

Low	Instruction is confined to whole-class or individual work. Example: The teacher delivers a lecture on the solar system, students read about it in their textbooks, and complete an individual worksheet. Example: Students watch a video about the anatomy of spiders.
Indeterminate	

Justification:

Materials. The extent to which the lesson uses sufficient quantities of appropriate instructional materials to provide access to information, enhance observation, support investigations, and help students develop scientific understanding.	
High	A diverse and complete set of appropriate instructional materials are used, such as print materials, measuring equipment, video materials, and real-life objects. Example: In a lesson on photosynthesis, students see a video about the chemical process of photosynthesis. They plant seedlings in terrariums and observe their growth over the period of a week. Then they manipulate the terrarium ecosystem by varying the amount of water, oxygen, etc. and observe the changes that occur. They write up their findings in lab notebooks and make science posters summarizing their findings.
Medium	Instructional materials are not matched to the curriculum, do not provide sufficient variety of examples, or are not available in sufficient quantity for all students to interact with them. Example: In a lesson on photosynthesis, students observe plants in a terrarium and read about the chemical reactions involved in photosynthesis. However, students do not have adequate materials to vary aspects of the terrarium environment. Example: In a lesson on photosynthesis, students observe plants in a terrarium and read about the chemical reactions involved in photosynthesis. However, there is a single terrarium for the entire class so that, although the environment can be manipulated, only the teacher (or a select group of students) is able to conduct the investigation.
Low	Very few instructional materials are used to help students learn. Those that are used do not support the doing of science. Example: In a lesson on photosynthesis, students listen to a lecture and read a chapter from the textbook.
Nonexistent	
Indeterminate	

Justification:

Assessment. The extent to which the series of lessons includes a variety of approaches to gather information about student understanding, guide instructional planning, and inform student learning.	
High	Assessment has the following features: • It occurs in multiple forms. • It occurs throughout the unit. • It taps a variety of scientific thinking processes. • It provides feedback to students. • It informs instructional practice. Example: The teacher uses an initial activity to elicit students' prior knowledge about plate tectonics. Based on this information, the teacher selects a series of laboratory investigations for students to complete during the unit. Throughout the investigations, students use their science journals to reflect on their current understanding of plate tectonics and the teacher uses these reflections, as well as students' homework assignments, to monitor their developing understanding. At the end of the unit, students complete a performance assessment. Example: The series of lessons on chemical changes begins with a lab activity. Students' written lab observations are reviewed by the teacher who writes questions and gives suggestions for clarification. Students use their textbook, library materials, and their notebooks, including teacher comments, to prepare a short paper.
Medium	Assessment has some but not all of the features mentioned above.
Low	Assessment has none of the features mentioned above.
Nonexistent	
Indeterminate	

Justification:

Scientific Discourse. Extent to which the teacher and students "talk science" (Lemke, 1990)—that is, the teacher and students explicitly engage in discussions that promote scientific habits of mind and ways of knowing.	
High	Teacher consistently engages students in scientific discourse where the emphasis is placed on making thinking public, raising questions, and proposing and revising explanations. Value is placed on sense making rather than getting the "right" answer. Students' ideas are solicited, explored and attended to throughout the classroom discourse. Teacher guides students' reasoning by attending to their thinking, challenging conceptions when appropriate, and providing knowledge only when it serves to help students fill gaps, or equilibrate. Example: Following an investigation on plant growth, students present their findings to the class. Considered to be research peers, classmates actively engage with the presenters by raising questions, challenging assumptions, and verbally reflecting on their reactions to the findings presented. Behaving as a senior member of the research community, the teacher asks probing questions, and pushes the thinking of both presenters and peers. Example: In a class discussion during a unit on particle theory, the teacher asks students to share their thinking about what why the diameter of a balloon increases when placed in hot water and decreases when placed in cold water. Teacher uses wait time to allow students to formulate their thinking. When students share their ideas (explanations), the teacher listens carefully and then asks them to explain their thinking and the rationale behind their thinking. Teacher asks other students to reflect upon, build on, or challenge the ideas presented by their classmates. Teacher may offer suggestions, or alternative ways of thinking about the question when gaps in student thinking are evident, correct students' ideas, or give the "real/right" answer.
Medium	Teacher and students occasionally engage in scientific discourse where emphasis is placed on sharing ideas and proposing explanations. Teacher solicits students' ideas, attempts to attend to them, but may fail to push thinking further. When students struggle with gaps in their thinking, teacher asks leading questions, restates students' ideas to include missing information, or is quickly fills in the missing information. Classroom discourse can often focus more on procedural rather than conceptual issues. Example: Following an investigation on plant growth, students present their findings to the class. Their classmates listen to presentations, but do not ask questions, challenge results or react to the findings. The teacher tends to ask the presenters more procedural questions about their investigations, rarely pushing conceptual understanding. Teacher is quick to provide content if it is missing from the presentations, or asks leading questions trying to prompt presenters into filling in the missing content. Example: In a class discussion during a unit on particle theory, the teacher asks students to reflect on how air particles might be affecting the diameter of a balloon when it is moved from bowl of hot water to a bowl of cold water. One student suggests that it has something to do with the air particles slowing down in the cold. The teacher responds to the student by saying "Yes, and when the

	air particles slow down, they don't push against the balloon as much." Teacher follows this statement with a question like, "And how would that affect the diameter of the balloon… if the air isn't pushing as hard, would the diameter of the balloon increase or decrease?" When most of the class responds with "decreases," the teacher goes on to ask, "So why then do you think the diameter of the balloon increases when we place it in a bowl of hot water?"
Low	The teacher transmits knowledge to the students primarily through lecture, or direct instruction. Discussions are characterized by IRE (initiation, response, evaluation) or "guess-what's-in-my-head." The focus is on scientific facts, rather than students' reasoning. Example: Following an investigation on plant growth, students are asked to present their findings. After all of the presentations have been given, teacher holds a whole class discussion in which she asks students to recall important facts about plant growth that they learned in the process of their investigations. All of the teacher's questions have known answers, and teacher evaluates the "correctness" of each student response as it is given. If "correct" answers are not given, the teacher asks the question again or provides the answer. Example: The teacher gives a lecture on particle theory, followed by a demonstration of how the diameter of a balloon decreases when moved from hot to cold water. In a whole class discussion, she asks students to use the information that they learned in her lecture to explain why the diameter of the balloon decreased. When one student gives an explanation that is incorrect, she corrects the response by giving the "right" answer, and moves on to the next topic.
Nonexistent	
Indeterminate	

Justification:

Structure of Instruction. Extent to which instruction is organized to be conceptually coherent such that activities build on one another in a logical way.	
High	The organization and structure of instructional questions, tasks, and activities are intellectually engaging to students. Tasks and activities are designed and sequenced to be conceptually connected and build on one another in ways that are clear to the students. Example: A unit of instruction on air pressure begins by engaging students through a provocative event in which they experience the profound effects of air pressure (trying to drink orange juice out of a cup through two straws in which one straw is placed outside of the cup). This engaging activity includes opportunities for students to explore and raise questions about their experiences with the orange juice. The teacher then involves students in a sequence of tasks designed to shape students' scientific thinking, focus on sense making, and foster scientific understanding. Lessons culminate in conclusions or generalizations made through evidence gained during students' exploration of the affects of air pressure, current scientific explanations provided, and opportunities to apply their developing understanding of air pressure to new phenomena, events or activities.
Medium	The organization and structure of instructional questions, tasks and activities are designed to address a specific concept in science. While the tasks are designed to illuminate some aspect of that concept, it may not be clear to the students how each task relates to, or builds on the other. Example: A unit of instruction on air pressure begins with the teacher suggesting that air pressure has a profound affect on our lives. She explains that we live in a sea of air, and like the pressure we feel when we dive under water, air exerts a similar, yet not as great, pressure on us as we walk around on the face of the earth. The teacher goes on to explain why this is true. Following this explanation, the teacher involves students in a series of separate activities in which they experience or witness the effects of air pressure. Lessons culminate in opportunities for students to demonstrate what they have learned about air pressure.
Low	There is no apparent organization or structure to the instructional questions, tasks and/or activities, other than a loose connection to the topic under study. While scientific content may be present in each activity or task, there is no evidence that any of the tasks are conceptually connected. Example: In a unit on air pressure, students see a video on scuba diving one day, listen to a lecture on the ideal gas law the second day, and participate in the orange juice/straw experiment described above on the third day.
Nonexistent	
Indeterminate	

Justification:

Hands-On. Extent to which students are interacting with physical materials or models to learn science.	
High	Students' use of physical materials or models forms a regular and integral part of instruction throughout the series of lessons. Example: As part of an investigation of water quality in their community, students bring water samples into class. They set up the appropriate equipment and measure the pH levels of the samples. In class the next day, students discuss how pH is related to water quality. The following day, they perform the same tests at a local stream and observe aquatic life in the stream. Example: As part of a discussion of plate tectonics, students model plate boundaries by acting them out with their bodies. Later in the unit, students supplement their reading about faults by using wooden blocks to represent different fault types.
Medium	Students occasionally use physical materials or models during the series of lessons. Example: As part of an investigation of water quality in their community, the teacher brings water samples into class and sets up equipment to measure its pH. The teacher selects several students who then measure the pH levels of these water samples while the others observe. The following day, the teacher takes them outside to watch a few students test the pH of water in a local stream. Example: As part of a discussion of plate tectonics, students model plate boundaries by acting them out with their bodies. This is students' only chance to interact with physical materials or models in the series of lessons.
Low	Students rarely use physical materials or models during the series of lessons. When physical materials are used, they are used by the teacher or the students' interaction with the materials is not related to scientific content. Example: As part of a unit on water quality, the teacher brings water samples into class, sets up equipment to measure its pH, and performs the measurements while students observe. Example: Students cut out pictures of different types of plate boundaries, assemble them on a separate sheet of paper, and label and define each one.
Nonexistent	
Indeterminate	

Justification:

Minds-On. Extent to which students participate in activities that engage them in wrestling with scientific issues and developing their own understanding of scientific ideas.	
High	Students consistently assume an active role in developing their understanding of scientific ideas by formulating questions, identifying information sources, analyzing and synthesizing information, or drawing connections among scientific ideas. Example: Students prepare for a debate about the scientific and ethical issues associated with cloning by searching for documents that analyze the issues from different perspectives, reading the various analyses, synthesizing the arguments, deciding on a position they support, and developing a written statement arguing for that position. Example: Working in groups, students use a variety of materials to explore the relationship between the form of a bird beak and the bird's food source. They test various model beaks and compare and discuss results and hypotheses with their group members.
Medium	Students occasionally assume an active role in developing their understanding of scientific ideas. The teacher may direct their learning by pointing them toward appropriate information sources, providing guidelines for formulating explanations, or otherwise constraining opportunities for self-directed learning. Example: The teacher organizes a debate about whether or not scientists should be able to continue conducting research on cloning. He prepares a worksheet which specifies that the student list 3 reasons to support each position, provides materials that analyze the issue from different perspectives, and assigns each student to one side of the debate. Students read the materials, prepare the worksheets, and then use the worksheets to present their positions. Example: Working in groups, students follow a series of instructions to test specific model bird beaks and food sources. When they've completed the activity, the teacher leads a group discussion, facilitating students' exploration of relationships between the form of a bird beak and its function.

Low	Most or all activities do not require cognitive engagement on the part of the students. They may be passively receiving information presented by the teacher or instructional materials or following a prescribed procedure to arrive at pre-determined results. Example: The teacher lectures on the pros and cons of allowing scientists to continue conducting research on cloning, using an overhead transparency she has prepared that summarizes the most important arguments that support each position. Students copy points from the overhead into their science notebooks. Example: Working in groups, students follow a series of instructions to test specific model bird beaks and food sources. When they've completed the activity, the teacher asks for a show of hands regarding the best beak for each food source and confirms the right answer.
Nonexistent	
Indeterminate	

Justification:

Cognitive Depth. Extent to which the lessons promote students' understanding of important concepts and the relationships among them and their ability to use these ideas to explain a wide range of phenomena.	
High	The series of lessons focuses on deep understanding and integration of scientific concepts. Example: During a class discussion, students are pushed to use their understandings of a) the relative motions of the Earth, Sun, and Moon and b) how light is reflected between the Earth, Sun, and Moon to explain the phases of the Moon.
Medium	The series of lessons focuses on mastery of isolated scientific concepts or limited understanding of relationships. Example: During a class discussion, students are asked to explain the phases of the Moon. Students answer that this phenomena is related to the Moon's orbit around the Earth and the light from the Sun, but they are not pushed to elaborate on these "explanations" nor to put these notions together to form a complete explanation of the phenomena.
Low	The series of lessons focuses on recall of discrete pieces of information. Example: Students are asked to complete a fill-in-the blank worksheet, identifying the names for the different phases of the Moon.
Nonexistent	
Indeterminate	

Justification:

Inquiry. Extent to which students are actively engaged in formulating and answering scientific questions.	
High	The series of lessons engage students in formulating and answering a scientific question. The question and the procedure for answering it, as well as criteria for appropriate evidence, are crafted by the students. Example: As part of a unit on motion, students are designing an amusement park. One group has chosen to work on a swinging Viking ship ride, and they are worried that the number of people on the ride (and their weight) will affect how fast the ride swings. They construct a simple pendulum and design an experiment to answer the question, "How does the weight at the end of a pendulum affect the amount of time it takes to complete ten swings?" They conduct the investigation and use the results to inform their design. Example: The class has been discussing global warming. One student remarks that they have had a very mild winter, and cites this as evidence of global warming. Another student expresses doubt that global warming could have a noticeable effect in such a short period of time. The teacher suggests that the students ask a question which could help them to resolve this issue. As a class, they decide to investigate how the temperature in their city has changed over the past 100 years. Students debate about what data they should gather, and different groups of students end up approaching the problem in different ways.
Medium	The series of lessons focuses on answering question(s). However, the students are only partially responsible for clarifying the question, designing procedures for answering it, and criteria for evidence. Example: Students are asked, "What is the relationship between the length of a pendulum and the period of its swing? Between the weight at the end of the pendulum and the period?" To answer the questions, students follow a carefully scripted lab manual, taking measurements and graphing the data. They use their results to formulate an answer to the question. Example: As part of a series of lessons on global warming, the teacher asks the students to make a graph, showing how the temperature of their city has changed over the past 100 years. They are given tables of data and told to make a graph. Each student writes a paragraph, describing what his/her graph says about global warming.

Low	Lesson focuses on the step-by-step verification of information presented in a lecture or textbook. Example: Students perform an experiment to verify the formula for the period of a pendulum's swing given in a lecture the day before. They follow a carefully scripted lab manual, taking specific measurements and making specific graphs of their data. They compare their results to the information presented to them in the lecture. Example: Students read in their textbook that the temperature of the Earth is rising x degrees per decade. At the back of the book, there is a table of data on which this statement was based. Following specific instructions, students graph this data to verify the statement in their book.
Nonexistent	
Indeterminate	

Justification:

Overall. How well the series of lessons reflect a model of instruction consistent with the *NSES*.	
High	
Medium	
Low	
Nonexistent	
Indeterminate	

Justification:

Starred Pieces. Extent to which the starred artifacts portray the same impression of instruction as the full notebook.	
High	The starred artifacts give the same impression as the full notebook, with little or no loss of detail.
Medium	The starred artifacts portray a similar impression as the full notebook, but lack some detail or impart a different impression on some dimensions.
Low	The starred artifacts portray a different impression as the full notebook or fail to reveal important aspects of instruction.
Nonexistent	
Indeterminate	

Justification:

APPENDIX B

SCORING GUIDES FOR MATHEMATICS PILOT

In all dimensions...

- *Go with the highest level you can be sure you observed in the data we collected.*
- *Consider the whole series of lessons rather than any individual lessons.*
- *If you can determine some features of the dimension but not others, rate on what you do see with a note in the justification about what you could not tell.*

Collaborative Grouping. The extent to which the series of lessons uses student groups to promote the learning of mathematics. Extent to which work in groups is collaborative, addresses non-trivial tasks, and focuses on conceptual aspects of the tasks. Note: groups typically will be of varying sizes (e.g., whole class, various small groups, individual), although the structural aspect is less important than the nature of activities in groups.	
High	Students work in groups of different sizes and composition over the series of lessons. The group activities appear to foster communication, collaborative problem solving, etc. to promote the learning of mathematics. Over time, students have varying responsibilities within the groups. (NOTE—most important aspect is nature of activity that goes on in the groups; variety in size and structure is secondary.) Example: The class is divided into groups of 3 or 4 students. Students are asked to compare the average monthly temperature for 4 cities in different parts of the world. Each group decides how to represent the data in both tabular and graphic forms, and prepares an overhead transparency with its table and graph. The class reconvenes; one member of each group shows the group's transparency and explains its decisions about how to display the data. All group members participate in answering questions that their classmates raise about the table and graph. Each student then answers the following question in his or her journal: "Which of these cities has the best climate? Why do you think so?"
Medium	Students work in groups of different sizes and composition over the series of lessons. However, the group activities do not take full advantage of the learning opportunities afforded by the groups, so that opportunities for communication, collaborative problem solving, etc are limited. Example: Students are given a table with information about average monthly temperatures for 4 cities in different parts of the world, and a blank graph with the x-axis and y-axis defined. They work in groups to display the tabular information by plotting points on their graphs. They then work individually to answer the following questions in their journals: "Which of these cities has the best climate? Why do you think so?" The class reconvenes and several students volunteer to read their journal entries aloud.
Low	Students work in pairs or groups solely for the purpose of reviewing mathematical facts or procedures, or for instructional management. Example: Students work in pairs using flash cards to test each other on equivalencies between fractions and percents. Example: Students work in pairs to check each other's math notebooks for completion before turning them in.

Nonexistent	Instruction is confined to whole-class or individual work. Example: The teacher delivers a lecture on how to create a graph from a table. Students read the relevant section of their textbook and work individually on problems at the end of the section.
Indeterminate	

Justification:

Structure of Instruction. Extent to which instruction is organized to be conceptually coherent such that activities build on one another in a logical manner leading toward deeper conceptual understanding and are enacted in ways that scaffold students' current understanding.	
High	The organization and structure of the questions, tasks and activities in the series of lessons moves students along a trajectory towards deeper conceptual understanding, using information about students' current knowledge and understanding. Example: In a unit on fractions, instruction begins with a discussion on where students have seen fractions before in their everyday lives in order to elicit students' prior knowledge. The teacher then involves students in an activity where they are required to use fractions for following a recipe and figuring out the price of gasoline. The lesson culminates with a discussion of the different strategies that students used to approach and complete the activity. This lesson acts as a springboard for the subsequent days' lessons on the relationship between fractions and decimals.
Medium	The organization and structure of the questions, tasks and activities in the series of lessons address one or more related mathematical skills or concepts but either does not build toward deeper conceptual understanding or pays less attention to students' current knowledge and understanding. Example: In a unit on fractions, instruction begins with a discussion on where students have seen fractions before in their everyday lives. Then the teacher presents students with a recipe. Students are instructed to read the recipe, which includes several fractions (i.e. 1/3 cup of sugar), and answer questions about the quantities involved. Next, the teacher demonstrates how to add fractions and discusses why a common denominator is needed. The lesson culminates in an activity in which students add together two fractions and describe a situation where they might have to add fractions together.
Low	The organization and structure of the questions, tasks and activities in the series of lessons follows a pre-determined path, focusing on discrete skills and procedures. Little attention is given to students' current knowledge and understanding. Example: In a unit on fractions, instruction begins with a presentation of the algorithms needed to solve problems dealing with fractions. Students are instructed to complete worksheets using the algorithms displayed in class.
Nonexistent	
Indeterminate	

Justification:

Multiple Representations. Extent to which the series of lessons promotes the use of multiple representations (pictures, graphs, symbols, words) to illustrate ideas and concepts, as well as students' selection, application, and translation among mathematical representations to solve problems. [Note: given the specific content, you may not see all forms of representation in a given set of lessons.]	
High	Students are regularly exposed to quantitative information in a variety of forms and are expected to use multiple representations to present data and relationships, select them appropriately, and translate among them. Example: In a series of lessons on patterns and functions, the teacher presents sequences in a variety of formats, including numerical lists, geometric patterns, tables, and graphs. Students are expected to identify functions that describe the underlying numerical sequence. Students are also asked to come up with different representations for a variety of functions presented by the teacher.
Medium	Students are sometimes exposed to quantitative information in a variety of forms and sometimes use multiple representations or translate among representations in the work they produce. Example: In a series of lessons on patterns and functions, the teacher presents sequences as numerical lists and also as geometric patterns, Students are expected to write functions that describe the underlying numerical sequence. Students are also asked to come up with geometric patterns for specific functions presented by the teacher.
Low	Most presentation of numbers and relationships are done in a single form, and most of the work produced by students follows this form. Example: In a series of lessons on patterns and functions, the teacher presents numerical sequences and asks students to write functions that describe the sequence.

Justification:

Hands-On. Extent to which students participate in activities that are "hands-on". Extent to which the series of lessons affords students the opportunity to use appropriate instructional materials (including tools such as calculators, compasses, protractors, Algebra Tiles, etc.), and that these tools enable them to represent abstract mathematical ideas.	
High	Students' use of instructional materials forms a regular and integral part of instruction throughout the series of lessons. The students are encouraged to use manipulatives in ways that express important mathematical ideas and to discuss the relationships between the manipulatives and these ideas. Example: Students are engaged in the task of modeling a long-division problem. Different types of manipulatives are assigned to groups; some use base-ten blocks, some use play money, and some use loose centimeter cubes. Each group has a piece of chart paper on which they represent the way they modeled the problem. The students present their solution and the class discusses the affordances of each representation. On the following day, the students model a division problem with a remainder and discuss different ways to represent the remainder. Later in the week students create their own division story problems which other groups represent with manipulatives of their choice, explaining that choice.
Medium	Students are encouraged to use manipulatives to solve problems with little or no explicit connection made between the representation and mathematical ideas. Example: Students are asked to solve a long division problem. Students may use manipulatives to solve the problem. When most of the students are finished, the class convenes and the teacher chooses students to explain their solutions to the class. Students may comment that they chose one method of solution over another because "it's faster," but the mathematical concepts behind these choices are left undeveloped.
Low	Students are permitted to use manipulatives if they are having difficulty with a mathematics procedure. Example: Students are asked to solve a long division problem. Several students ask for the teacher's help and he suggests they try using base-ten blocks to model the problem. The teacher stays with the students and assists them in using the materials.
Nonexistent	Student use only pencil and paper, textbooks and the chalkboard during mathematics lessons.
Indeterminate	

Justification:

Cognitive Depth. Extent to which the series of lessons promotes command of the central concepts or "big ideas" of the discipline and generalizes from specific instances to larger concepts or relationships.	
High	Series of lessons focuses on deep understanding and use of central concepts (the "why" something works). • Unifying concepts are those that require students to integrate two or more smaller concepts (such as those characteristics of a "medium" classroom). Example: Students are asked to use their understandings of variable to symbolically represent the word problem "There are 6 times as many students as teachers at Lynwood School. Write a number sentence that shows the relationship between the number of students and the number of teachers." After generating an equation, each student graphs her equation and writes an explanation of the relationship and what this means practically. The teacher then leads a discussion about the relationships the students have found and how this relates to a linear relationship between 2 variables.
Medium	Series of lessons focuses on mastery of isolated concepts. The lesson may require students to explain or describe the concept but not to use it or derive it from particular cases. Thus, their understanding may be limited to repeating an essentially memorized version of the concept. • "Isolated concepts" are ideas that, while larger than procedures and formulas, do not engage students with the unifying concepts of the discipline. Example: Students are asked to represent the above word problem in an equation. The students then have to plug in 5 sets of numbers to see if their equation works. The teacher selects two or three equations as anonymous examples and leads the class in comparing the equations and determining whether they are correct.
Low	Series of lessons focuses on procedural mathematics, e.g., disconnected vocabulary, formulas, and procedural steps. These are elements of mathematics that can be memorized without requiring an understanding of the larger concepts. Students are not asked why procedures work or to generate generalizations based on the given problems. Example: The teacher defines the terms variable and linear relationship and tells the students they will be working on these concepts. Students are then given the equation $6 \times t = s$ and told that it represents the same word problem as above. The students have to plug in 5, 10, 20, 50, and 100 for t to see how many students would be at the school.
Nonexistent	
Indeterminate	

Justification:

Mathematical Communication. Extent to which the teacher and students "talk mathematics." Extent to which students are expected to communicate their mathematical thinking clearly to their peers and teacher, and to use the language of mathematics to express their ideas. Extent to which the classroom social norms foster a sense of community so that students feel free to express their ideas honestly and openly.	
High	Students are challenged to express their mathematical thinking to both other students and the teacher, orally and in writing. The use of appropriate mathematical language is considered a classroom norm. Students' ideas are solicited, explored and attended to throughout the classroom discourse. The classroom is characterized by social norms that foster a sense of community, encouraging students to express their ideas honestly and openly. Example: Students are using reallotment to find "fair prices" for different sizes and shapes of floor tile. As the students work in groups, the teacher moves around the room listening to their discussions and, at times, joining them. In answer to student questions, the teacher responds with suggestions or her own questions, keeping the focus on thinking and reasoning. Later, each group is expected to show the whole class how they used reallotment to find the prices of the tiles. The teacher encourages the use of appropriate mathematical language during this discussion. Classmates actively engage with the presenters by raising questions, challenging assumptions, and verbally reflecting on their reactions to the findings presented. The teacher asks probing questions, and pushes the thinking of both presenters and peers.
Medium	Students are expected to communicate about mathematics in the classroom, but communication is typically teacher-directed. The emphasis placed on student communication serves primarily as a way for the teacher to find out what the students are thinking. When students struggle with gaps in their thinking, the teacher asks leading questions, restates students' ideas to include missing information, or is quick to fill in the missing information. Classroom discourse can often focus more on procedural rather than conceptual issues, and the use of mathematical language is not a classroom norm. Example: Students are using reallotment to find "fair prices" for different sizes and shapes of floor tile. As the students work in groups, the teacher moves around the room listening to their discussions. When students stop her and ask for help or ask a question about the assignment, the teacher tells students how to reallot portions of the tiles in order to calculate their areas. At the end of the activity, students from each group are asked to show how they reallotted the tile areas. Their classmates listen to presentations, but do not ask questions, challenge results or react to the findings. Although students participate in the discussion, the teacher takes responsibility for developing the mathematical content. Teacher is quick to provide content if it is missing from the presentations, or asks leading questions trying to prompt presenters into filling in the missing content.

Low	The teacher transmits knowledge to the students primarily through lecture, or direct instruction. Discussions are characterized by IRE (initiation, response, evaluation) or "guess-what's-in-my-head". The focus is on answers rather than students' reasoning. Example: The teacher works on the overhead projector to show students how to use reallotment to find "fair prices" for pieces of floor tile in different sizes and shapes. As she works, she calls on students to suggest reallotment possibilities, evaluating the correctness of each student's response as it is given. All of the teacher's questions have known answers. If "correct" answers are not given, the teacher asks the question again or provides the answer.
Nonexistent	
Indeterminate	

Justification:

Explanation and Justification. Extent to which students are expected to explain and justify their reasoning and how they arrived at solutions to problems, and extent to which students' mathematical explanations and justifications incorporate conceptual, as well as computational and procedural arguments.	
High	Teachers expect students to explain their thinking and strategies. Students' explanations use generalized principles or previously proved conjectures rather than examples or an appeal to authority. Example: For the problem 125x+137=127x+135, a student explains that she knew there were two more groups of x on the right side and that 137 is two more than 135. So she simplified the equation to 2=2x. But the only way you can get the same number back in multiplication is to multiply that number by one. Therefore x has to be one. Example: A student justifies that a×b÷b=a is always true by saying that dividing a total by the same number of groups will give you one in each group. And that any number times one gets you the original number, from a previously proven conjecture.
Medium	Teachers sometimes expect students to explain their thinking and strategies. Students only sometimes provide explanations, or their explanations are usually procedural rather than conceptual. Their justifications are based more on concrete examples than generalizable principles. Example: A student explains that she subtracted 125x from both sides like she did on the previous problem. That gave her 137=2x+135. Then she subtracted 135 from both sides because she can only subtract the smaller number from the larger one. That gave her 2=2x. Next she divided 2 into both sides and that gave her 1=x. Example: In proving whether a×b÷b=a is true a student generates the example of 5×4÷4=5. But he makes no reference to conjectures or properties about dividing a number by itself or multiplying a number by one. No justification is made on whether the equation is always true or not.
Low	Student explanations are completely procedural and their justifications are strictly an appeal to authority. Example: "I subtracted the same number from both sides and divided to get one." Student explains the steps but never why he did them. Example: "It's true because the book says it is" or "it just is." "You (the teacher) said yesterday that it was true."
Nonexistent	
Indeterminate	

Justification:

Problem Solving. Extent to which instructional activities enable students to identify, apply and adapt a variety of strategies to solve problems. Extent to which problems that students solve are complex and allow for multiple solutions.	
High	Students work on problems that are complex, integrate a variety of mathematical topics, and draw upon previously learned skills. Problems lend themselves to multiple solution strategies and have multiple possible solutions. These multiple solutions and strategies are the focus of classroom discussion. Problem solving is an integral part of the class' mathematical activity, and students are regularly asked to formulate problems as well as solve them. Example: During a unit on measurement, students regularly solve problems such as: "Estimate the length of your family's car. If you lined this car up bumper to bumper with other cars of the same size, about how many car lengths would equal the length of a blue whale?" After solving the problem on their own, students compare their solutions and discuss their solution strategies. Example: At the end of a unit on ratio and proportion, pairs of students are asked to create problems for their classmates to solve. Several pairs produce complex problems such as the following: "Baseball Team A won 48 of its first 80 games. Baseball Team B won 35 of its first 50 games. Which team is doing better? "
Medium	Students regularly work on solving mathematical problems, most of which have a single correct answer. The problems incorporate one or two mathematical topics, require multiple steps for completion, and can be solved using a variety of strategies. Example: During a unit on ratio and proportion, students solve problems such as: "A baseball team won 48 of its first 80 games. How many of its next 50 games must the team win in order to maintain the ratio of wins to losses? Justify your answer." The teacher gives the right answer and students present their strategies.
Low	Problem-solving activities typically occur only at the end of instructional units or chapters. The mathematical problems that students solve address a single mathematical topic, have a single correct answer, and provide minimal opportunities for application of multiple solution strategies. Example: At the end of a textbook chapter on ratio and proportion, students solve problems such as: "A baseball team won 48 of its first 80 games. What percent of the 80 games did it win?"
Nonexistent	No evidence of students engaging in solving mathematical problems.
Indeterminate	

Justification:

Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies to support the learning of important mathematical ideas and furnish useful information to both teachers and students.	
High	Assessment has the following features: • It occurs in multiple forms. • It occurs throughout the unit. • It taps a variety of mathematics thinking processes. • It provides feedback to students. • It informs instructional practice. Example: Students in an algebra class are asked to represent graphically a race in which two contestants begin at different starting points. The students are also required to write a paragraph explaining their choice of graph and their justification for it. The teacher discovers that only two students have been able to justify their responses adequately, and that most graphs are flawed. She changes her plan for the next day's lesson and engages the class in a discussion of the various representations focusing on several specific examples from the students' work. The following day she gives students a quiz in which they are asked to explain the meaning of a graph which she provides for them.
Medium	Assessment has some but not all of the features mentioned above. Example: Students are asked to graph the same race as in the high example, but are not asked to explain their mathematical thinking. When the teacher looks at the graphs, she sees that most students were not able do the assignment. Nevertheless, she continues with a new graphing lesson on linear equations. The following day she gives students a quiz in which they are asked to explain the meaning of a graph which she provides for them.
Low	Assessment has none of the features mentioned above. Example: At the end of a unit on linear equations, the teacher gives the student a multiple-choice test. Students get their Scantron answer forms back with their grades written on the top.
Nonexistent	
Indeterminate	

Justification:

Connections/Applications. The extent to which the series of lessons helps students connect mathematics to their own experience, to the world around them, and to other disciplines.	
High	Students are regularly asked to make connections between the math they are learning in class and the math used in the world around them. They learn to apply classroom math in contexts that are relevant to their own lives. They also learn how mathematics is used in other academic disciplines. Teacher and students bring in concrete examples showing how the mathematics learned in class is being used in non-school ways. Example: In a lesson on percentages, students are engaged in a discussion around where they have seen or used percentages before. Students give the example of sales tax. The next day, a student brings to class a newspaper article discussing the sales tax. Teacher uses this article to engage students in an activity demonstrating how taxes are decided upon and how they are computed. During the lesson, one student comments that sometimes the article shows the sales tax as a percentage and at other times as a decimal. Teacher poses a final question asking students when each of the differing representations would be used and why.
Medium	Students have some opportunities to apply the mathematics they are learning to real-world settings and to other academic subjects, but this happens only occasionally or the examples are not really relevant to the students' own lives. Example: In a lesson on computing percentages, the teacher relays to students through a newspaper article that the income tax has risen. Teacher discusses that the new tax will mean that higher income families will pay an extra 3% on earning over $100,000. The teacher demonstrates how the new sales tax will be computed. Lesson culminates with an activity where students compute the new income tax on different household incomes.
Low	Students are rarely asked to make connections between the math learned in the classroom and that of the world around them or the other subjects they study. When connections are made, they are through happenstance not a planned effort on the part of the instructor. Example: In a lesson on calculating percentages, students are told to convert their percentage into a decimal and then to multiply. Students are given a worksheet of problems that require the use of this procedure. While working on the worksheet, one student shouts out that he has seen percentages before on the back of cereal boxes. The teacher confirms that percentages can be found on cereal boxes and then tells student to proceed with their worksheet.
Nonexistent	
Indeterminate	

Justification:

Overall. How well the series of lessons reflect a model of instruction consistent with the NCTM Standards. This dimension takes into account both the curriculum and the instructional practices we observe.	
High	
Medium	
Low	
Nonexistent	
Indeterminate	

Justification:

APPENDIX C

TABLES OF SCIENCE AND MATHEMATICS RATINGS BY DIMENSION, NOTEBOOK, AND RATER

Table C1

Science Ratings by Dimension, Notebook, and Rater

	Lebett		Onker		Mason		Glebe		Hammer			
	A	D	B	F	A	F	C	E	B	C	D	E
Collaborative Grouping	1	1	4	2	5	3	2	2	3	2	3	3
Materials	1	1	3	1	3	2	4	2	4	4	3	5
Assessment	2	0	1	3	1	2	1	1	1	2	2	4
Scientific Discourse	1	-	-	3	2	1	1	1	-	1	-	2
Structure of Instruction	2	2	5	3	3	3	4	2	3	3	3	3
Hands-On	1	0	3	1	3	3	2	2	3	5	5	5
Minds-On	1	1	2	3	3	2	2	1	1	1	2	1
Cognitive Depth	1	1	3	3	3	2	1	1	1	1	2	2
Inquiry	0	0	2	2	1	2	1	1	1	1	1	1
Overall	1	1	3	3	3	2	2	1	1	1	2	2
Starred Pieces	5	1	1	5	3	3	5	5	3	5	5	3

Table C2

Mathematics Ratings by Dimension, Notebook, and Rater

	Watson			Wainright			Boatman			Caputo			Hill		Lever		Mandell			Young	
	A	B	C	D	A	B	C	D	E	E	C	A	F	G	B	G	G	F	D	E	F
Collaborative Grouping	3	3	3	4	4	4	4	5	4	4	2	3	1	1	0	2	3	4	5	1	1
Structure of Instruction	4	4	4	5	3	1	3	5	4	4	4	4	3	1	2	4	3	2	3	1	2
Multiple Representations	5	5	5	4	3	1	4	5	5	3	4	3	1	1	5	5	3	3	3	3	3
Hands-On	3	0	4	5	3	0	4	5	4	3	5	4	0	1	3	4	1	4	4	3	3
Cognitive Depth	4	3	4	4	4	1	3	5	4	4	4	5	1	1	3	3	1	3	3	2	2
Mathematical Communication	3	3	3	3	3	-	3	4	4	-	4	3	1	1	1	3	1	2	3	2	3
Explanation and Justification	3	3	4	2	3	1	3	4	3	4	4	3	1	0	2	2	1	1	1	1	1
Problem Solving	3	3	3	5	3	3	2	5	3	4	5	3	0	1	2	2	3	3	4	2	3
Assessment	4	3	3	5	3	2	4	5	4	5	1	4	3	2	2	3	2	3	5	2	2
Connections/ Applications	4	1	1	5	3	1	1	5	2	4	4	4	0	1	3	5	3	4	5	1	3
Overall	4	3	4	4	3	2	4	5	4	4	4	4	2	1	2	3	2	3	3	2	2