

**Using Artifacts to Describe Instruction: Lessons Learned from Studying
Reform-Oriented Instruction in Middle School Mathematics and Science**

CSE Technical Report 705

Hilda Borko, Karin L. Kuffner, Suzanne C. Arnold, and Laura Creighton
CRESST/University of Colorado, Boulder
Brian M. Stecher, Felipe Martinez, Dionne Barnes, and Mary Lou Gilbert
CRESST/RAND

January 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education and Information Studies
University of California, Los Angeles
GSEIS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2007 The Regents of the University of California

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

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

**USING ARTIFACTS TO DESCRIBE INSTRUCTION: LESSONS LEARNED
FROM STUDYING REFORM-ORIENTED INSTRUCTION
IN MIDDLE SCHOOL MATHEMATICS AND SCIENCE¹**

**Brian Stecher, Hilda Borko, Karin L. Kuffner, Felipe Martinez, Suzanne C.
Arnold, Dionne Barnes, Laura Creighton and Mary Lou Gilbert**

Abstract

It is important to be able to describe instructional practices accurately in order to support research on “what works” in education and professional development as a basis for efforts to improve practice. This report describes a project to develop procedures for characterizing classroom practices in mathematics and science on the basis of collected classroom artifacts. A data collection tool called the “Scoop Notebook” was used to gather classroom artifacts (e.g., lesson plans, instructional materials, student work) and teacher reflections. Scoring guides were developed for rating the Notebooks (and observed classroom behaviors) along ten dimensions of reform-oriented practice in mathematics and science. Field studies were conducted in middle school science and mathematics classrooms to collect information about the reliability, validity, and feasibility of the Scoop Notebook as a measure of classroom practice. The studies yielded positive results, indicating that the Scoop Notebooks and associated scoring guides have promise for providing accurate representations of

¹ Many people contributed to the success of this project. Although they are unnamed in this report, we would like to acknowledge the importance of the 96 middle school mathematics and science teachers who assembled Scoop Notebooks as part of our pilot- and field-studies. We asked more of them than is typically required in educational research, and they responded thoughtfully and graciously. Many researchers were part of our team during the five years of the project. In addition to the named authors on this report, the following people contributed directly to the research effort: Alicia Alonzo, Daniel Battey, Victoria Deneroff, Elizabeth Dorman, Sherrie McClam, Shannon Moncure, Joi Spencer, and Alice Wood. Linda Daly from RAND produced the formatted Scoop Notebook pages, sticky notes and other project materials. We are also grateful for the patient and reliable assistance provided by Donna White and Lisa Loranger. Finally, we wish to acknowledge the support and encouragement of Eva Baker and Joan Herman, who were more than just CRESST administrators, but were interested and provocative colleagues.

selected aspects of classroom practice. The report summarizes these results and discusses lessons learned about artifact collection and scoring procedures.

Project Goals and Rationale

This report describes lessons learned from a five-year project, funded through the Center for Evaluation, Standards, and Student Testing (CRESST), to develop an alternative approach for characterizing classroom practice (Borko, Stecher, Alonzo, Moncure, & McClam, 2005; Borko et al., 2006; Stecher et al., 2005). This approach, called the “Scoop Notebook,” consists of a one- or two-week process in which teachers collect artifacts of instructional practice (e.g., lesson plans, instructional materials, student work), take photographs of the classroom set-up and learning materials, write answers to reflective questions, and assemble the results in a three-ring binder. The goal of the Scoop Notebook is to represent classroom practice well enough that a person unfamiliar with the teacher or the lessons can make valid judgments about selected features of practice on the basis of the notebook alone. The target of description for the project was “reform-oriented” practice, in particular ten dimensions of instruction that were consistent with the national mathematics and science standards and were likely to be manifest in artifacts. The project sought to answer the question, “Can accurate judgments about reform-oriented instructional practice be made based on the classroom artifacts and teacher reflections assembled in the Scoop Notebook?” This report describes the background and rationale for the project, the composition of the Scoop Notebook, the development of the dimensions of reform-oriented practice and scoring procedures, the quality of the resulting information, and the lessons learned about the use of classroom artifacts and teacher reflections as tools for describing classroom practice.

Four concerns motivated this effort to develop an alternative method for describing classroom practice. First, teachers’ actions mediate the success of school improvement efforts, so it is important to be able to describe what teachers do as they attempt to implement reforms (Ball & Rowan, 2004; Blank, Porter, & Smithson, 2001; Fullan & Miles, 1992; Mayer, 1999; Spillane, 1999). At present, the accountability provisions of the No Child Left Behind legislation are prompting schools to undertake a variety of efforts to improve student achievement (e.g., scripted curriculum materials, interim or benchmark

assessments), and information on classroom practices is key to understanding why such efforts are effective or not. Second, it is necessary to have accurate measures for understanding instructional practice to be able to support teachers in their efforts to improve. We can only provide feedback to teachers and support their attempts to reflect on and change their practice if we have accurate ways to obtain descriptions of their practice (Firestone, Mayrowetz, & Fairman, 1998). Third, the standards-based reform movement relies on claims about the value of reform-oriented practice for improving student achievement, but the research base to support these claims is limited (e.g., Laguarda, 1998; McCaffrey et al., 2001). The lack of evidence is due, in part, to the absence of an efficient, reliable procedure for gathering data about instructional practices in a large number of classrooms as a basis for examining the relationship between practice and student achievement. Researchers need reliable and efficient ways to analyze classroom practice in order to determine if those features of instructional practice that have been claimed as “making a difference,” do indeed make a difference in student learning and achievement. Fourth, research shows that surveys and case studies—the most-widely used methods for studying instructional practice—have limitations in terms of reliability and validity, particularly when they are used to measure new and evolving approaches to instruction (Burstein et al., 1995). Artifacts—naturally occurring products of instruction—seemed like a reasonable basis for developing descriptions of practice that address these concerns, particularly the latter two.

The following sections provide background information about methods that are commonly used to gather information about classrooms, including artifacts, and about the key features of reform-oriented instruction that were the focus of this study.

Methods of Collecting Information about Classroom Practice

Researchers have used many methods to acquire descriptive classroom data, including surveys, case studies, vignettes, teacher logs, and artifacts. Each method has strengths and weaknesses, which were considered as we developed the Scoop Notebook. (See Borko, Stecher, Alonzo, Moncure & McClam [2003] for an extended discussion of alternative methods.) The following paragraphs briefly summarize the key features of the different approaches.

Surveys are perhaps the most widely used methods for describing classroom practices. They are a cost-effective way to include a large number of classrooms in a study and to explore broad patterns of practice and change (Mayer, 1999; Patton, 1990). Surveys have been used to measure many aspects of practice, including coverage of subject matter, cognitive demands on students, instructional techniques and strategies, the allocation of time, and teachers' beliefs and attitudes. However, surveys have limitations. Like all self-report measures, they are subject to biases. Teachers may be reluctant to report accurately on their own behavior, coloring their responses in ways they believe to be socially desirable. Even when attempting to report correctly, respondents have imperfect memories; they may find it difficult to respond to survey questions that ask them to judge the quantity and frequency of specific actions, to recall events from earlier in the school year, or to summarize events over a long period of time. Another problem with surveys is the difficulty of describing new or evolving practices (e.g., cooperative groups, formative classroom assessment) so that both researchers and respondents have the same understanding (Antil, Jenkins, Wayne, & Vasdasy, 1998). Finally, there are some elements of classroom practice, such as interactions between teachers and students, which can only be captured through direct observation.

Case studies are in-depth investigations that typically entail extended in-person visits to classrooms to observe instruction and student learning as it occurs, as well as interviews to provide insights from the perspective of the teachers. Because of the extended engagement with each classroom, case studies are usually conducted in small numbers of sites, although some researchers exchange fewer visits to each classroom for visits to more classrooms. Case studies are widely used in educational research; Mayer (1999) points out that much of what we know about instructional practice comes from studies of only a handful of classrooms. The case study approach overcomes some of the limitations of surveys for describing instructional practices. For example, an independent observer records information about the classroom as it occurs, reducing potential biases and memory errors. Because observers can be carefully trained and monitored to recognize specific features and nuances of practice, case studies can be used in situations where teachers are attempting to implement reform-oriented teaching or other instructional innovations (Spillane & Zeuli, 1999). Yet, case studies also have limitations. The generalizability of

findings and conclusions to classrooms other than those investigated is unknown (Knapp, 1997). Case studies are also very time- and labor-intensive, so they are typically not feasible tools for policy research on the impact of large-scale educational reform efforts. Hill, Ball and colleagues have developed a rubric for coding videotapes of classroom practice, which may make case studies or other approaches that rely on classroom observation more feasible in the future (Learning Mathematics for Teaching, 2006).

Recognizing the limitations of surveys and case studies, a number of researchers have used alternative approaches to gather information about classroom practice, including daily logs and vignettes. Several researchers have asked teachers to record information about selected classroom events or interactions on a daily basis in a structured log (e.g., Burstein et al., 1995; Camburn & Barnes, 2004; Porter, Floden, Freeman, Schmidt, & Schwillie, 1988; Rowan, Camburn, & Correnti, 2004; Rowan, Harrison, & Hayes, 2004; Smithson & Porter, 1994). Various formats have been used for logs, but most focus on a few specific actions, events, and/or students to make recall easier, and most use simple selected-response questions to reduce the reporting burden on teachers. Logs work well when the researcher is interested in daily events rather than cumulative information, although repeating logs over multiple occasions may provide longer-term perspective. Logs also suffer from the potential for self-report bias and different interpretations of terms by researchers and respondents (Hill, 2005).

Other researchers have begun to explore the use of vignettes of practice to obtain insights into teachers' thinking about their instruction (Kennedy, 1999; Ma, 1999; Le et al., 2006; Stecher et al., 2006). Teachers are asked to respond to written or oral descriptions of real or hypothetical classroom events, and their responses reveal their attitudes, understanding and pedagogical skills. When used with open-ended response formats, vignettes provide an opportunity for teachers to provide detailed descriptions about the instructional strategies they use (or do not use) and to explain the decisions they make when planning and implementing their lessons. Such insights into teachers' thinking are unlikely to be revealed by surveys, case studies, or logs. However, vignette-based methods rely exclusively on teacher self-report, so respondents can color their responses to suit the perceived desires of the researcher.

More recently, several researchers have incorporated instructional artifacts into their studies of classroom practices (e.g., Burstein et al., 1995; Clare & Aschbacher, 2001; Clare, Valdes, Pascal, & Steinberg, 2001; Matsumura, Garnier, Pascal, & Valdes, 2002; Matsumura, Slater, Wolf, Crosson, Levison, Peterson, & Resnick, 2005; McDonnell & Choisser, 1997; Resnick, Matsumura, & Junker, 2006; Ruiz-Primo, Li, & Shavelson, 2002). Artifacts are raw records of classroom practice, which reveal teachers' instructional efforts and students' learning products. 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. The materials are usually specified by researchers based on their interest (e.g., all assignments in a specified time period, students' science notebooks, "typical" and "challenging" language arts assignments). In many cases, researchers have been able to make valid judgment about practice on the basis of artifacts. For example, sampled language arts assignments yielded a reliable and valid measure of the quality of classroom assignments in language arts, which was positively related to the quality of observed instruction and student work (Clare, 2000). However, there are also limitations to the use of artifacts. They place an additional burden on teachers, who must save, copy, assemble, and often annotate classroom work products. The artifacts by themselves are often not clear, and they require additional documentation or interpretation from the teacher. Furthermore, artifacts reveal very little about instructional interactions between teachers and students.

The Scoop Notebook (which is described in the next section entitled "Characterizing Reform-Oriented Practice") represents a hybrid of several methods, designed to maximize their good features while minimizing their shortcomings. The Scoop Notebook asks teachers to collect all materials associated with a lesson (e.g., worksheets, overheads, in-class work and assignments) on a daily basis. Teachers annotate the artifacts to clarify the instructional context. The use of artifacts is designed to capture practice broadly while reducing self-report bias. As in the case of logs, teachers are asked to react only to each day's events, which are recent and familiar. Teachers are also asked to provide daily reflections on the lesson. Like vignettes, teachers are prompted by specific questions about their intentions for the lesson, reactions to it, and plans to follow up; and they give open-ended responses revealing their thoughts and beliefs in their own words. Other elements of the notebook, such as the

calendar and the photographs, are included to address specific information needs. The design of the notebook reflects both a concern about high-quality data collection and an interest in a particular approach to instruction.

Focus On Reform-Oriented Instruction

The study was motivated not only by the goal of developing a good descriptive instrument but also by the desire to describe a particular set of instructional practices, i.e., those associated with the visions of reform-oriented classrooms portrayed in the *National Science Education Standards* (National Research Council [NRC], 1996) and *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000). These national standards and the curriculum materials that were produced to support them embody a particular approach to teaching mathematics and science that encompasses both content (“what is taught”) and pedagogy (“how it is taught”). For example, reform-oriented science emphasizes skills, such as observation, measurement, and experimentation, that characterize the actual work of scientists; and instruction is designed to help students learn how to ask and pursue questions, construct and test explanations, form arguments, and communicate their ideas with others (NRC, 1996). Similarly, reform-oriented mathematics emphasizes mathematical thinking and reasoning skills, problem solving, and interconnections among mathematical concepts; and the associated instruction should help students learn to make conjectures, develop deductive arguments, and use mathematics to model and predict real-world phenomena (NCTM, 2000).

To accomplish these goals, teachers are encouraged to create learning communities that mimic the ways in which mathematicians and scientists develop knowledge and learn from one another. For example, students should be provided with opportunities to express their understandings and work with their peers to defend and dispute their ideas. In reform-oriented mathematics and science classrooms, less emphasis is placed on students’ abilities to memorize information or follow pre-set procedures to arrive at a solution; greater emphasis is placed on students’ abilities to explain and justify a claim, to communicate their thinking, to design and implement a systematic approach to answer a question, or to use tools and resources to assist their learning.

This study focused on the presence or absence of selected features of reform-oriented instruction. To be useful as a basis for describing classrooms, this broad, comprehensive vision of reform-oriented practice had to be translated into concrete, operational terms. As a starting point, we used the elements of reform-oriented practice identified by panels of science and mathematics experts in an earlier study (Le et al., 2006; Stecher et al., 2002). That study identified about two-dozen elements of instruction in each subject, and contrasted reform and non-reform manifestations of each element. For the purpose of this study, the list was narrowed to focus on aspects of reform-oriented teaching that might be instantiated in artifacts, such as subject matter content or assessment; aspects that were unlikely to be characterized via the contents of the Scoop Notebook were eliminated. This consideration led to the identification of ten dimensions of practice in science, and similar but not identical 10 dimensions in mathematics (which are described in the section entitled “The Scoop Notebook”).

In reality, there was considerable back and forth in the development of the Scoop Notebook and the dimensions of instructional practice. The process of refining the dimensions helped us to make decisions about the types of artifacts and reflective questions to include in the Scoop Notebook. The process of generating the notebook procedures helped us decide which dimensions to include and how to describe them clearly. The goal of the study was not to produce a comprehensive description of reform oriented instruction, nor to describe every aspect of classroom practice, but to test whether selected elements of practice could be described accurately using an artifact-based procedure.

Classroom observation was used to test the quality of the description provided by the notebooks, i.e., judgments based on the notebooks were compared to judgments based on direct observation of classes. In addition, we obtained “Gold Standard” judgments based on all the information available from observations and notebooks combined, and we compared these judgments to the other two sets.

Research Questions

The project was guided by three main research questions:

- 1) Can raters make reliable judgments about selected dimensions of reform-oriented instructional practice on the basis of the Scoop Notebook and on the basis of classroom observations?

- 2) Are raters' judgments of reform-oriented practice based on Scoop Notebooks similar to their judgments based on classroom observations? Specifically,
 - Do scores assigned on the basis of the Scoop Notebook agree with those assigned on the basis of classroom observations (and with "Gold Standard" scores that use all available information about a classroom)?
 - Are the relationships among ratings on the dimensions of reform-oriented instruction consistent across notebooks and observations?
- 3) Is it feasible to use artifact collection and scoring procedures on a large scale to characterize classroom practice?

Characterizing Reform-Oriented Practice

As noted in the previous section, this study adopted reform-oriented mathematics and science instruction as its pedagogical focus. We used previous research as a basis for identifying 10 dimensions of reform-oriented instructional practice in each content area. The science dimensions reflect the emphasis in the science standards (NRC, 1996) on the active engagement by students in inquiry-based activities in which they explain and justify their scientific thinking both verbally and in writing, to develop a knowledge base in science. The mathematics dimensions reflect the focus in the mathematics standards (NCTM, 2000) on students in mathematics classrooms solving problems with multiple solutions and solution strategies, explaining and justifying their solutions, and communicating their mathematical thinking to others.

We developed short definitions of the dimensions as well as longer, more detailed scoring guides. We found that both definitions and scoring guides were necessary to describe the elements of reform-oriented instruction clearly enough to be used to characterize classroom practice on the basis of either the Scoop Notebook or classroom observations.

We made changes to the definitions and scoring guides during the pilot phase and between the field tests, in order to improve the reliability and validity of our measures. The following section contains the final definitions of the 10

dimensions of instructional practice in science and mathematics.² The subsequent section describes and illustrates the scoring guides derived from these definitions.

Science

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on scientific tasks that are directly related to the scientific goals of the lesson, and to enable students to work together to complete these activities. An active teacher role in facilitating group interactions is not necessary.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent, such that activities are related scientifically and build on one another in a logical manner.

3. Use of Scientific Resources. The extent to which a variety of scientific resources (e.g., computer software, internet resources, video materials, laboratory equipment and supplies, scientific tools, print materials) permeate the learning environment and are integral to the series of lessons. These resources could be handled by the teacher and/or the students, but the lesson is meant to engage all students. By variety we mean different types of resources OR variety within a type of scientific resource.

4. “Hands-On”. The extent to which students participate in activities that allow them to physically engage with scientific phenomena by handling materials and scientific equipment.

5. Inquiry. The extent to which the series of lessons involves the students actively engaged in posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.

6. Cognitive Depth. Cognitive depth refers to a focus on the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and making connections and relationships among science concepts. This dimension includes two aspects of cognitive depth: lesson design

² We also revised the scoring guides over the course of the project. Only the final versions are included in this report. Earlier versions are available in the reports of the mathematics and science field studies (Borko et al., 2006; Stecher et al., 2005), or from the authors upon request.

and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently promotes cognitive depth.

7. Scientific Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their scientific ideas honestly and openly. The extent to which the teacher and students “talk science,” and students are expected to communicate their scientific thinking clearly to their peers and teacher, both orally and in writing, using the language of science.

8. Explanation/Justification. The extent to which the teacher expects, and students provide, explanations/justifications, both orally and on written assignments.

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important scientific ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making).

10. Connections/Applications. The extent to which the series of lessons helps students connect science to their own experience and the world around them, apply science to real world contexts, or understand the role of science in society (e.g., how science can be used to inform social policy).

Mathematics

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on mathematical tasks that are directly related to the mathematical goals of the lesson and to enable students to work together to complete these activities. An active teacher role in facilitating group interactions is not necessary.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related mathematically and build on one another in a logical manner.

3. 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. The extent to which students select, use, and

translate among (go back and forth between) mathematical representations in an appropriate manner.

4. Use of Mathematical Tools. The extent to which the series of lessons affords students the opportunity to use appropriate mathematical tools (e.g., calculators, compasses, protractors, Algebra Tiles), and that these tools enable them to represent abstract mathematical ideas.

5. Cognitive Depth. Cognitive depth refers to command of the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and making connections and relationships among mathematics concepts. This dimension includes two aspects of cognitive depth: lesson design and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently and effectively promotes cognitive depth.

6. Mathematical Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their mathematical ideas honestly and openly. The extent to which the teacher and students “talk mathematics,” and students are expected to communicate their mathematical thinking clearly to their peers and teacher, both orally and in writing, using the language of mathematics.

7. Explanation/Justification. The extent to which the teacher expects and students provide explanations/justifications, both orally and on written assignments.

8. 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.

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important mathematical ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making).

10. 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. The extent to which the series of lessons helps

students apply mathematics to real world contexts and to problems in other disciplines.

Scoring Guides

In order to be able to rate a teacher's instructional practice on each dimension we created detailed scoring guides for mathematics (see Appendix A) and science (see Appendix B). Each dimension is rated on a five-point scale, ranging from *low* (1) to *high* (5). The associated scoring rubric contains both detailed general descriptions of each level and specific classroom examples of *high* (5), *medium* (3), and *low* (1) practice. For most dimensions, the intermediate ratings, *medium-high* (4) or *medium-low* (2), are not described and do not have classroom examples.

In addition to the 10 dimensions of instructional practice, raters were asked to assign an Overall Reform rating representing the rater's holistic impression of the instructional practices of the teacher. The rating is not an average of the 10 dimensions but represents a cumulative answer to the question: "How well does the series of lessons reflect a model of instruction consistent with dimensions previously described, taking into account both the curriculum and the instructional practices?"

Figure 1 shows an example of the scoring guide for a science dimension ("hands-on") and Figure 2 shows an example of the scoring guide for a mathematics dimension (multiple representations).

4. “Hands-On”. The extent to which students participate in activities that allow them to physically engage with scientific phenomena by handling materials and scientific equipment.

NOTE: The emphasis is on direct observation and interaction with scientific equipment and physical objects, to address the substance of the science lesson. Acting out a scientific phenomenon does count. Computers don’t unless use involves equipment such as probes.

High: During a series of lessons, all students have regular opportunities to work with materials and scientific equipment.

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.

Medium: During a series of lessons, some of the students work regularly with materials or scientific equipment OR all students work with materials or scientific equipment but only occasionally.

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. Later in the unit, students supplement their reading about faults by using wooden blocks to represent different fault types.

Low: There are no activities that require students to handle or work with materials or scientific equipment (other than pencil and paper).

Figure 1. Sample dimension from science scoring guide.

3. 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. The extent to which students select, use, and translate among (go back and forth between) mathematical representations in an appropriate manner.

NOTE: dimension includes both exposure (by teacher or curriculum) and use by students.

High: Students are regularly exposed to quantitative information in a variety of forms. Students use multiple representations to present data and relationships, select representations appropriate to the tasks, and translate among them.

Example: In a lesson 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, and to indicate which representations are most appropriate for answering different questions about the mathematical functions.

Medium: Use of multiple representations has some but not all of the features mentioned above. Students are sometimes exposed to quantitative information in a variety of forms. Students sometimes use multiple representations, select representations appropriate to the tasks, or translate among them.

Example: In a lesson 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

Figure 2. Sample dimension from mathematics scoring guide.

The scoring sheet that accompanied the guide also asked raters to provide a justification for each rating, i.e., a brief written description of the evidence used for arriving at the given rating. For example, a rater who assigned a value of “3” on the multiple representations dimension might list examples of artifacts that incorporated the use of multiple representations, as well as artifacts that did not. A complete justification would also explain how the evidence corresponded to the rating criteria, e.g., “a rating of 3 was assigned because the teacher and curriculum appeared to promote the use of multiple representations but there was no evidence of use by students.” When the raters had problems assigning a rating, they often described the inconsistencies in the evidence that caused problems. For example, a rater might report that the teacher’s reflections indicated the use of multiple representations by students, but that there was no evidence of multiple representations in the examples of student work.

The 11 dimensions of instructional practice in science and mathematics discussed in this section (including the Overall Reform dimension) were also considered during the development of the Scoop Notebook procedures for mathematics and science classrooms, which are described in the next section.

The Scoop Notebook

The goal of our studies was to develop a procedure to collect instructional artifacts and teacher reflections that would allow us to characterize a series of lessons in manner similar to an observer in the classroom. The guiding question was whether similar judgments about the extent to which the lessons embody the 10 dimensions of reform-oriented instructional practice could be obtained from artifacts and from direct observation. We sought to collect classroom materials—such as lesson plans, student work, photographs of the classroom—supplemented by teachers’ answers to reflection questions that would support reliable judgments about practice. We developed the Scoop Notebook 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 it back to their laboratories for analysis, we planned to “scoop” materials from classrooms to be examined later. 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.

There were two main challenges we faced in the development of the Scoop Notebook. First, we needed to make our expectations clear enough to the teachers so that they would collect information that was revealing of their practice in ways that we could understand. Clarity was also important to insure that we obtained comparable materials from all participating teachers, not assemblages that reflected teachers' individual preferences about artifacts. To communicate clearly to teachers, the notebook instructions were written with some redundancy. They contained an initial short overview outlining all the pieces. The overview was followed by detailed instructions for each of the elements. The notebook was divided into sections, and separate instructions were also included at the beginning of each section.

Second, we wanted to collect enough information to allow us to make accurate inferences about each of the 10 dimensions of classroom practice, without over-burdening participating teachers. The desire for completeness thus had to be balanced against the practical realities of asking teachers to retrieve, copy, annotate, and assemble artifacts in the Scoop Notebook. We attempted to strike an appropriate balance by asking teachers to collect information for a unit of instruction lasting five to seven days and by specifying reasonable numbers of assignments, student responses, etc.

Anticipating that notebook development would require some trial and error, we conducted two pilot studies with small numbers of mathematics and science teachers before embarking on the main study (Borko et al., 2005). As expected, a few important changes were made as a result of the pilot studies. The directions were re-organized and rewritten a number of times to make them easier to understand. Sample answers were provided to accompany the pre-Scoop reflection questions to model the kinds of information and level of detail we wanted. In order to reduce the burden on teachers, we reduced the number of days on which artifacts were collected from seven to five, the number of assignments from all assignments to three, and the number of examples of student work from each assignment from four to three. The final version of the notebook used in the main study and described here reflects these changes.

For the most part, the mathematics and science Scoop Notebooks are the same. They differ primarily in terms of the examples included as illustrations in the instructions. An example in the mathematics notebook might be “directions for pattern and function activity;” while in the science notebook it would be “directions for lab activity.” Only the science Scoop Notebook is included in this document (see Appendix C); the mathematics notebook will be available in a forthcoming CRESST report.

Contents of the Scoop Notebook

We packaged the Scoop Notebook as a three-ring binder, consisting of the following main components (see Appendix C):

- project overview
- teacher directions for collecting and labeling artifacts
- materials for assembling the notebook

The first section of the notebook contains the project overview, which briefly introduces the teachers to the Scoop Notebook. We present the analogy of a scientist “scooping” materials for examination and the rationale for the assembly of a Scoop Notebook. These introductory pages provide the teachers with a checklist highlighting the procedures to follow before, during, and after the Scoop collection period. In addition, there is a final checklist for teachers to review before handing in the assembled Scoop Notebook.

The second section provides the detailed instructions for collecting the classroom Scoop. This section includes explicit instructions and examples of how to:

- select a class and timeframe for the Scoop
- complete the daily calendar
- take photographs and complete the photograph log
- collect classroom artifacts and daily instructional materials
- select student work, assignments and a formal classroom assessment

- label daily instructional materials and student work
- respond to daily reflection questions

The third section of the notebook contains all the materials teachers need for assembling the Scoop Notebook:

- the pre-Scoop, daily and post- Scoop reflection questions, with accompanying examples
- the daily calendar form
- the photograph log
- pocket folders for classroom artifacts (one for each day of the Scoop)
- a pocket folder for student work and an assessment example

In addition to these three sections, the Scoop Notebook contains a zipper pocket with sticky notes for labeling the artifacts and student work, a disposable camera, and a cassette tape. The cassette tape is provided for use by teachers who prefer to audio tape the reflections, rather than write them by hand or computer.

Procedures for Assembling the Scoop Notebook

Detailed directions in the Scoop Notebook explain which artifacts to collect and how to label them. The directions also instruct the teachers to answer reflection questions prior to, during, and after the Scoop period.

Artifact collection instructions. Directions in the notebook ask teachers to collect three categories of artifacts: materials generated prior to class (e.g., lesson plans, handouts, scoring rubrics), materials generated during class (e.g., writing on the board or overheads, student work), and materials generated outside of class (e.g., student homework, projects). The teachers are encouraged to include any other instructional artifacts not specifically mentioned in the directions. They are asked to label each artifact with a “Classroom Artifact” sticky note indicating what the artifact is (e.g., copy of overhead, lesson, lab activity).

In addition, teachers are requested to make entries in the daily calendar, briefly describing the length of the lesson, the topic, and the instructional materials used. A disposable camera is included in the Scoop Notebook for

teachers 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 cannot be included in the notebook (e.g., posters and 3-dimensional projects prepared by students). Teachers also fill out a photograph log, identifying the subject and date of each picture.

Teachers are also asked to select three different instances of student-generated work (e.g., in-class assignments, homework). For each selection, they are asked to collect three examples representing a range of student work from high to low quality. Because we are interested in teachers' judgments about the quality of student work, directions specify that their selections should be based on the quality of the work rather than the ability of the students. Teachers are instructed to make an independent selection of student work for each assignment, rather than tracking the same students throughout the artifact collection process. The Scoop Notebook contains "Student Work" sticky notes to be filled out and attached to each example of student work. The sticky note asks the teachers to rate the quality of the work (high, medium, or low), to describe the reason for giving that rating, and to explain what they learned about the student's understanding of the material from the work.

Finally, teachers are asked to include a recent formal classroom assessment task (e.g., test, quiz, prompt or task description for a portfolio piece, paper, performance, final project, demonstration) that is representative of the assessments they used.

Reflections. In addition to collecting instructional artifacts, teachers are asked to respond to three different sets of reflective questions. These questions attempt to elicit information about a teacher's classroom practice which classroom artifacts alone might not provide (e.g., the context of the series of lessons within the curriculum, a description of student interactions during a lesson). Prior to the beginning of the Scoop period teachers respond to these pre-Scoop reflection questions:

- 1) What about the context of your teaching situation is important for us to know in order to understand the lessons you will include in the Scoop?
- 2) What does a typical lesson look like in your classroom?
- 3) How often do you assess student learning, and what strategies/tools do you use?

- 4) What are your overall plans for the set of lessons that will be included in the Scoop?

During the Scoop period, teachers respond to the following daily reflection questions “as soon as possible” after completion of each lesson:

- 1) What were your objectives/expectations for student learning during this lesson?
- 2) Describe the lesson in enough detail so we understand how the Scoop materials were used or generated.
- 3) Thinking back to your original plans for the lesson, were there any changes in how the lesson actually unfolded?
- 4) How well were your objectives/expectations for student learning met in today’s lesson? How do you know?
- 5) Will today’s class session affect your plans for tomorrow (or later in the “unit”)? If so, how?
- 6) Is there anything else you would like us to know about this lesson that you feel was not captured by the Scoop?
- 7) Have you completed the Daily Calendar and Photograph Log entries for today?

After the conclusion of the Scoop period, teachers answer the following post-Scoop questions:

- 1) How does this series of lessons fit in with your long-term goals for this group of students?
- 2) How representative of your typical instruction was this series of lessons (with respect to content, instructional strategies and student activities)? What aspects were typical? What aspects were not typical?
- 3) How well does this collection of artifacts, photographs, and reflections capture what it is like to learn science in your classroom? How “true-to-life” is the picture of your teaching portrayed by the Scoop?
- 4) If you were preparing this notebook to help someone understand your teaching, what else would you want the notebook to include? Why?

Finally, at the end of the Scoop period, teachers are asked to make sure their notebooks are complete by comparing the materials against a final checklist. After the completion of the Scoop period, a member of the research team contacted the school and made arrangements to retrieve the completed notebook, camera, and other materials from the teacher.

Field Study Procedures and Analysis Methods

This research investigated use of artifacts to measure reform-oriented instructional practices in middle-school mathematics and science. The Scoop Notebook, our vehicle for collecting artifacts, entailed a product and a process, and both were analyzed as part of this study. The product is the collection of artifacts and additional materials teachers prepared at our request that were assembled in the Scoop Notebook and formed the basis for judgments about reform-oriented practice. The process aspect of the notebook consists of the directions, designed to elicit similar information from all teachers, and the scoring procedures, designed to yield comparable judgments from all readers. The research studies explored the quality of the judgments based on the artifacts as well as the quality of the procedures used to obtain them.

Initially, we conducted pilot studies in a small number of middle school science and mathematics classrooms (Borko, Stecher, Alonzo, Moncure, & McClam, 2003; 2005). The pilot studies were primarily formative in nature, i.e., they provided information to enable us to improve our data collection and scoring procedures. Insights from the pilot studies led to changes in the Scoop Notebook and scoring guides used in the field studies.

Subsequently, we conducted three field studies to evaluate the Scoop Notebooks in middle school classrooms in Colorado and California—one in mathematics (Stecher, Borko, Kuffner, Wood, Arnold, Gilbert, & Dorman, 2005) and two in science (Borko et al., 2006). Procedures were similar although not identical in the three field studies. This report draws upon results from the three field studies to present a picture of the reliability and validity of the Scoop Notebook ratings as well as insights into the notebook collection process.

Participants

Ninety-six teachers from Colorado and California participated in Scoop Notebook field studies between the spring of 2003 and spring of 2005 (see Table 1). A few teachers participated in multiple field studies. The sample was selected in a multi-step process. First, a small number of school districts in the Los Angeles and Denver metropolitan areas were approached and agreed to participate in the study. Taken together the districts in each geographic area contained a diverse set of schools with respect to size, student demographic

characteristics, student academic performance, and approach to teaching mathematics and science. Second, a diverse subset of 8-12 middle schools was identified in each state. We informed the school principals about the study and asked for permission to invite teachers to participate. Where possible we selected some schools known to be using reform-oriented curricula and others known to be using more traditional curricula. The purpose of this selection criterion was to insure that a range of teaching practices were represented in the final sample. Third, teachers were contacted directly to solicit volunteers to participate in the study. In most cases, researchers visited the schools and described the study to groups of mathematics or science teachers. In some cases, information was distributed in school mailboxes. No screening was done at the teacher level; all volunteers were included. Teachers were paid a \$200 honorarium for participating in the study and completing the notebooks.

Table 1
Classrooms Sampled for Validation Studies

| Subject | Date | Location | Number of Teachers / Scoop Notebooks |
|----------------|----------------|-----------------|---|
| Mathematics | Spring 2003 | CA | 13 |
| | | CO | 23 |
| Science | Fall 2003 | CA | 16 |
| | | CO | 23 |
| Science | Spring 2005 | CA | 11 |
| | | CO | 10 |

Data Collection

Two primary sources of data were used in this study: the Scoop Notebook and classroom observations.³ A member of the research team met individually or

³ A small sample of classrooms (n=7 for the mathematics field study and n=8 for the first science field study) was audiotaped during the days they were observed. The teachers wore wireless lapel microphones to record their voices as well as those of students in close proximity. The audiotapes were later transcribed to provide a record of classroom discourse. These transcripts were rated with the same scales and scoring guides and compared to the information obtained from notebooks and observations. Given the small sample of audiotaped classrooms the results

in small groups with participating teachers to go over the Scoop Notebook and procedures. When we described the Scoop Notebook to participating teachers, we framed the task in terms of the question: “What is it like to learn mathematics/science in your classroom?” Teachers were told that they would be assembling a variety of materials that captured different aspects of classroom practice, including samples of student work, lesson plans, photographs, and teachers’ responses to reflective questions. They were asked to select a class comprised of students who were fairly typical of all the students they taught, and to pick a series of lessons that was fairly typical of instruction in that class. They were instructed to 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 non-traditional scheduling were instructed to “scoop” for an amount of instructional time approximately equivalent to five days on a normal schedule. The researcher gave each teacher a copy of the notebook, and reviewed the contents with them (as noted in the previous section, the Scoop Notebook contained detailed directions to teachers for collecting and assembling data). Feedback from teachers during these meetings indicated that the directions in the notebook were clear and comprehensive, and teachers reported that they understood the procedures.

Each classroom that participated in the study was observed by one or two members of the research team on two or three occasions during the period the Scoop was being collected. In most cases, the same person observed the class on all occasions, providing an extended exposure to classroom practices. During each classroom visit, the observer wrote open-ended field notes describing the classroom, the subject matter addressed during the lesson, and the interactions that occurred.

Scoring

The scoring guides described in the section on “Characterizing Reform-Oriented Practice” were used for rating the Scoop Notebooks, the classroom observations, and the audiotape transcripts. Rater training occurred prior to each rating session to ensure that raters were using the scoring guides in a similar

can only be considered as suggestive. They are available in the technical reports of each field study (Borko et al., 2006; Stecher et al., 2005).

manner. During these calibration sessions, participants reviewed and rated the same material, compared scores, discussed differences, and reached consensus on the application of the rating criteria. When necessary, revisions were made to clarify the scoring guide. Notebook and transcript raters were calibrated using Scoop Notebooks that were set aside for this purpose; observers were calibrated using videotapes of mathematics or science lessons.

Notebooks were read and rated by two or three researchers who were not familiar with the teacher or the classroom. Their ratings were based solely on the materials in the Notebook. The readers assigned a value on each dimension and an Overall Reform rating. They also wrote justifications for each rating based on the evidence in the Notebook. During two of the validation studies, the readers also assigned a rating on a five-point scale for completeness (the degree to which the teacher included all requested materials in the Scoop Notebook) and confidence (the rater's level of confidence in assigning values on each dimension).

Observations were rated using the same scales as the notebooks. Ratings were assigned each day on the basis of that day's lesson, and a Summary Observation rating was completed after the series of observations. Field notes taken during the observation were used to review the sequence and events of the observed lesson and inform decisions about the value assigned to each dimension.

Finally, researchers who observed in classrooms were given access to the Scoop Notebook and were asked to rate the classroom again on the basis of a combination of information sources—classroom field notes, daily and Summary Observation ratings, and the materials in the Scoop Notebook. These “Gold Standard” ratings represented our best estimate of the “true” status of the lessons on the ten dimensions reform-oriented practice.

Data Analysis Procedures

Reliability. We employed two complimentary approaches to assess the reliability of ratings of the 11 dimensions of instructional practice in science and mathematics. First, for each dimension, agreement indices were computed as simple 0-100 percentages reflecting the proportion of time that raters agreed in their judgments of instructional practice. The level of inter-rater agreement

provides initial evidence of consistency (i.e., the extent to which multiple raters agree on their assessment of individual teachers' practices) and can help pinpoint teachers or raters with unusual or problematic ratings.

However, agreement indices do not provide a sense of measurement precision—i.e., the confidence that could be placed on any individual teacher's rating. For this purpose, reliability indices provide an assessment of the accuracy of ratings, and the extent and sources of measurement error involved in judging an individual teacher's practice. These 0 to 1 coefficients indicate the extent to which the observed ratings reflect measurement error, or correspond closely to teachers' true scores. For each dimension the reliability indices were estimated by means of generalizability theory designs where the object of measurement (the classroom or teacher) receives ratings from multiple raters (in the case of notebook ratings), or multiple raters and multiple observations (for observation ratings).⁴ Technically the design consists of an object of measurement crossed with one facet (raters) or two facets (raters and observations). Estimates of the variance associated with each of these facets were obtained using the VARCOMP procedure in the statistical package SAS (SAS Institute Inc., 2003). The variance components were then used to estimate reliability coefficients.

Two kinds of reliability coefficients were estimated: generalizability coefficients—which are appropriate for situations where interest centers on ordering or ranking teachers relative to each other (i.e., relative decisions, or norm-referenced interpretations of a score), and dependability coefficients—which are appropriate where a judgment of individual performance with reference to a standard is sought (i.e., absolute decisions, or criterion-referenced interpretations). With the Scoop Notebook, interest centers on assessing individual teachers' instructional practice in the classroom with reference to the criteria specified on the rating guides; consequently we present only absolute reliability (i.e., dependability) coefficients here.

Generalizability theory further allows researchers to estimate the levels of reliability that would be obtained under different hypothetical scenarios (i.e., varying numbers of raters and observation occasions). In evaluating the

⁴ Strictly speaking, the design is *incomplete*, because all raters could not rate all notebooks or visit all classrooms. In the mathematics field study only one observer visited each classroom and thus a quantitative indicator of reliability for ratings based on classroom observations could not be computed.

reliability of ratings of instructional practice based on classroom observations it is important to consider that for most dimensions some day-to-day variation in teacher instructional practice is expected: for example, while students in a reform-oriented science classroom will gain experience manipulating materials and substances, it is not expected that such hands-on activities would be carried out every day during the school year. Thus, an assessment of instructional practice based on a single visit to the classroom during the school year would naturally involve a great deal of uncertainty (i.e., measurement error). To address this concern, we collected information for several days of instruction. In practice, however, daily fluctuations in instruction are not captured directly in the notebook ratings because raters considered the evidence in the notebook as a whole and assigned a single score for each dimension for the whole period. Thus, the notebook reliability estimate is not affected by day-to-day variance in the ratings. On the other hand, classroom observations cover a similar slice of time, but raters assign both individual scores after each visit and a Summary Observation score at the end of the week. Comparing the reliability of ratings based on the Scoop Notebook to that of ratings based on individual (daily) observations would unfairly favor the Notebook. For the purposes of our study, Summary Observation ratings (assigned for each dimension after multiple visits to the classrooms) are the most appropriate comparison to notebook ratings because both present a single global assessment based on evidence collected throughout the entire series of lessons.

Validity. To investigate the validity of ratings of reform-oriented instructional practice in mathematics and science based on the Scoop Notebook we considered multiple sources of evidence. First, we analyzed the pattern of inter-correlation among the dimensions of reform instruction in science and mathematics (i.e., their internal structure or dimensionality) to determine the extent to which they reflect a common trait, in this case reform-oriented instruction. We conducted Factor Analysis (using the SAS software with promax rotation) and compared the internal structure of ratings based on the notebook to the internal structure of ratings based on classroom observations; attention here centers on the proportion of variance accounted for by the first factor, and the loadings of the dimensions on the first and subsequent extracted factors.

Second, we considered evidence of convergent validity based on comparisons to other measures. We investigated the degree of correspondence

between ratings based on the notebook and ratings based on other sources of information, specifically Summary Observation ratings and Gold Standard ratings. For each of these comparisons, preliminary evidence was obtained by analyzing the level of agreement between the two ratings being compared. Then we considered the correlation between the two ratings. We computed a Pearson correlation between the two ratings for each dimension, across all teachers. These correlations indicate the strength of the linear relationship between the two sources of information. We also computed Pearson correlations between ratings for each teacher across dimensions to obtain a general sense of the linear relationship between overall judgments of a classroom obtained using both sources of information.

Finally, we investigate the degree to which ratings of instructional practice based on the Scoop Notebook reflect differences known a-priori to exist between California and Colorado in the curricula and standards used in the two states. This approach rests on the notion that a reliable and valid measure of any kind should be able to detect important differences known to exist between two subjects (in this case states, although this could be extended to teachers or schools) in the trait being measured.

Analyses of the Scoop process. In addition to investigating the reliability and validity of ratings of the notebook, we examined how well the Scoop Notebook procedures operated and how well the various components of the notebook functioned in practice. To answer these questions, we conducted post-hoc analyses of the quantity (in this case, completeness) and quality (in this case, usability) of the artifacts that were collected, we analyzed teachers' responses to questions about the notebook, and we reviewed raters' reflections on the scoring process.

To investigate quantity/completeness we reviewed the contents of the 96 Scoop Notebooks collected during the five years of the project. For each notebook, we tallied the presence or absence of each product (e.g., weekly calendar, photo logs) or artifact (e.g., assignments, or assessments), and the number of examples of each artifact included in the notebook. From these data, we computed the percentage of notebooks containing each product as well as the average number of examples of each artifact.

To examine quality/usability, pairs of researchers were assigned to review each type of product or artifact and answer the questions: How well did this component work? How might it be made to work better? The pair of researchers reviewed a random sample of 20 notebooks, took notes on what they found, discussed their observations together, and prepared a summary with thoughts about effectiveness and improvement. In addition, individual researchers reflected on their experience rating the Scoop Notebooks over the series of field studies and wrote memoranda detailing their ideas for improvement.

Teachers' judgments about the notebooks were obtained from their answers to two of the post-Scoop reflection questions, "How well does this collection of artifacts, photographs, and reflections capture what it is like to learn mathematics/science in your classroom?" and "If you were preparing this notebook to help someone understand your teaching, what else would you want the notebook to include and why?" We analyzed these responses in the same 20 percent random sample of notebooks used for the previous analysis.

In addition, members of the research team rated the usefulness of the 11 sources of information contained in the notebooks for making judgments about the 10 dimensions of practice. Each source of information was rated on a three-point scale (0=not helpful, 1=somewhat helpful, 2=very helpful) reflecting the extent to which it was helpful in rating each of the 10 dimensions of practice. We then computed the average perceived usefulness of the various sources of information for rating each dimension of practice.

Finally, researchers' perspectives on the scoring procedures were obtained from group discussions and individual reflections. The group discussions focused primarily on ways to improve the descriptions of the dimensions of reform-oriented practice and the associated scoring guides. Some discussions occurred during rater calibration/training sessions and resulted in revisions to the scoring guide during the pilot and field studies. In addition, the research team met for approximately two days at the end of the project and reviewed all the evidence we had obtained on each dimension, including all versions of the dimension descriptions and scoring guide, information about the technical quality of each dimension from each of the field studies, and all of the researchers' own experience observing classes and rating notebooks on the dimension. The next section presents the results of the analyses investigating the reliability and validity of ratings of the Scoop Notebook. We follow it up with a

look at the results of our post-hoc investigation of the notebook collection and scoring procedures in the section entitled “Effectiveness of the Scoop Notebook Procedures and Scoring Process.” The final section concludes this report with an exploration of the implications of these results, and the lessons that can be extracted for research and practice.

Reliability and Validity of Scoop Ratings

This section summarizes the results of three field studies conducted to determine the extent to which the Scoop Notebook can be used to produce reliable and valid assessments or judgments of the use of reform-oriented instructional practices in mathematics and science classrooms. As described in the previous section, we employed a range of complementary methodologies to address the research questions, including an examination of agreement and reliability indices, factor analytic techniques to address questions of dimensionality, and comparisons and correlations among different measures of the same classroom. We present the results organized by the research question they address: we first consider the results pertaining to the reliability of ratings based on the Scoop Notebook and classroom observations. Second, we address the validity question, presenting results that investigate the dimensionality of notebook and observation ratings, and the relationships among them. At each stage we discuss the main pattern of results observed across studies, presenting results specific to science or mathematics when notable differences exist between subjects; two technical reports for the mathematics and science field studies (Borko et al., 2006; Stecher et al., 2005) contain the complete reliability and validity results and additional analyses for each subject.

Reliability of Ratings of the Scoop Notebook

The first research question was whether the Scoop Notebooks could be scored with adequate reliability. For most dimensions in the mathematics and science notebooks the results point to adequate to high levels of reliability. Across dimensions, inter-rater agreement within 1 point was consistently over 75 percent; reliability (dependability) coefficients for the hypothetical average over three raters were consistently over 0.70, and often over 0.80 (see Figure 3).⁵

⁵ As noted in previous sections, the definitions of some dimensions and descriptions in the accompanying scoring guides changed slightly between studies. The section entitled

A few dimensions were rated with lower levels of reliability. Lower reliability was typically associated with a lack of variance in the ratings. This was the case with the Structure of Lessons dimension, for example: Nearly all teachers were judged to structure their lessons appropriately and rated with values of 4 or 5, and consequently there is little variance in the ratings. In this situation the scale contains very little information about *true* differences between teachers to begin with, so the reliability of any measure for distinguishing one teacher from another is understandably low. Similarly, there was little variability in ratings of Assessment for the mathematics notebook, resulting in low reliability for that dimension.

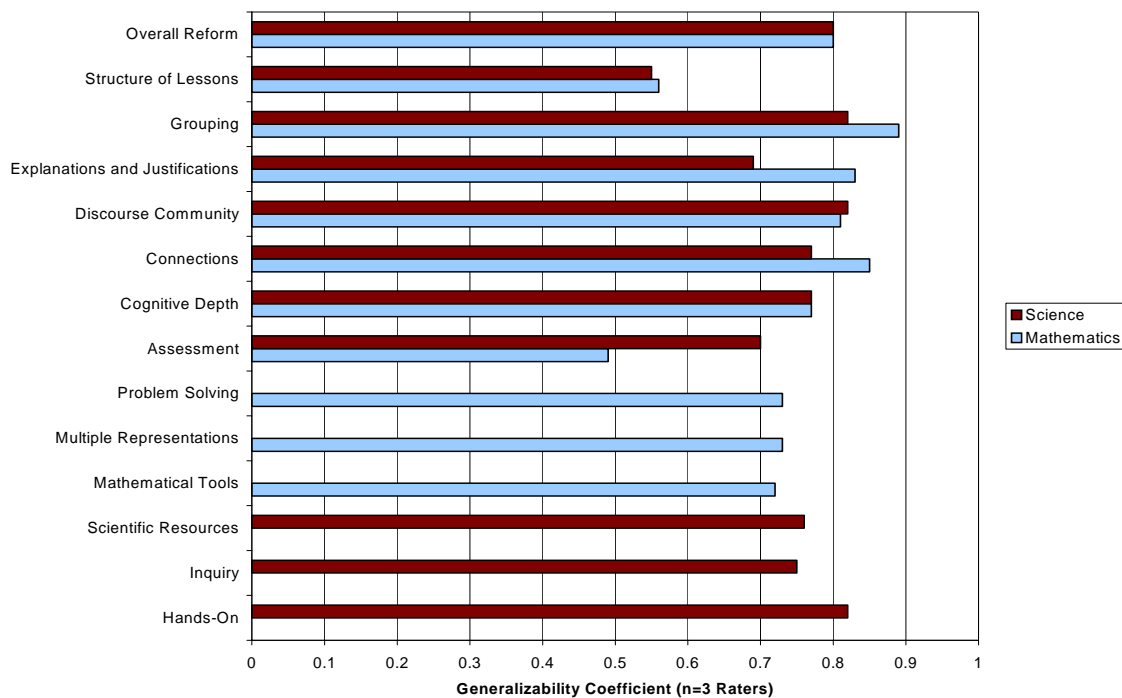


Figure 3. Reliability of Ratings of the Mathematics and Science Scoop Notebooks

There is no conclusive evidence that reliability was higher for either the mathematics or the science notebooks. Of eight dimensions that were shared by

“Characterizing Reform-Oriented Practice” presented the final versions of the dimensions after all changes resulting from the pilot studies, data analysis, and researchers’ and teachers’ insights. However, findings related to specific dimensions in this chapter refer to the definitions in use at the time of that particular study. These definitions and associated scoring guides are available in the technical reports for the mathematics and science field studies (Stecher et al., 2005; Borko et al., 2006).

both subjects, in six we observed comparable levels of reliability in mathematics and science. For these dimensions the differences across subjects (and studies) were under 0.10, and for the Overall Reform dimension a reliability of 0.80 was achieved in both subjects – that is, the science and mathematics notebooks exhibit equal levels of reliability in capturing the extent to which the series of lessons reflects a model of reform-oriented instructional practice consistent with that described by all the dimensions in the notebook taken together.

The dimensions with the most noticeable differences across subjects are Assessment, and Explanation/Justification. Assessment was rated considerably more reliably in the science notebook (0.70) compared to mathematics (0.49). A slight variation in the definition of this dimension across studies could partly explain the difference. In the mathematics study the dimension referred to the use of “a variety of formal and informal assessment strategies to support the learning of important mathematical ideas”; in the science study the description of the assessment dimensions was modified to say “to measure student understanding of important scientific ideas”. This change clarified and narrowed the dimension, likely leading to the improved reliability observed in the science study.

With Explanation/Justification the reverse pattern was observed: While the reliability of ratings was over 0.80 in mathematics, it decreased to slightly under 0.70 in the science study. An explanation for this divergence is less apparent in the definition of the dimension across studies, because only minor wording changes were made after the mathematics study. Nevertheless, these changes could have resulted in greater ambiguity in the distinction between Explanation/Justification and Cognitive Depth in the science study.

Reliability of Ratings Based on Classroom Observations (Science) ⁶

The reliability of Summary Observation ratings assigned after three classroom visits was adequate or high for most dimensions of instructional practice in science: The results in Figure 4 indicate that for eight of 11 dimensions of instructional practice adequate, reliability of 0.70 or higher was achieved. As

⁶ The goal of the mathematics study was to estimate the reliability of notebook ratings, and therefore the design did not provide for multiple raters visiting each classroom; consequently the reliability of observation ratings can only be investigated fully for science classrooms.

with notebook ratings, the dimensions that exhibited lower levels of reliability (Explanation/Justification, Discourse Community, and Assessment) were among those with the least variance in the ratings. As before, the lack of true variance between teachers for these dimensions reduces the ability of any rating or score to distinguish among them (i.e., its reliability).

As noted in the section entitled “Field Study Procedures and Analysis Methods,” Summary Observation ratings are the most appropriate point of comparison for notebook ratings because both take into account the whole of the information collected during the entire observation (and scooping) period. However, one limitation of these global ratings is that neither provides useful information to investigate the optimal number of days of observation (or scooping) needed to achieve adequate levels of reliability. On the other hand, individual observation (daily) ratings assigned after each visit to the classroom can be used to gain insights into this question. We used daily ratings to estimate the reliability that would be achieved with different hypothetical numbers of classroom visits. For each dimension of instructional practice in science, Figure 5 shows the projected level of absolute reliability (dependability) for increasing numbers of classroom observations.

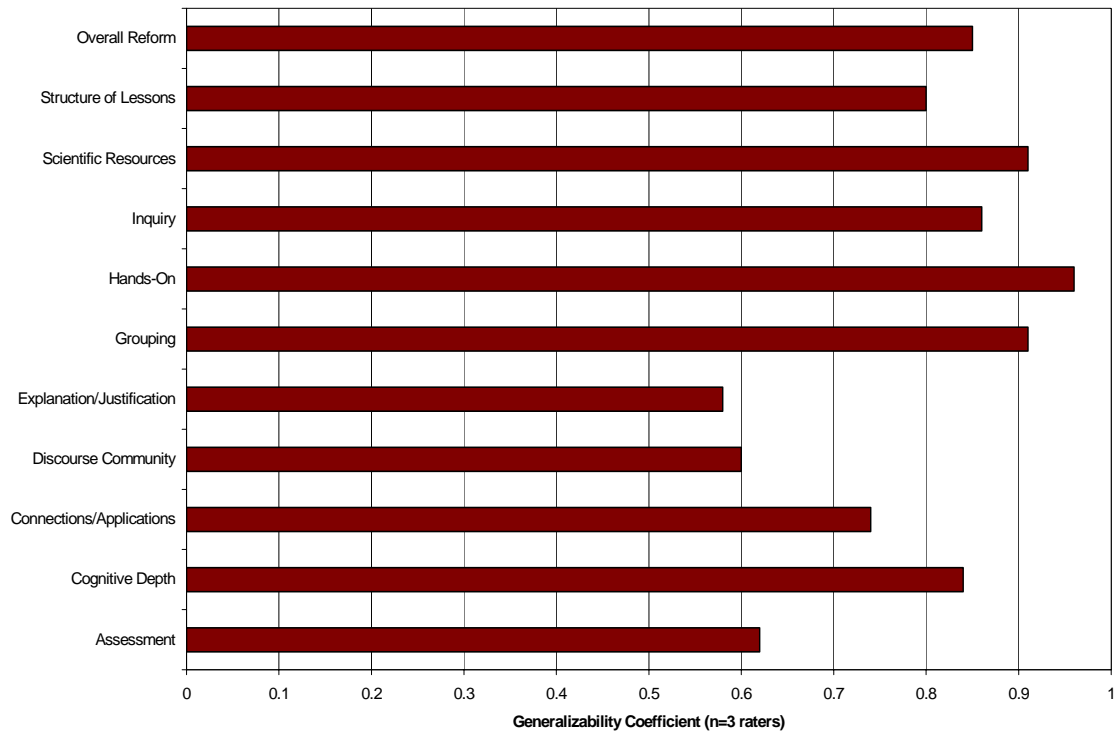


Figure 4. Reliability of Summary Classroom Observation Ratings (Science)

One important question of interest when considering the use of artifacts is how many days of “scooping” are optimal to achieve useful and reliable insights about instructional practice, without imposing an unreasonable burden on teachers. We cannot answer this question directly, because all notebooks were collected for a similar period of time (five to seven days), and it was not possible to disaggregate the notebooks and score the artifacts on a daily basis. However, classroom observations were rated on a daily basis, and the information in Figure 5 provides related evidence suggesting that for most dimensions (and for the Overall Reform dimension in particular) the incremental gains in terms of reliability increased considerably for the first four or five observations; after that additional observations did not produce sizeable gains in the reliability of the ratings. This suggests that a five-day period of “scooping” may be reasonable for most individual dimensions and the Overall Reform dimension.

Importantly, it does not follow from this analysis that adequate reliability will be attained after five observations (or a five-day period of scooping) in every

case. As Figure 5 shows, for some dimensions with large day-to-day variance (e.g., Grouping, Hands-on) additional observations could be justified. In addition, other sources of measurement error not related to the number of occasions a classroom is observed (e.g., raters, interactions between raters and features of the classroom or notebook) could result in low levels of reliability for some dimensions even after five observations (or five days of scooped materials).

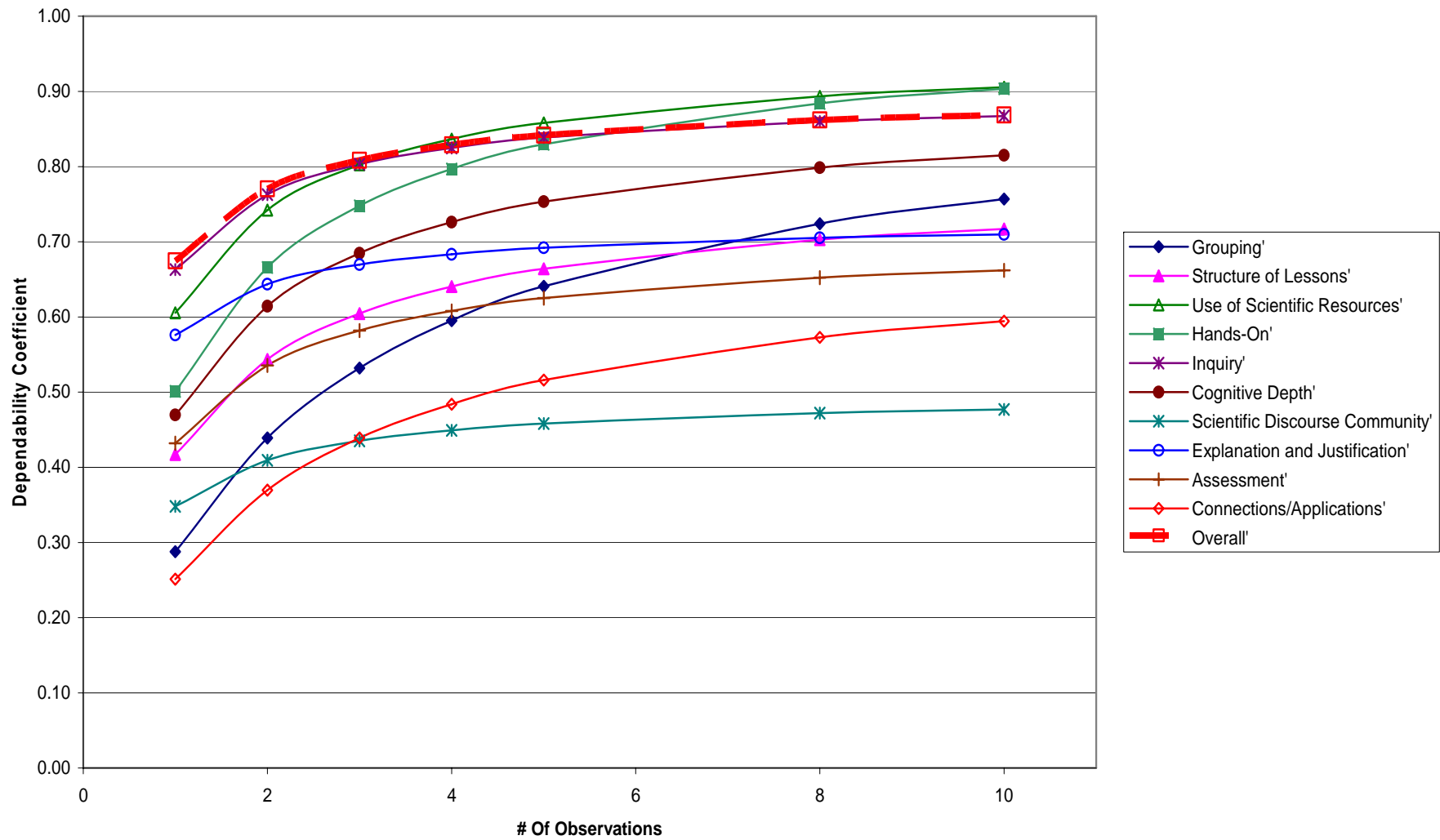


Figure 5. Dependability of Science Classroom Observation Ratings (by number of observations; n=3 raters assumed)

Comparing the Reliability of Observation and Notebook Ratings (Science)

Figure 6 indicates that the reliability of notebook ratings for each dimension of reform instruction in science was generally adequate and comparable to that of Summary Observation ratings. The reliability of ratings for the Overall Reform dimension was over 0.80 with both notebooks and classroom observations. Similarly, for Grouping, Connections/Applications, and Cognitive Depth the reliability of notebook and classroom observation ratings was nearly identical.

For four dimensions (Hands-on, Inquiry, Scientific Resources, and Structure of Lessons) reliability was somewhat higher for Summary Observation ratings compared to notebook ratings. This finding likely reflects the fact that these facets of practice stand out when observing in classes and can be more difficult to discern on the basis of the notebook. For example, Hands-On instruction is quite apparent in class but evidence about it might be more indirect in the notebook, e.g., a copy of a worksheet. Furthermore, two readers might not interpret the worksheet in the same manner when rating Hands-On instruction. The same is true for Scientific Resources.

For the other three dimensions (Assessment, Discourse Community, and Explanation/Justification) the reliability of notebook ratings was higher than that of ratings based on classroom observations. Assessment might be more consistently gauged on the basis of the notebook because the notebook contains examples of quizzes or tests drawn from the entire Scoop period, and teacher reflections discuss it explicitly. On the other hand, classroom observers might not be in class on the day an assessment activity is carried out and might respond differently to indirect clues about a previous test or quiz.

In addition, some caution should be used in comparing the reliability of ratings of instructional practice in science based on notebooks and observations because the data come from different studies (Fall 2003 and Spring 2005). As mentioned previously, some dimensions were modified slightly between studies. As a result, observed differences in reliability between notebook and observation ratings could reflect real differences between data collection methods, changes in the dimensions across studies, or a combination of both. However, we believe it is reasonable to compare the reliability of notebooks and observations because, for most dimensions, the definitions and scoring guides changed only minimally. Further, both studies sampled classrooms from common pools in the two states (in some cases the same

classrooms participated). Despite these cautions, the general trend in the results is clear: the reliability of ratings based on notebooks and observations was adequately large, and for most dimensions the differences across methods were small.

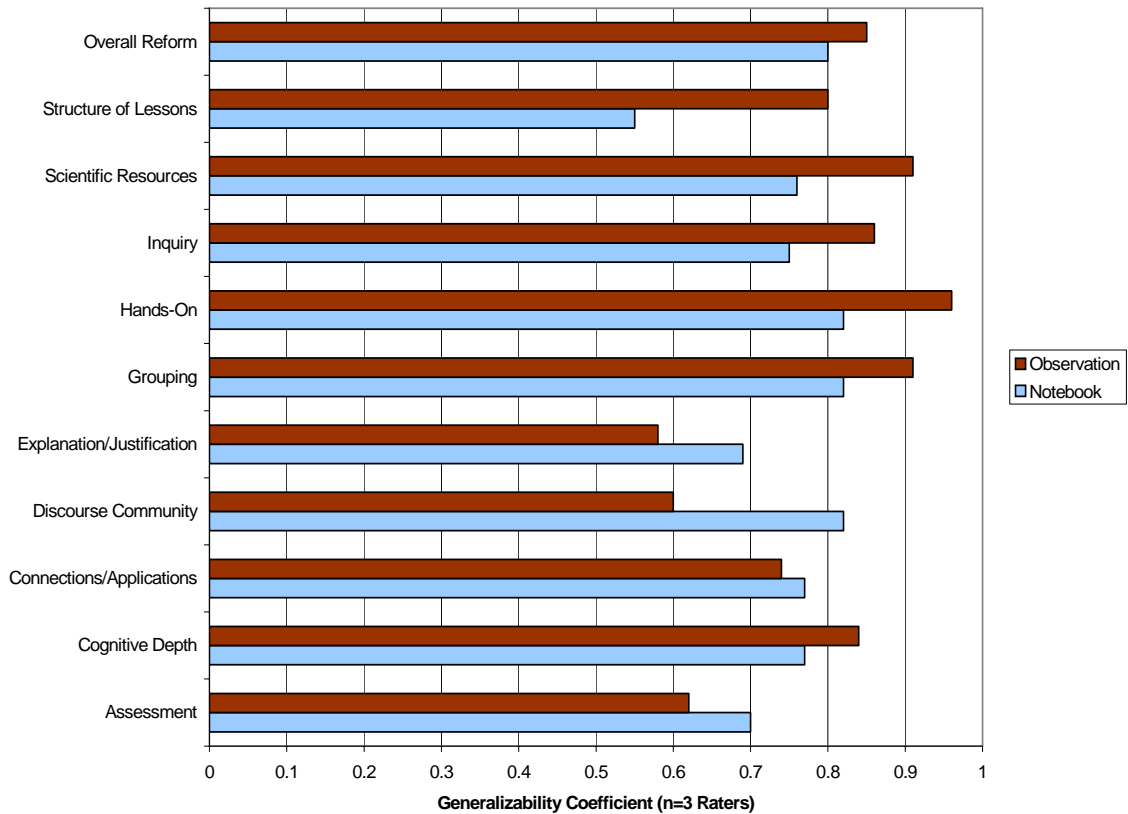


Figure 6. Reliability of Notebook and Summary Observation Ratings (Science)

Validity of Ratings of the Scoop Notebook

We bring multiple sources of evidence to bear when investigating the validity of ratings based in the Scoop Notebook (AERA, APA, NCME, 1998). First, we investigate the internal structure of the ratings to determine the degree to which a common trait or factor underlies the 11 dimensions of reform instruction, as measured by notebooks and observations. As noted before, evidence of a dominant common factor would provide indication that the different dimensions of instructional practice relate (or conform) to a common reform instruction dimension. Second, we investigate the correlations between notebook ratings and ratings of the same classrooms based on different sources of information. High correlations would

provide evidence of convergent validity of the notebook (in relation to the other data collection methods). Finally, we investigate the extent to which notebook ratings reflect differences known a-priori between the curricula and instructional practice in the two states where data were collected (California and Colorado).

Dimensionality of Notebook and Observation Ratings

Factor analysis of the notebook and observation ratings in mathematics produced similar results indicating that a dominant factor can be identified underlying the 10 dimensions of reform instructional practice. This reform-based instruction factor accounts for 56 percent and 67 percent of the total variance in notebook and observation ratings, respectively. Thus, with both methods of data collection it would be feasible to condense the separate ratings for each of 10 dimensions into a single global index of reform instruction in mathematics classrooms. Nevertheless, additional factors are needed to fully explain the variance of some of the dimensions of reform practice in mathematics: As shown in Table 2, some distinctive aspects of classroom practice exhibit lower levels of cohesiveness with (i.e., a weaker loading on) the dominant reform instruction factor. This trend is more evident with notebook ratings than observation ratings. In particular, Cognitive Depth and Problem Solving stand out as dimensions of reform instruction in mathematics that may be best described separately from the others.

In science it is also possible to identify a common reform trait underlying the 10 dimensions of instructional practice. However, the dominant trait in science accounts for only 49 percent and 42 percent of the total variance in notebook and observation ratings, respectively—a smaller proportion than that observed in mathematics. While a single index of reform instruction in science can also be computed, the various dimensions of reform practice may not be as closely interrelated in science as in mathematics. Moreover, as was the case with mathematics, Table 2 indicates that some of the dimensions of teacher practice in science are less closely related to the dominant reform instruction trait than others. With observation ratings in particular, low loadings were observed for Grouping, Hands-On, and Structure of Lessons. If interest centers on a specific dimension, a global index (or similarly, the Overall Reform dimension) would not fully capture all the relevant features of instruction and classroom environment, and it would be preferable to consider the ratings for those dimensions separately.

Table 2
 Factor Loadings for Notebook and Summary Observation Ratings in Mathematics and Science.

| Dimension | Mathematics | | Science | |
|--------------------------------|-------------|--------------|------------|--------------|
| | Note-books | Observations | Note-books | Observations |
| Mathematical Tools | 0.95 | 0.97 | | |
| Multiple Representations | 0.86 | 0.90 | | |
| Problem Solving | 0.55 | 0.75 | | |
| Hands-On | | | 0.58 | 0.45 |
| Inquiry | | | 0.82 | 0.64 |
| Scientific Resources | | | 0.59 | 0.66 |
| Assessment | 0.77 | 0.66 | 0.80 | 0.59 |
| Cognitive Depth | 0.34 | 0.90 | 0.81 | 0.83 |
| Connections | 0.71 | 0.80 | 0.60 | 0.57 |
| Discourse Community | 0.58 | 0.69 | 0.81 | 0.79 |
| Explanation/Justification | 0.89 | 0.89 | 0.76 | 0.70 |
| Grouping | 0.77 | 0.80 | 0.76 | 0.37 |
| Structure of Lessons | 0.65 | 0.58 | 0.36 | 0.47 |
| Percent variance accounted for | 56% | 67% | 49% | 42% |

Thus, in both mathematics and science there is evidence of unidimensionality indicating that the individual dimensions of instructional practice conform to (or are part of) a more general dimension that we have termed Reform-Oriented instruction. One interpretation of this pattern is that Reform-Oriented instructional practices, although distinct from one another, are compatible and hence often found in the same classrooms: Each of our dimensions encompasses several practices, and some practices are associated with more than one dimension. At the same time, each dimension has distinctly unique characteristics that differentiate it from the others. Thus, it might be possible to distill a general reform construct, with a few additional dimensions reflecting unique aspects of classroom practice that exhibited lower

levels of cohesiveness with the dominant factors. Of course, for some purposes, such as staff development, it might be important to distinguish among the dimensions.

Finally, it is interesting to note that Factor Analysis of notebooks and classroom observation ratings produced similar results, lending support both to the idea of a dominant reform factor, and to the notion that the two data collection procedures provide similar evidence about instructional practice traits.

Convergent Validity

Table 3 shows the correlation between ratings based on the Scoop Notebook and those based on classroom observations. In general, the table reflects moderate to high correlations between ratings of the same classroom based on the two sources of information. In mathematics the correlations were 0.65 or higher for nine of eleven dimensions, which indicates that both information sources tend to produce similar judgments of teacher practice.

In mathematics, the dimensions with the lowest correlations were Assessment (0.37) and Structure of Lessons (0.38); this could partly reflect the low reliability in the notebook ratings for these dimensions (see Figure 3). In addition, it may be a result of specific features of each of the dimensions. This possibility, and its implications for future development of the Scoop Notebook, is considered further in the concluding section of this report.

The low correlation for Structure of Lessons likely reflects the high skewness in the distribution of that variable—i.e., most teachers attained the highest ratings. In addition, the scoring guide was changed between the scoring of observations and notebooks; the criterion that the activities in the series of lessons lead toward deeper conceptual understanding was eliminated. This slight change in the description of the dimension could have also decreased the correspondence among ratings.

In science classrooms, the correlations between notebook and observation ratings were in the moderate range and tended to be lower than those observed in mathematics: For five dimensions the correlations were 0.60 or higher—and over 0.50 for 10 dimensions. The lowest correlation was again for Structure of Lessons (0.26) which, as with mathematics, is likely related to the highly skewed distribution and low reliability of ratings of this dimension.

Table 3
Correlation Between Notebook and Summary Observation Ratings.

| Dimension | Pearson Correlation | |
|---------------------------|---------------------|---------|
| | Mathematics | Science |
| Mathematical Tools | 0.70 | |
| Multiple Representations | 0.78 | |
| Problem Solving | 0.66 | |
| Hands-On | | 0.76 |
| Inquiry | | 0.69 |
| Scientific Resources | | 0.55 |
| Assessment | 0.37 | 0.54 |
| Cognitive Depth | 0.70 | 0.53 |
| Connections | 0.65 | 0.55 |
| Discourse Community | 0.69 | 0.64 |
| Explanation/Justification | 0.75 | 0.62 |
| Grouping | 0.86 | 0.61 |
| Structure of Lessons | 0.38 | 0.26 |
| Overall Reform | 0.78 | 0.57 |

We computed an overall correlation coefficient to indicate the correspondence between notebook and observation ratings across dimensions. For each classroom we computed the average notebook rating and average observation rating averaging across the 11 dimensions (10 dimensions plus Overall Reform). In mathematics the correlation between these classroom averages based on notebooks and observations was very high at 0.85. In the science study this correlation decreased noticeably to 0.71, although it remained high in absolute terms. Thus, the evidence points to a considerable degree of convergent validity between broad judgments of the extent of reform-oriented practice in individual classrooms made on the basis of notebooks and observations.

Additional evidence of validity can be obtained by investigating the correlation between Notebook and Gold Standard (GS) ratings. Days after assigning their final observation ratings, observers also assigned a Gold Standard rating taking into account both their observations and the materials and reflections collected by the teacher in the notebook. Table 4 shows that the correlations between notebook and

GS ratings were in the moderate to high range for most dimensions (as before, Assessment and Structure of Lessons exhibit the lowest correlations). On the other hand, the correlations between observation and GS ratings were consistently very high.⁷ This suggests that when information available from the two data sources does not overlap, raters may be more persuaded by, or inclined to rely on, information collected through their own observations than on the information collected by the teacher in the notebooks.

Table 4
Correlation Between Notebook and Gold Standard Ratings.

| Dimension | Pearson Correlation | |
|---------------------------|---------------------|---------|
| | Mathematics | Science |
| Mathematical Tools | 0.82 | |
| Multiple Representations | 0.70 | |
| Problem Solving | 0.74 | |
| Hands-On | | 0.85 |
| Inquiry | | 0.62 |
| Scientific Resources | | 0.59 |
| Assessment | 0.43 | 0.54 |
| Cognitive Depth | 0.81 | 0.41 |
| Connections | 0.78 | 0.70 |
| Discourse Community | 0.79 | 0.70 |
| Explanation/Justification | 0.80 | 0.54 |
| Grouping | 0.87 | 0.67 |
| Structure of Lessons | 0.45 | 0.26 |
| Overall Reform | 0.81 | 0.59 |

Divergent Validity: Differences Between California and Colorado

The last source of validity information for the Scoop Notebook arises from differences in the curriculum between California and Colorado. Our *a priori* observation was that the mathematics curriculum and textbooks in use in Colorado

⁷ For all dimensions the correlations were 0.8 or higher. The results are not presented here in detail; for complete results refer to the technical reports for the mathematics and science field studies (Borko et al., 2006; Stecher et al., 2005).

were more reform-oriented than those in use in California. If the notebooks are capturing the teachers' instructional approaches accurately, then any state-to-state differences in reform-oriented practices should be reflected in the ratings assigned to the Colorado and California notebooks. In mathematics, the information collected through the notebooks (as well as classroom observations) confirmed our *a priori* perception that the textbooks and curricula used by the Colorado teachers embody more of the reform principles than the textbooks and curricula used by teachers in California. That is, Colorado notebooks received consistently higher ratings, particularly on the dimensions most closely associated with reform curricula (detailed results are available in Stecher et al., 2005). These results provide further evidence of the validity of ratings of the Scoop Notebook.

Effectiveness of the Scoop Notebook Procedures and Scoring Process

The previous section focused on the scores assigned to the Scoop Notebooks—were they reliable and did they provide a valid description of reform-oriented teaching? This section discusses the notebook collection and scoring procedures—were they effective, and how might they be improved? Data to answer these questions come from tallies of notebook contents, responses from teachers to evaluative questions, feedback from members of the research team, and results reported in the preceding section. The first part of the section focuses on the procedures we used to collect data from teachers; the second part discusses the usefulness of different data sources for rating the notebooks; the third part examines insights gained during the scoring process. Brief recommendations for modifications to the Scoop Notebook are included where appropriate.

Effectiveness of the Scoop Notebook Data Collection Procedures

Information about the quantity and quality of the materials provided by teachers came from three sources: a review of the contents of the notebooks, teachers' answers to questions about the Scoop portrayal of their instructional practices, and insights from readers who scored many notebooks. These sources were analyzed to answer two questions:

- 1) How complete were the notebooks with respect to each type of artifact or material?
- 2) What suggestions did researchers and teachers offer for improving each element of the Scoop Notebook?

The answers to these questions are presented separately for artifacts, constructed materials, and teacher reflections.

Artifacts of Practice. Artifacts of practice include instructional materials, assignments, assessments, and student work that were part of normal classroom activities. These materials are artifacts in the usual sense of the word, i.e., normal elements of practice that existed independent of the study, which were copied and made available for research purposes.

Teachers provided most of the artifacts of practice we requested. Eighty-nine percent of the notebooks contained at least one annotated artifact. On average, there were nine assignments or other artifacts per notebook. The artifacts of practice typically included worksheets (that students completed during the lesson), handouts (with information, tables, diagrams, or descriptions of procedures), quizzes, copies of textbook pages, and copies of overhead transparencies.

All notebooks but one also contained examples of student work. On average, each notebook had four examples of student work deemed to be of high quality, and three each of middle and low quality. Most of the student work had been done by individual students, but some notebooks included work completed by groups of students. On the other hand, only 58 percent of teachers provided examples of classroom assessments, and only 39 percent included completed assessments showing students' answers. The dearth of assessments made it difficult to rate some classrooms on the Assessment dimension. Failure to include assessments in all the notebooks may have been due to a lack of clarity and emphasis in the Scoop instructions, the fact that no assessments were given during the Scoop period, or an unwillingness on the part of teachers to share assessments.

In general, lesson plans, worksheets and assignments were easy to read and provided useful information about the content of the lessons. However, photocopies of overhead transparencies, student class work, and homework were sometimes difficult to decipher. In addition, some teachers failed to annotate materials (such as pages from textbooks or supplemental materials) using yellow sticky labels so it was difficult to tell when or how these materials were used during the lesson (discussed below). Similarly, it was often difficult to tell whether student work had been done in class, at home, or both. This distinction was probably not critical in interpreting the work, but it was an obstacle to making sense out of the collected artifacts.

Clearer directions and instruction might improve the quality of copied material and annotations.

Constructed Materials. This category includes descriptive materials constructed by teachers solely for the purposes of the notebook (e.g., weekly calendar, photographs and photo log, sticky labels). In general, teachers were good at providing the constructed materials requested as part of the study.

Ninety percent of the teachers completed the weekly calendar. Most entries were clear, although a few teachers were quite brief and terse. The calendar was helpful in providing a broad view of the Scoop period, particularly the sequence of lessons, the use of student groups, and any school events or special activities that interrupted the normal daily pattern of lessons. However, the calendar entries were too sketchy to provide details about lesson demands or instructional practices.

Eighty-nine percent of the teachers completed the photograph log and included classroom photographs. On average, each teacher who provided photographs took 17 pictures; the number of photographs ranged from 5 to 49; and only 6 teachers failed to include any photographs. Fortunately, the photographs were usually easy to interpret, despite the fact that many teachers did not complete the photograph logs very thoroughly. Most teachers remembered to use the flash, and their photographs provided valuable information about the organization of the classroom, the availability of materials, the equipment used for specific lessons, etc. Even when teachers failed to use the flash, the grainy images still revealed some useful information about these features of instruction.

Ninety-three percent of the teachers labeled the student work they gathered with white sticky labels. In almost all cases, teachers explained on the label why they rated the student work as they had and what the work revealed about student understanding. However, in many cases teachers' comments were not as revealing as we anticipated. For example, in response to the question, "Why did you rate it this way?" teachers often provided comments that we could have deduced from the work sample itself, such as "didn't answer the question," or "this student clearly understands." Similarly, in response to the question, "What does this tell you about student understanding?" teachers often wrote, "understands well," or "needs extra help." In contrast, some teachers did provide the kinds of insights into learning and teaching we were hoping to prompt, such as "Lack connection between data and graphs. More work on this required." or "Needs to write down all math steps to

have an understanding.” Perhaps the first question could be eliminated and the second refined to make the information more helpful.

Similarly, most of the assignments/materials included in the notebooks had yellow sticky labels affixed. In many cases, the labels were of limited value because they did not add information that was not already apparent from the work itself. In the case of overhead transparencies, teacher generated notes and in-class materials, the stickies were often useful for identifying the dates and lesson context.

Teacher Reflections. Teachers were asked to answer reflective questions prior to the start of the collection period, during the Scoop data collection, and afterwards; in general, they answered these questions clearly and completely.

Pre-Scoop Reflections. Over 95 percent of teachers answered the pre-Scoop questions, which included the context of teaching, typical lessons, their use of assessments, and their plans for the Scoop period. Readers reported that teachers’ answers to the pre-Scoop questions contained useful information for interpreting the contents of the Scoop Notebooks and making judgments about the dimensions of reform-oriented practice. For example, when describing the context, some teachers discussed the ability level of their students, the availability of equipment and materials, holidays and special events that might affect classroom activities during the Scoop period, etc. Responses to the question about typical lessons revealed information about the organization of lessons and the kinds of activities students engaged in; they also often revealed information about teachers’ expectations for student behavior and student learning. For example, Teacher 40 described his inclinations toward having students work in groups as part of a discussion of seating arrangements, “sometimes students choose own seats and sometimes assigned—always sit in groups—good for students to learn to work with a wide variety of students.” Teacher 90 described using crossword puzzles as a type of formative assessment of key concepts. “For them to get the correct word, based on the clue, they really have to know the concepts.”

In some cases, teachers went far beyond our questions and discussed their interactions with students, their feelings about the school, and their opinions about the community. We offered teachers the opportunity to audio-tape responses, and the few teachers who used this method often provided pages of information, both relevant and incidental.

Daily Reflections. Response rates to the teacher reflection questions remained high during the Scoop period; 82 percent of teachers provided answers to all five daily reflection questions on each of the days they collected materials for the notebook. The other teachers skipped some of the questions on some of the days, but only three teachers skipped most of the questions on most of the days.

Both the quantity and the quality of the daily reflections varied among teachers. At their worst, teachers could be terse, incomplete, or non-responsive. Some offered rambling comments that reflected their concerns at the moment rather than the questions we asked. At their best, the responses were quite revealing of many elements of the lesson. For example, in answering the first two daily questions about long-term goals and representativeness,⁸ many teachers listed clear, detailed objectives for the lesson, described the planned sequence of activities, provided references to textbook pages or handouts, and indicated which artifacts from the notebook were related to the lesson.

Many of the daily reflections were directly germane to the dimensions of practice we were investigating. For example, the following comment from Teacher 10 was directly relevant to classroom discourse, “Students solved the problem and then shared strategies... I challenged them to talk with their table or neighbor to see if they could come up with one more way to solve the problem.” Similarly, this note from Teacher 20 provided information about student grouping “Some [students] were not working well with their groups – I know because they were physically removing themselves from the group/not talking to each other, etc.”

Teachers’ responses to the next three daily reflection questions were somewhat less revealing.⁹ The modal responses were, essentially, that the lesson unfolded as planned, that students learned what was expected, and that no changes were planned for the next day. Some responses showed that teachers had a clear sense of which concepts or procedures students had learned and which were causing problems, of how well different students had mastered them, and of how to address learning deficiencies in subsequent lessons. It was also common to read simply that

⁸ 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 (with respect to content, instructional strategies and student activities)?

⁹ Thinking back to your original plans for the lesson, were there any changes in how the lesson actually unfolded? How well were your objectives/expectations for student learning met in today’s lesson? Will today’s class session affect your plans for tomorrow (or later in the “unit”)?

students did not finish the activities or learn the material as rapidly as anticipated and, as a result, the teacher was planning to complete the lesson the next day, come back to it later, or leave it incomplete. Most teachers seemed to differentiate between topics that were important enough to demand review and topics that would be left incomplete if time ran out.

There were few useful answers to the sixth daily question;¹⁰ most teachers simply reported, “no,” they had nothing to add. In the future it might be possible to eliminate some of these latter questions with no major loss of information.

Looking at responses across days revealed that daily reflections often grew shorter as the Scoop period continued. This shrinkage may reflect the fact that many of the answers were unchanged from one day to the next, that the lessons themselves were similar or even repeated from day to day, or that teachers grew tired of responding to the questions.

Post Scoop Reflections. At the conclusion of the Scoop period, teachers were asked four post-Scoop reflection questions concerning the representativeness of the Scoop and ways to improve it, and 84 percent of the teachers answered them all.¹¹ The first question, which concerned the fit of the Scoop lesson with the teacher’s long-term goals for the class, did not prompt many responses that added to our understanding of classroom practices. The second question about the representativeness of the lessons included in the Scoop period yielded generally positive judgments. That is, the vast majority of teachers said that the lessons during the Scoop period were “very typical” or “very representative” of lessons during the year. This is important evidence regarding the validity of inferences drawn from the sample of lessons included in the notebook. A few teachers explained why they believed the notebooks were not representative, and these comments were quite germane, as well. For example, Teacher 40 reported, “Not typical. Mostly I try to be ‘hands on/minds on.’ The periodic chart and atoms are hard to do so.” This

¹⁰ Is there anything else you would like us to know about this lesson that you feel was not captured by the Scoop?

¹¹ 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 (with respect to content, instructional strategies and student activities)? What aspects were typical? How well does this collection of artifacts, photographs, and reflections capture what it is like to learn science in your classroom? If you were preparing this notebook to help someone understand your teaching, what else would you want the notebook to include?

comment suggests that judgments based on the notebook would not be typical of this teacher's instruction in some dimensions, particularly, Hands On.

Similarly, most teachers said the collection of artifacts, photographs and reflections in the notebook captured what it was like to learn mathematics or science in their classroom very well. The response of Teacher 85 was typical, "I think the project does a pretty good job at getting a snapshot of life [sic] of an 8th grade science student in my classroom." A few teachers commented that the artifacts were not enough to provide a true picture, indicating that the Scoop period should be extended or the notebooks should be supplemented with classroom observations. For example, Teacher 25 reported, "Artifacts and photos show the outline of instruction, but observation is the only way to see how student interact, understand and work together." Teacher 15 mentioned the lack of information on discourse, "What you don't see in the binder is any dialogue between students. That dialogue, between students and between teacher and students, is a pivotal part of my mathematics instruction. I am always asking student to explain their thinking..." This comment is consistent with the judgments of the research team regarding the evidence in the notebooks about discourse (see next section).

Teachers had many ideas regarding ways to supplement the notebooks to provide a more complete picture of instruction in their classroom. For example, Teacher 55 wrote, "Have students write their reflections about how they thought the lesson was and whether they thought the aim for the day was met and why [sic]." For both practical and ethical reasons we did not involve students in the data collection, although it is clear that they have an important perspective to offer on our defining question, "What is it like to learn mathematics/science in this classroom?" Other suggestions from teachers included including information about state standards, district curriculum, school mission or philosophy, additional examples of student work, pictures of the students,¹² information about classroom discipline, a follow-up interview after reviewing the materials, information about individual students' learning needs and abilities, and videotapes of the teaching. Some teachers also suggested a longer Scoop period, arguing that "it really takes a week and a half to go over a topic and make sure they really understand it" (Teacher 70).

¹² Because we did not consider students to be the objects of this research, we did not obtain consent from students and their parents, and we directed teachers not to take identifiable pictures of students.

Finally, a few teachers despaired of artifacts ever being able to capture their practices fully. For example, Teacher 35 reported

“I don’t come to school everyday to teach mathematics, I come to school everyday to teach people... young people... I learn with whom I must use gentle language and with whom I must be forceful. I watch their faces ever so intently to see if they understand. None of this can be put into a notebook.”

Similarly, Teacher 25 said, “I think it is crucial to understand the students and what works for them in order to see if instruction is effective.” Of course, we concur with both of these teachers. The Scoop project was designed to see whether artifacts could serve as rough surrogates for direct observations of classroom instruction, but we acknowledged at the outset that we did not expect them to provide the same level of detail as observations. Another teacher suggested that the notebooks did not reflect “the organizational and support structures that I use (Teacher 45),” including “grade record sheets, classroom guidelines and expectations, homework help brochures, and computers, that help students find and manage information...” While most of these materials could have been included in the notebook, it may be that this teacher did not include them because they were not generated during the Scoop week. Her comments suggest that it might be useful to broaden the list of artifacts to include materials that would help raters understand the “organization and support structures” in a teacher’s classroom.

Usefulness of Scoop Contents for Rating Specific Dimensions of Practice

As part of the evaluation of the notebook procedures, the research team members rated the usefulness of the 11 sources of information contained in the notebooks for making judgments about each of the 10 dimensions of practice. Figure 7 provides a visual summary of the ratings.

| | | | | | | | | | | | |
|-------------------------------------|----------------------|------------------|-----------------------|----------|-----------|--------|----------------------|---------------------|-------------------------------|--------------|-----------------------------|
| 1. Grouping | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 2. Structure of Lessons | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 3. Scientific Resources | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 4. Hands-On | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 5. Inquiry | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 6. Cognitive Depth | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 7. Discourse Community | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 8. Explanation/Justification | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 9. Assessment | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |
| 10. Connections/Applications | Pre-Scoop Reflection | Daily Reflection | Post-Scoop Reflection | Calendar | Photo Log | Photos | Yellow Sticky Labels | White Sticky Labels | Daily Instructional Materials | Student Work | Formal Classroom Assessment |

| | |
|-------------------------|---------------------------|
| Very helpful | Mean ≥ 1.5 |
| Somewhat helpful | $0.5 < \text{Mean} < 1.4$ |
| Not helpful | Mean ≤ 0.4 |

Figure 7. Average usefulness of artifacts for rating each dimension.

The daily reflections received the highest average usefulness ratings on all the dimensions, indicating that readers found these teacher responses to be very revealing on all features of instructional practice. Photographs were perceived to be highly useful for three dimensions (grouping, scientific resources, and hands-on) reflecting the importance of visual elements in gaining an accurate sense of the extent to which these features of instruction were present in the classroom. Daily instructional materials received high usefulness ratings on four dimensions, and student work on three. In contrast, the photo log was judged to be of limited value on most dimensions. The white sticky labels that accompanied assignments and classroom materials and the formal classroom assessments were also rated of limited value in most cases. However, both the white labels and the assessments were rated as highly useful on the Assessment dimension.

The pattern of ratings of usefulness reveals two broad findings. First, the teacher reflections were judged to be as revealing or more revealing of the dimensions of reform-oriented practice than any of the naturally occurring artifacts in the notebook. In fact, teachers' reactions to each day's lesson were perceived as more helpful in rating the dimensions than any of the instructional artifacts. This suggests that scooping "pure" artifacts of practice might not have produced as accurate a depiction of these dimensions of practice as we anticipated when we initiated this project. Second, the usefulness of each artifact varied across dimensions. That is, the contents of the notebook that were most (and least) revealing with respect to Structure of Lessons were different than the contents that were most (and least) revealing about Hands-On Activities. If we were to limit our interest in describing practice to a subset of the dimensions, we might be able to reduce the types of information we collected without any loss of clarity. The converse would also be true, eliminating certain artifacts from the Scoop would reduce our ability to describe some dimensions, but not others. Unfortunately, no single source of information was so uninformative and no single dimension was so indescribable that it would be an obvious choice for elimination.

Effectiveness of the Scoop Scoring Procedures

The section "Reliability and Validity of Scoop Ratings" contained evidence about the technical quality of the scores assigned to the Scoop Notebooks and to classroom observations. Here we present information about the effectiveness of the scoring process—including the definition of the dimensions of reform-oriented

practice and the application of the scoring guides. The findings come from discussions among members of the research team over the course of the study (including the pilot studies and the field studies) as well as from results reported under the earlier section entitled “Reliability and Validity of Scoop Ratings”. The analyses were designed to answer two questions:

- 1) How well did the scoring procedures function?
- 2) What suggestions did researchers offer for improving the scoring process?

The results are presented in two parts. First, we discuss insights gained during the development and revision of the scoring guides, rater training and calibration, and debriefing. Second, we present information obtained by comparing scoring activities across data sources—notebooks compared with observations, and mathematics scoring compared with science scoring. Many of the conclusions are presented in the form of recommendations.

Insights from the Development and Revision of Scoring Guides

Having multiple pilot and field tests allowed us to revise scoring materials and procedures based on experience. Most revisions occurred during the training/calibration meetings that were held before researchers observed in classrooms and again before they rated the Scoop Notebooks. In general, these revisions were made in response to questions and concerns that surfaced in discussions of disagreements among raters.

Define dimensions in terms of fewer features. Multiple features are necessary to fully characterize the complex constructs represented by the dimensions of instructional practice. However, including multiple features in the description of a dimension can lead to differences in the judgments made by different raters. To address this concern, we revised definitions to focus on fewer features when it was possible to do so without compromising the essential nature of a dimension. For example, the original scoring guide for the Grouping dimension took into account substantive features of the work done in groups, i.e., whether group work addressed non-trivial tasks, focused on conceptual aspects of the task, and was collaborative in nature. We revised the definition of Grouping to focus only on the extent to which groups were used during instruction without judging the activities done in groups. Similarly, the scoring guide for Cognitive Depth was simplified to focus on lesson design and teacher enactment but not student performance.

Reduce overlap among dimensions. It proved helpful to eliminate overlap among dimensions that was present in earlier versions of the scoring guide. For example, deep conceptual understanding of subject matter is a key characteristic of reform-oriented instruction that is related to several of the dimensions, including Cognitive Depth, Structure of Lesson, and Discourse Community. We reduced redundancy and facilitated scoring by eliminating conceptual understanding from the definition and the criteria for rating Structure of Lessons and Discourse Community.

Provide precise decision rules for determining performance levels. Inconsistencies occurred when raters placed emphasis on different components or features of a dimension, such as quality versus quantity. For example, consider Connections/Applications: In some classrooms there might be one high quality (or highly relevant) connection across a series of lessons; in others there might be several connections that are more superficial in nature. Raters who placed more emphasis on the quality of the connection and those who placed more emphasis on quantity or frequency would likely have differed in their judgments. In situations where a dimension was multi-faceted (e.g., the description involved both the frequency of an activity and some qualitative feature of the activity), we found it helpful to provide specific quantitative guidelines for assigning ratings. For example, for Connections/Applications, we asked raters to focus on frequency. In contrast, the Grouping dimension includes two aspects: the frequency of working in groups and the activities done by students in the group. After working with it in a less precise form, we rewrote the description to specify three possible combinations of frequency and activities that would be given a rating of 3: that the teacher frequently designs activities to be done in groups but many students work independently on the activities, that students occasionally work in groups doing activities that are directly related to the mathematical/scientific goals of the lesson, or that students regularly work in groups doing rote or routine activities. Similarly, we specified that for a rating of 3 on Use of Mathematical Tools, either “students are regularly encouraged to use mathematical tools ... with little or no explicit connection made between the representation and the mathematical ideas or they use tools occasionally to express important mathematical ideas.”

Identify relevant information sources. Rating inconsistencies may have occurred because raters relied on different sources of information within the Scoop Notebook. For example, two researchers might assign different ratings for Grouping,

depending on whether they rely primarily on photographs or teachers' reflections; or they might assign different ratings for Discourse Community depending on whether they attend mostly to written work or teacher reflections. To address this concern, we added specific suggestions about information sources in the notebook to draw upon when assigning ratings. For example, readers were informed that indirect evidence for Discourse Community might be found in teacher reflections such as "I had students compare their solution strategies with one another" or "I walked around and listened to students' conversations." Evidence might also come from lesson plans showing discussion of mathematical topics.

Ultimately, efforts to improve reliability by tightening the scoring rules had to be balanced with the need to avoid narrow rules that would not be responsive to the range and complexity of the Scoop Notebooks or the situations we encountered in the classroom. It is difficult to resolve this tension; our experience suggests it is possible to strive for descriptions that reduce differences in emphasis but impossible to completely eliminate such differences. Given that different sources of information may provide "better" evidence for some dimensions than others, our experience suggests that in constructing an artifact-based descriptive tool, one should start with aspects of practice, consider sources of evidence, and select those that balance evidentiary power with practicality.

Insights from Comparing Scores across Data Sources and Subjects

Comparing notebook scores with observation scores. As reported in the section entitled "Reliability and Validity of Scoop Ratings," correlations between ratings based on notebooks and those based on observations were moderate to high for most dimensions; however, agreement was notably lower for Assessment and Structure of Lessons. We investigated these dimensions further and found that differences in scores could be explained by differences in the information provided by the notebooks and the observations, i.e., the two data sources were not equally informative about these dimensions. For example, the Assessment dimension includes both formal and informal strategies, which are not equally easy to identify by observers in a classroom and readers of a notebook. It is easy to judge teachers' formal assessment strategies from the notebooks but not their informal assessment methods. Despite the fact that teachers are asked to use sticky notes to provide explicit reflections on student work, their comments typically are very brief. Thus, evidence for making judgments about informal assessment strategies is often very

limited. In contrast, evidence of informal assessment is readily apparent from teacher-student interactions in the classroom, yet, on any given day, there might not be a formal assessment activity such as a test or quiz.

We also asked raters about the ease with which they could rate each dimension. Their replies provided additional insights about dimensions that were easier to rate based on observations than Scoop Notebooks. For example, researchers observing in classrooms reported that they found it easy to assess whether students were actually working in groups (as opposed to just sitting in groups), and to determine the nature of the group tasks. On the other hand, determining a Grouping rating from the Scoop Notebook was more difficult; readers relied primarily on photographs and teacher reflections. Similarly, raters found Cognitive Depth easier to rate based on observations than Scoop Notebooks. We did not ask teachers to identify the concepts or “big ideas” of their lesson in the notebook, so raters had to rely primarily on assignments, worksheets, and teacher reflections to make judgments about Cognitive Depth, and these artifacts were not always clear. In these cases, it may be inappropriate to expect artifacts to be as revealing of practice as observations, and it may be necessary to change the Scoop process to be able to assess them adequately.

Comparing mathematics scoring with science scoring. We also found that some dimensions seemed to operate better in one subject area than the other although we thought they were appropriate to both. For example, ratings of Explanation/Justification were less reliable in science than mathematics. On further investigation, we found that there was considerable rater-by-classroom interaction in science. This interaction suggests that the specific characteristics of each science classroom may interact with raters’ understanding of the dimension to generate inconsistencies in the ratings. These inconsistencies may arise because some raters were less familiar with the content of the science lessons than others. They may also reflect the fact that explanation and justification are prominently featured in mathematics reform initiatives, but they are not an explicit component of the science reform standards (NCTM, 2000; NRC, 1996). In developing the mathematics education standards, the mathematics community has considered issues such as the nature of explanation, how to differentiate types of explanations, etc. that are not present in science standards. Perhaps the Explanation/Justification scoring guides were better defined in mathematics than in science, or raters were better able to draw upon their understanding of the professional community’s standards in mathematics than in science.

A similar situation occurred with the Connections/Applications dimension. Connections in mathematics are well defined in the mathematics reform documents in terms of “real-world” applications, and reviewers were consistent in noting when this occurred. Connections in science were “everywhere and nowhere,” and raters found it difficult to decide whether or not to give credit to naturally occurring connections in the lessons.

The problems with these two dimensions raise some questions about the wisdom of pushing for parallel dimensions in mathematics and science. The advantage, from the perspective of the research, was the ability to ask parallel questions about artifact collection in mathematics and science. We determined that the Scoop Notebook and scoring guides do function similarly for the two subject areas, but we identified several features that were more problematic in one subject area than the other. Therefore, for uses that do not require parallel measures of mathematics and science instruction, it would make sense to revise the scoring guide for each subject area to more closely align with descriptions of desired instructional practices set forth in the Standards documents for that discipline (NCTM, 2000; NRC, 1996).

The differences between the psychometric characteristics of mathematics and science ratings also highlighted the importance of the subject matter expertise of the readers. Some researchers on this project had stronger mathematics backgrounds while others’ had stronger backgrounds in science. For several dimensions, these differences in subject matter expertise were associated with differences in perceived difficulty of ratings and in rater confidence. The clearest example of this is Cognitive Depth. Raters reported that their own depth of understanding of the content affected their ability to judge the extent to which the lessons promoted command of central ideas in the discipline. As a result, some raters felt more confident of their ratings of mathematics lessons whereas some were more confident in science. We found that subject matter expertise was important both when developing artifact-based strategies for measuring instructional practices and when using these strategies. Our initial answer to the question “How much subject matter expertise is enough?” is that the level of expertise must be related to the complexity of the instructional practices being measured. In other words, the sophistication of the development and scoring processes should match the sophistication of the dimensions of practice. If we wanted to measure homework completion, for example, we might be able to construct a simpler process and use raters with lower levels of expertise.

Artifact-Based Measures of Classroom Practice

In this section, we recap what we learned about appropriate ways to use the Scoop Notebook in its present form, note additional questions about the Scoop raised by the field studies, and consider what the Scoop project suggests about the use of artifacts in research, in general, and about possible directions for future development and research.

Appropriate Uses of the Scoop Notebook

Aggregate description of Reform-Oriented Practice. We believe that the Scoop Notebook, as presently constituted, provides a reasonably accurate portrayal of selected dimensions of instructional practice. The reliability and validity are sufficient to support using notebook ratings in situations where aggregate descriptions are needed, such as for describing practices of groups of teachers, or as part of evaluations of instructional or curricular program reform efforts. For example, a Scoop Notebook might be useful for providing an indication of changes that occur over time as a result of adopting a new science or mathematics curriculum.

On the other hand, our analyses indicate that the reliability and validity of the ratings are currently not sufficient to justify the use of the Scoop Notebook for making high-stakes decisions about individual teachers. It might be possible, however, to use the notebook in combination with other measures of teacher background, attitudes, or content knowledge, as part of a system of indicators that could support valid inferences about individual classroom practice in high stakes situations. Moreover, the quality of the information collected by teachers in the notebook (and thus conceivably the technical quality of the ratings of this information) could improve if it were part of a process endorsed by teachers, or in situations where teachers are personally invested in providing the most complete and detailed information possible about their classroom practices. While such systems might provide sufficiently more robust evidence than the materials and procedures we field tested, further research would clearly be needed to validate using the Scoop Notebook in combination with other data sources for high stakes purposes.

Professional development. We also believe that the Scoop Notebook can be valuable as a learning tool for teachers, particularly if it is incorporated into a

professional development program directed at helping teachers understand their own practice. In the context of such a program, teachers may find the Scoop Notebook helpful for describing and reflecting on their own instructional practices, and for collaborating with colleagues and researchers to better understand the nature of mathematics and science learning and teaching in their classrooms. As one example, it could help teachers trace changes in their instructional practices over time or across instructional units. Similarly, it could help teachers to think systematically about the ways in which different aspects of reform instruction might be more or less suitable for lessons with different objectives or content foci.

Importantly, when the Scoop Notebook is used as a tool within a professional development program, the reflection questions can be modified to better fit the specific learning goals of that program. The teachers can use these questions to guide their personal reflections on the learning goals in relation to their own practice (as they collect classroom artifacts and assemble the Scoop Notebook), as well as their discussions with colleagues within the professional development program.

Unanswered Question About the Scoop Notebook

The study raised some questions about the Scoop Notebook and features of the rating guidelines and procedures that require further investigation. These include questions about the dimensions used to characterize classroom practice, time sampling, data collection procedures, and costs.

Dimensions of Reform-Oriented Practice. As noted earlier, factor analyses of notebook and observation ratings point to considerable overlapping variance among the dimensions. Although each dimension has distinctly unique characteristics that differentiate it from the others, the dimensions are also closely related conceptually. Thus, it could be possible to characterize instructional practices in mathematics and science reliably and validly, using fewer dimensions, to provide an overall indication of the extent to which a classroom resembles the model of reform-oriented instruction that underlies the Scoop Notebook—indeed, the Overall Reform dimension in our Mathematics and Science studies consistently showed good levels of reliability and validity. Moreover, simplifying the process of rating Scoop Notebooks by using fewer, less overlapping dimensions, could facilitate rating consistency and result in improved reliability. At the same time, nuances of classroom practice reflected in the unique aspects of the dimensions may not be adequately captured by aggregate indices.

We leave open the question of what is the best way to characterize reform-oriented instructional practice. The answer depends in part on the purpose and the kinds of inferences sought. For example, some researchers or practitioners may not be interested in all the dimensions that currently comprise the construct in our rating system and may choose to focus on aggregate indices or a reduced set of variables of interest. In general, for evaluative purposes (or others where the reliability of broad measures of practice is of central interest) we would caution against using large numbers of related dimensions in favor of approaches that involve fewer, more conceptually distinct dimensions. On the other hand, in some situations aggregate indices may not be useful or even desirable. For example, if interest centers on professional development or on providing feedback about classroom practice to individual teachers in a way that prompts thoughtful comparison and conversation, retaining dimensions that reflect specific and meaningful aspects of instruction would be desirable, even if they were highly correlated with each other.

Time sampling. Teacher practice in the classroom may vary from day to day based on curriculum, progress within a lesson, and features of the classroom environment like student discipline, teacher mood, or school-wide activities. Another set of issues that remains to be resolved relates to the number and timing of occasions of measurement (e.g., observations, scoop collection) needed to provide a stable portrayal of instruction in a classroom. Research suggests that observation occasion (i.e., time sampling) is an important source of error when measuring instruction (e.g., Shavelson, Ruiz-Primo, & Wiley, 1999; Shavelson, Webb, & Burstein, 1986), and the evidence from our study indeed points to important variability in terms of teacher practices in the classroom from one day to another. We found that for most dimensions five days of data collection can provide enough information to capture a unit of instruction with adequate levels of reliability and validity. While fewer days might be acceptable for other purposes, (e.g., for describing homework procedures) we suspect that most interesting aspects of practice will not be reliably revealed in less than a week of instruction or its equivalent.

Importantly, the extent to which the evidence presented here may support generalizing from a single unit of instruction to draw implications about classroom practice in general is uncertain. We observed classrooms at different times during fall or spring, so there is some assurance that our data captures variation in

instructional practices throughout the school year in the aggregate. However, because we observed a single unit of instruction in each classroom we were not able to investigate the extent to which there may be systematic fluctuations in individual teachers' instructional practices between units, or at different time points during the year. Without additional research we do not know how many days of Scooping would be needed to produce a comprehensive and accurate description encompassing an individual teacher's practice over the course of an entire school year.

Other factors to consider in time sampling decisions include the range of activities that occur in mathematics and science classrooms, and the extent to which ratings may be sensitive to the particular activities, lessons, or units observed. For example, researchers remarked that some of the mathematics units observed seemed to lend themselves more easily than others to the use of multiple representations; similarly, some science units were particularly well-suited to the incorporation of hands-on activities. Feedback from teachers and researchers suggests that variability due to instructional activities and units may be more of a concern in science than mathematics, and for some dimensions of instructional practice than others.

Interestingly, most teachers in our field studies indicated in their post-Scoop reflections that the materials in their notebooks were representative of what it was like to learn mathematics or science in their classrooms. Nevertheless, additional research would be needed in exploring questions such as how many days of instruction to include in a Scoop Notebook, how frequently and when during the school year these materials should be collected, and whether a sampling plan should take unit content into account.

Other aspects of practice. We believe the Scoop Notebook could be modified to measure many other aspects of practice, such as classroom management, press for achievement, student engagement, and motivation. Additional research is needed, however, to determine whether this is indeed the case, what artifacts and scoring dimensions should be used to characterize these features of classroom life, and what features might not be characterized well by any artifacts. It is also worth exploring what modifications would be necessary to adapt the Scoop Notebook and scoring guidelines for use in studying instruction in other subject areas or grade levels. More would need to be done, however, to study which artifacts would be revealing of which features, and which features might not be evidenced well by any artifacts.

Procedures and costs. We realize that the efficacy of the notebook will depend on its cost as well as the reliability and validity of the information it yields. Thus, it would be useful to have systematic data regarding various types of costs, as well as to explore potential ways to reduce these costs—for example, by streamlining the processes of data collection—without negatively affecting the reliability and validity of judgments. Similarly, approaches to improving reliability and validity without increasing costs would be of interest. In other words, it would be worthwhile to investigate how much could be learned with less intensive data collection methods, and how much more could be obtained with the same amount of resources.

There are a number of considerations and possibilities that might help in streamlining the process of data collection and making it more effective. One suggestion that warrants investigation for improving the quality of the information collected through the Scoop Notebook is the provision of additional training for teachers. Prior to data collection, we met with participating teachers individually or in small groups for approximately one-half hour to review the notebook and instructions. Researchers might ask teachers to complete notebook materials for one day and then review their materials and provide feedback. Such additional guidance and support might encourage teachers to provide photographs and logs that better depict classroom dynamics and activities, more complete selections of assignments and student work, and more detailed reflections. However, it might not be practical for most large-scale research studies.

More generally, it would be useful to have systematic information regarding the costs of the Scoop Notebook in terms of time, resources, and money. For example, additional research would be necessary to investigate the extent to which increasing training time is a cost-effective allocation of resources in large-scale data collection efforts.

Similarly, based on our informal conversations with field study teachers we know that completion of the Scoop Notebook imposed a moderate burden on them: We estimate that they spent an average of about 10 hours compiling materials and completing reflections. However, we do not have systematic data on burden to respondents of the notebook in general and of individual parts of the notebook specifically.

Artifacts in General

Our experience with the Scoop Notebook also suggests some general lessons about the use of artifacts for studying classroom practices. Artifacts capture many important aspects of instruction, and they can help outsiders understand what occurs in classes they cannot attend. Perhaps their most valuable attribute is that they reflect actual instructional activities rather than teachers' interpretations of those activities. But everything that is produced is not equally revealing, and different artifacts are more or less revealing about different features of classroom life. In this section, we explore some of the lessons we learned with respect to these issues.

Artifacts Can Be Collected Systematically "At a Distance"

This study demonstrates that it is possible to collect comparable artifacts from multiple teachers, and it is possible to do this even at some distance. The ability to meet with teachers in person and explain the procedures was beneficial, but we believe the notebook process could have been successful with group instructions or telephone instructions, as well. More personal contact may be necessary the further the data collection process strays from activities that are familiar.

Some Aspects of Practice Are Not Revealed in Artifacts

Everything that happens in a classroom is not necessarily reflected in an artifact, and artifacts do not magically reveal hidden secrets of practice. With respect to the dimensions of classroom practice that we investigated, artifacts were more informative about structural features such as use of mathematical tools or scientific resources, and less informative about interactive aspects of instruction such as patterns of discourse and the nature of explanations. Moreover, within individual dimensions, artifacts may be more revealing of certain aspects of practice (e.g., formal assessment tools) than others (e.g., informal assessment strategies).

Recognizing that artifacts alone might not be sufficient for our purposes, we also asked teachers to generate additional materials for us, including daily calendars and reflections. Interestingly, despite our concerns about the limitations of teacher reflections including their susceptibility to social desirability, raters consistently judged the reflections to be as informative or more informative about the dimensions than other components of the Scoop Notebook. This finding should not be interpreted to suggest that teacher reflections could replace the other components

of the notebook. Quite the contrary, we found that the notebooks provided a richer and more complete portrayal of classroom practice to the extent that the reflections and artifacts reinforced and complemented each other. For example, while photographs of teams of students conducting a science experiment may provide useful information about the extent of Grouping and Hands-On practices in a classroom, teacher reflections describing the learning goals of the experiment and the way the lesson developed would provide valuable context to improve our interpretation of the photographs. Thus, in combination the information from the pictures and the reflections can provide better information about many of the dimensions (not just Grouping and Hands-On) than either source alone.

Choice of Artifacts Should Be Based on Descriptive Purposes

There is no *a priori* best artifact or set of artifacts for every purpose. The design and choice of artifacts should be driven by the aspects of classroom practice of interest, and the goals of the investigation. Certain artifacts revealed the most about one of our dimensions and other artifacts revealed the most about another. For example, for our purposes photographs were very revealing. We found that they showed a lot about classroom arrangements, displays, educational resources, activities-based lessons, and student participation. They are also relatively easy and inexpensive to include in an artifact collection using disposable cameras.¹³ We also found assignments to be very revealing, providing information related to a number of dimensions including Cognitive Depth, Explanation/Justification, and Assessment. However, the sticky notes teachers completed to provide comments on these assignments were less helpful than we anticipated. Graded homework assignments and completed formal assessments could be a valuable addition to the set of artifacts in the notebook.

Teacher reflections are a valuable source of information about a variety of instruction aspects including classroom routines, teacher expectations, and instructional improvement, but they are most useful when tied to specific queries. Journal entries in response to general prompts may not be as useful for describing specific elements of practice as are responses to more targeted questions. One shortcoming of reflections, like surveys, is their self-reported nature. There may be a

¹³ Researchers who include photographs in artifact collections must be careful to protect identities of students. Despite our instructions, several teachers took pictures that showed student faces. We had to black them out to protect students' identities.

component of social desirability in teachers' reflections, particularly when they are aware of the researchers' views about the features of instructional practice being studied. In addition, judgments based on reflections may be biased by teachers' verbal ability, expressiveness, or thoroughness. In other words, while some teachers write more and therefore reveal more, this does not necessarily imply that they do more in the classroom or that their teaching is better.

While other researchers can learn from our experiences with the types of artifacts and materials contained in the Scoop Notebook, they may also find it necessary to conduct additional research to discover how much is revealed by one method, one set of dimensions, or one set of artifacts, compared with another.

In conclusion, perhaps the most important insight we can offer regarding artifact-based tools, such as the Scoop Notebook, is that when used judiciously as part of a data collection system, they can offer valuable information about many aspects of reform-oriented practice. We set out to develop and field test a system for measuring instructional practices that represents a hybrid of several methods, in an attempt to maximize the strengths of these methods while minimizing their shortcomings. We believe that we achieved this goal with the Scoop Notebook. It capitalizes on the strengths of both artifacts and self-reports; as is the goal with most hybrids, it has advantages over the separate sources from which it was derived.

References

- Antil, L. R., Jenkins, J. R., Wayne, S. K., and Vasdasy, P. F. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. *American Educational Research Journal*, 35, 419-454.
- Ball, D. L., and Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal*, 105(1), 3-10.
- Blank, R. K., Porter, A., and Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science: Results from survey of enacted curriculum project*. Final report published under a grant from National Science Foundation/HER/REC. Washington, DC: CCSSO.
- Borko, H. and 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., Stecher, B. M., Alonzo, A., Moncure, S., and McClam, S. (2003). *Artifact packages for measuring instructional practice*. (CSE Technical Report No. 615). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Borko, H., Stecher, B. M., Alonzo, A., Moncure, S., and McClam, S. (2005). Artifact packages for measuring instructional practice: A pilot study. *Educational Assessment*, 10(2), 73-104.
- Borko, H., Stecher, B. M., Martinez, F., Kuffner, K. L., Barnes, D., Arnold, S. C., Spencer, J., Creighton, L., and Gilbert, M. L. (2006). *Using classroom artifacts to measure instructional practices in middle school science: A two-state field test*. (CSE Report No. 690). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., and Guitton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.
- Camburn, E., and Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal*, 105(1), 49-73.
- Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice*. (CSE Technical Report No. 532). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards and Student Testing (CRESST).

- Clare L. and 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., and Steinberg, J. R. (2001). *Teachers' assignments as Indicators of instructional quality in elementary schools* (CSE Technical Report No. 545). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Firestone, W. A., Mayrowetz, D., and 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.
- Fullan, M. G. and Miles, M. B. (1992). Getting reform right: What works and what doesn't. *Phi Delta Kappan*, 73, 745-752.
- Hill, H.C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, 19(3), 447-475.
- 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.
- Laguarda, K. G. (1998). *Assessing the SSIs' impacts on student achievement: An imperfect science*. Menlo Park, CA: SRI International.
- Le, V, Stecher, B. M., Lockwood, J. R., Hamilton, L. S., Robyn, A., Williams, V. L., Ryan, G., Kerr, K. A., Martinez, J. F., and Klein, S. P. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement*. Santa Monica, CA: RAND Corporation.
- Learning Mathematics for Teaching (2006). *A coding rubric for measuring the quality of mathematics in instruction* (Technical Report LMT1.06). Ann Arbor, MI: University of Michigan, School of Education.
- 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., and Valdes, R. (2002). *Measuring instructional quality in accountability systems: Classroom assignments and student achievement* (CSE Technical Report No. 582). Los Angeles, CA: University of

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

Matsumura, L. C., Slater, S. C, Wolf, M., Crosson, A., Levison, A., Peterson, M., and Resnick, L. (2005). *Using the Instructional Quality Assessment Toolkit to investigate the quality of reading comprehension assignments and student work* (CSE Technical Report No. 669). Los Angeles: University of California, 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.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S.P., Bugliari, D., and Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education*, 32(5), 493-517.

McDonnell, L. M., and Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Technical Report No. 442). Los Angeles, CA: 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., and Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws and T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 96-113). Hillsdale, NJ: Erlbaum.

Resnick, L., Matsumura, L.C., and Junker, B. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the instructional quality assessment* (CSE Technical Report No. 681). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Rowan, B., Camburn, E., and Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: a study of literacy teaching in third-grade classrooms. *The Elementary School Journal*, 105, 75-102.

- Rowan, B., Harrison, D. M., and Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal*, 105, 103-128.
- Ruiz-Primo, M. A., Li, M., and Shavelson, R. J. (2002). *Looking into students' science notebooks: What do teachers do with them?* (CSE Technical Report No. 562). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- SAS Institute, Inc. (2002-2003). *SAS 9.1 Documentation*. Cary, NC: SAS Institute, Inc.
- Smithson, J. L., and 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., and 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., and Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (CSE Technical Report No. 525). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Stecher, B., and Borko, H. (2002). Integrating findings from surveys and case studies: examples from a study of standards-based educational reform. *Journal of Education Policy*, 17, 547-569.
- Stecher, B. M., Borko, H., Kuffner, K. L., Wood, A. C., Arnold, S. C., Gilbert, M. L., and Dorman E. H. (2005). *Using classroom artifacts to measure instructional practices in middle school mathematics: A two-state field test* (CSE Technical Report No. 662). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Stecher, B. M., 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.

- Stecher, B. M., Le, V., Hamilton, L. S., Ryan, G., Robyn, A., and Lockwood, J. R. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis*, 28(2), 101-130.
- Wolf, S. A. and 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

Mathematics SCOOP Rating Guide CRESST Artifact Project

The Rating Guide is designed to be used for three types of ratings: observation, Scoop Notebook, and Gold Standard. In general, the language applies equally well in all cases. However, in some cases, we modified language or added examples to reflect differences between ratings of one type or another. [Comments that apply just to observation, notebooks or gold standard ratings are noted in brackets.]

The Rating Guide consists of three parts: a quick reference guide; a description of all the rating levels with examples; and a reporting form for recording ratings and justifications/evidence.

For all dimensions (unless otherwise specified)...

- *Rate each dimension based on the highest level you found in the notebook or observed during the lesson. (Guiding principle: “When in doubt, be nice.” i.e., give the higher of the 2 ratings.)*
- *The rating should take into account teacher, students, and materials that are used.*
- *Remember, a rating of “5” does not mean perfection; it means that the lesson or series of lessons meets the description of a 5.*
- *One characteristic (limitation) of the Scoop rating scale is that there are many different ways a classroom can be a “medium” on each dimension.*
- *A rating of “medium” may be based on the frequency of multiple features of a dimension (e.g. assessment) and/or different levels of enactment by teachers and students (e.g. explanation/justification). In particular:*
 - *frequent occurrence of some features and limited occurrence of others*
 - *medium occurrence of all features*
 - *medium levels of enactment by both teacher and students*
 - *high level of enactment by one and low level by the other*

Specific Notes for Observation Ratings:

1. Take notes during the observation of each lesson.
2. Complete an observation rating each day and then a “summary rating” at the end of the series of observations (after all days of observation).
3. The summary rating completed at the end of the series of observations is a holistic rating (rather than mathematical average).
4. It is sometimes difficult to rate a dimension based on the observation of one lesson, especially when the dimension description includes a “series of lessons.”

Specific Notes for Gold Standard Ratings:

1. Use your rating forms and notes from the class observations, as well as the Scoop Notebook, in order to determine the rating for each dimension. In other words, use “everything that you know” to determine the gold standard ratings.
2. When explaining your rating of each dimension, describe the evidence you used from both your observations (overall rating) and the notebook (student work, reflections, pictures, lesson plan, etc.) to determine your rating.

3. Two dimensions are included in the Scoop Notebook and gold standard ratings (but not the observation ratings): notebook completeness and confidence.

Quick Reference Guide for CRESST Mathematics SCOOP Ratings

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on mathematical tasks that are directly related to the mathematical goals of the lesson and to enable students to work together to complete these activities. Active teacher role in facilitating groups is not necessary.

[The focus for observation of a single lesson is on the nature of the activities, and how integral the group activity is to the substance of the lesson (and not necessarily the amount of time spent in groups.)]

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related mathematically and build on one another in a logical manner.

NOTE: The focus of this dimension is on design, rather than enactment.

[Ratings of observations should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.]

3. 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. The extent to which students select, use, and translate among (go back and forth between) mathematical representations in an appropriate manner.

NOTE: dimension includes both exposure (by teacher or curriculum) and use by students.

4. Use of Mathematical Tools. The extent to which the series of lessons affords students the opportunity to use appropriate mathematical tools (e.g., calculators, compasses, protractors, Algebra Tiles), and that these tools enable them to represent abstract mathematical ideas.

NOTE: When students use equipment and/or objects to collect data that are later used in exploring mathematical ideas, the equipment/objects are not considered to be mathematical tools unless they are also explicitly used to develop the mathematical ideas.

5. Cognitive Depth. Cognitive depth refers to command of the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and connections and relationships among mathematics concepts. This dimension considers two aspects of cognitive depth: lesson design and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently and effectively promotes cognitive depth.

6. Mathematical Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their mathematical ideas honestly and openly. The extent to which the teacher and students “talk mathematics,” and students are expected to communicate their mathematical thinking clearly to their peers and teacher, both orally and in writing, using the language of mathematics.

NOTE: There is a “high bar” on this dimension because there is an expectation for students to have an active role in promoting discourse; this should not be only the teacher’s role. This is in contrast to

Explanation/Justification. The rating does take into account whether discourse focuses on mathematics content but not the cognitive depth of that content.

7. Explanation and Justification. The extent to which the teacher expects and students provide explanations/justifications, both orally and on written assignments.

NOTE: Simply “showing your work” on written assignments – i.e., writing the steps involved in calculating an answer – does not constitute an explanation. This is different from “cognitive depth” because it is not dependent on “big ideas” in the discipline.

8. 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.

NOTE: This dimension focuses more on the nature of the activity/task than the enactment. To receive a high rating, problems should not be routine or algorithmic; they should consistently require novel, challenging, and/or creative thinking.

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important mathematical ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making).

[Often the observer will need to make inferences about how the teacher is using the information, or plans to use the information, especially on a daily observation.]

10. 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. The extent to which series of lessons helps students apply mathematics to real world contexts and to problems in other disciplines.

NOTE: The experiences may be teacher-generated or student-generated, but they should relate to the students’ actual life situations.

11. Overall. How well the series of lessons reflects a model of instruction consistent with dimensions previously described. This dimension takes into account both the curriculum and the instructional practices.

NOTE: The rating on this dimension is implicitly a weighted average of the ratings on the first ten dimensions, with greater weight being given to Problem Solving, Cognitive depth, Mathematics Discourse Community, and Explanation/Justification.

[For Notebook and Gold Standard Ratings add:

12. Completeness. The extent to which the notebook contains all the materials we asked teachers to assemble.

13. Confidence. The degree of confidence you have in your ratings of the notebook across all dimensions.]

Description of CRESST Mathematics SCOOP Rating Levels

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on mathematical tasks that are directly related to the mathematical goals of the lesson and to enable students to work together to complete these activities. Active teacher role in facilitating groups is not necessary.

[The focus for observation of a single lesson is on the nature of the activities, and how integral the group activity is to the substance of the lesson (and not necessarily the amount of time spent in groups).]

High: Students consistently work in groups doing activities that are directly related to the mathematical goals of the lesson.

Example: Students do some work in groups on most days that information is scooped. For example, on one day 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 discusses and 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?"

Example: Lessons are organized so that students consistently work on their homework assignments in groups during class time, and all or almost all students are active participants during these sessions.

Medium: Either students occasionally work in groups doing activities that are directly related to the mathematical goals of the lesson OR students regularly work in groups doing rote or routine activities (e.g. checking each other's homework for accuracy and completeness, doing flashcards).

Example: Students do some work in groups on some days during which information is scooped. For example, on one day, 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.

Example: Students do some group work each day, primarily to review materials learned during whole class instruction. For example, on one day, they work in pairs using flash cards to test each other on equivalencies between fractions and percents.

Low: Students do not work in groups OR group activities do not involve mathematics.

Example: Students work in pairs to check each other's math notebooks for completion before turning them in.

Example: Instruction occurs only in whole-class or individual settings.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related mathematically and build on one another in a logical manner.

NOTE: The focus of this dimension is on design, rather than enactment.

[Ratings of observations should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.]

High: The series of lessons is conceptually coherent throughout; activities are related mathematically and build on one another in a logical manner.

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. The teacher assigns the students to bring in newspaper examples of fractions and decimals for the next day's lesson.

Medium: The series of lessons is conceptually coherent to some extent, but some activities appear to not be related to one another, OR some activities do not appear to follow in a logical order.

Example: In a unit on fractions, instruction begins with a discussion on where students have seen fractions before in their everyday lives. Next, the teacher demonstrates how to add fractions and discusses why a common denominator is needed. Before continuing with fractions, the teacher hands back a test on area and perimeter taken the previous week and answers questions about it. Then the teacher presents students with a recipe. Students are instructed to read the recipe, which includes several fractions (i.e. $\frac{1}{3}$ cup of sugar), and answer questions about the quantities involved. 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 series of lessons does not appear to be logically organized and mathematically connected.

Example: One day there is a lesson on graphing linear equations. The next day, the teacher hands back homework from yesterday that involved area and perimeter, and teacher answers questions about it. On the third day, students work on a worksheet introducing graphing calculators. They then practice graphing linear equations. On the final day of the scoop, they work on an unrelated Problem of the Day, in preparation for the state assessment.

3. 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. The extent to which students select, use, and translate among (go back and forth between) mathematical representations in an appropriate manner.

NOTE: dimension includes both exposure (by teacher or curriculum) and use by students.

High: Students are regularly exposed to quantitative information in a variety of forms. Students use multiple representations to present data and relationships, select representations appropriate to the tasks, and translate among them.

Example: In a lesson 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, and to indicate which representations are most appropriate for answering different questions about the mathematical functions.

Medium: Use of multiple representations has some but not all of the features mentioned above. Students are sometimes exposed to quantitative information in a variety of forms. Students sometimes use multiple representations, select representations appropriate to the tasks, or translate among them.

Example: In a lesson 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.

Example: In a lesson on patterns and functions, the teacher presents sequences in a variety of formats, including numerical lists, geometric patterns, tables, and graphs. However, student work focuses exclusively on translating between tables and graphs, or between functions and graphs. The problems indicate which representations to use.

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 lesson on patterns and functions, the teacher presents numerical sequences and asks students to write functions that describe the sequence.

4. Use of Mathematical Tools. The extent to which the series of lessons affords students the opportunity to use appropriate mathematical tools (e.g., calculators, compasses, protractors, Algebra Tiles), and that these tools enable them to represent abstract mathematical ideas.

NOTE: When students use equipment and/or objects to collect data that are later used in exploring mathematical ideas, the equipment/objects are not considered to be mathematical tools unless they are also explicitly used to develop the mathematical ideas.

High: Students' use of mathematical tools (including manipulatives) forms a regular and integral part of instruction throughout the series of lessons. The students are encouraged to use these tools in ways that express important mathematical ideas and to discuss the relationships between the tools and these ideas.

Example: Students are engaged in the task of modeling a long-division problem. Different types of tools 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 that other groups represent with manipulatives of their choice, explaining that choice.

Example: In a unit on linear equations, students use a graphing calculator to graph linear equations. Next, they use graphing calculator to explore how the slope and intercept of a the line are affected by changing the coefficient of the x-term and the constant term in the point-slope form of the equation.

They then use tables of values to identify coordinates of points for graphing equations on graph paper, and they compare the understandings fostered by these two approaches to graphing.

Medium: Students are encouraged to use mathematical tools (including manipulatives) to solve problems with little or no explicit connection made between the representation and mathematical ideas OR they use tools occasionally to express important mathematical ideas.

Example: Students are asked to solve a long division problem. Students are encouraged to 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 and associated manipulatives over another (or chose not to use manipulatives) because “it’s faster,” but the mathematical concepts behind these choices are left undeveloped.

Example: In a unit on linear equations, students use tables of values to identify coordinates of points for graphing equations on graph paper. They graph a set of lines to explore how the slope and intercept of the lines are affected by changing the coefficient of the x-term and the constant term in the equations. Then they confirm the accuracy of the graph using a graphing calculator

Low: Students are permitted to use mathematical tools (including manipulatives) if they are having difficulty with a mathematics procedure, or the only tools that are used are the standard paper, textbooks, and chalkboard.

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.

5. Cognitive Depth. Cognitive depth refers to command of the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and connections and relationships among mathematics concepts. This dimension considers two aspects of cognitive depth: lesson design and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently and effectively promotes cognitive depth.

High: Lessons focus on central concepts or “big ideas” and promote generalization from specific instances to larger concepts or relationships. The teacher consistently promotes student conceptual understanding. The teacher regularly attempts to engage students in discussions or activities that address central mathematical ideas and principles.

Example: The teacher designs a series of lessons in which 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. The students volunteer their conjectures, which are written on the board. Then the teacher facilitates a discussion about the students’ conjectures and linear relationships between two variables.

Medium: The series of lessons has some but not all of the features mentioned above. Lessons may focus on mastery of isolated concepts, but not on connections among them (e.g., they may require students to explain or describe the concept but not to use it or apply it), OR, the teacher sometimes attempts to engage students in discussions about

connections between mathematical concepts and sometimes demonstrate these connections, but not consistently.

Example: Students are asked to represent the above word problem in an equation. The students then are asked 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: Lesson 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. The teacher rarely attempts to engage students in instructional activities that demonstrate the connectedness of mathematical concepts and principles. The teacher's interactions with students focus on correctness of their answers rather than on conceptual understanding.

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.

6. Mathematical Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their mathematical ideas honestly and openly. The extent to which the teacher and students “talk mathematics,” and students are expected to communicate their mathematical thinking clearly to their peers and teacher, both orally and in writing, using the language of mathematics.

NOTE: There is a “high bar” on this dimension, because there is an expectation for students to take an active role in promoting discourse; this should not be only the teacher's role. This is in contrast to Explanation/Justification. The rating does take into account whether discourse focuses on mathematics content but not the cognitive depth of that content.

[The kind of indirect evidence we might find in the notebook includes:

- teacher reflections, such as:
 - I had students compare their solution strategies with one another;
 - I consciously try to get students to voice their ideas;
 - I walked around and listened to students' conversations
 - I encourage students to ask each other questions when they present their solutions to the class.
- peer reflections on student written work
- lesson plans showing discussion of mathematical topics]

High: Students consistently are encouraged to express their mathematical thinking to other students and the teacher, and they are supported by the teacher and other students in their efforts to do so. Students' ideas are solicited, explored, and attended to throughout the lesson, and students consistently use appropriate mathematical language. Emphasis is placed on making mathematical reasoning public, raising questions and challenging ideas presented by classmates.

Example: Students are using reallocation 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 reallocation 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. These discourse patterns appear to be the norm.

Medium: Students are expected to communicate about mathematics in the classroom, but communication is typically teacher-initiated e.g., the teacher attempts to foster student-to-student communication but students don't communicate with each other without teacher mediation). The use of appropriate mathematical language may or may not be consistent.

Example: Students are using reallocation 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 reallocate 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 reallocated 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. The students are typically giving answers to the teacher's questions, rather than engaging in student-to-student communication. The 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. Those discussions that occur are typically characterized by IRE (initiation, response, evaluation) or "guess-what's-in-my-head" discourse patterns. Students rarely use appropriate mathematical language. Student-to-student communication, when it occurs, is typically procedural and not about mathematical thinking.

Example: The teacher works on the overhead projector to show students how to use reallocation to find "fair prices" for pieces of floor tile in different sizes and shapes. As she works, she calls on students to suggest reallocation 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.

7. Explanation and Justification. The extent to which the teacher expects and students provide explanations/justifications, both orally and on written assignments.

NOTE: Simply "showing your work" on written assignments – i.e., writing the steps involved in calculating an answer – does not constitute an explanation. This is different from "cognitive depth" because it is not dependent on "big ideas" in the discipline.

High: Teacher consistently expects students to explain their mathematical thinking and problem solving strategies, both orally and on written assignments. Students' explanations show their understanding of generalized principles or previously proved conjectures, rather than examples or an appeal to authority. NOTE: We need to see evidence not only of teacher expectations, but also of a variety of students giving explanations and justifications.

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 $2=2x$. She pointed out that 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: In a whole class discussion, one student justifies that $a \times b \div b = a$ is always true by saying that when you divide a number by itself, you get one. So it's really like multiplying by one, and any number times one gets you the original number. Several other students also share the explanations they had written on their papers.

Medium: Teacher sometimes expects students to explain their mathematical thinking and problem solving strategies. Students sometimes provide explanations/justifications that include mathematical ideas. Students' explanations are usually procedural rather than conceptual. Or, teacher consistently expects students to explain their mathematical thinking, but students often do not provide such explanations.

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 $1=x$. Although the teacher consistently asks students to explain their mathematical thinking, students typically describe the procedures they used to get their answers.

Example: In proving whether $a \times b \div b = a$ is true a student generates the example of $5 \times 4 \div 4 = 5$. 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: Students rarely provide explanations. When they do, their explanations typically are procedural, and their justifications are generally 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."

8. 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.

NOTE: this dimension focuses more on the nature of the activity/task than the enactment. To receive a high rating, problems should not be routine or algorithmic; they should consistently require novel, challenging, and/or creative thinking.

High: Problem solving is an integral part of the class' mathematical activity. Students work on problems that are complex, integrate a variety of mathematical topics, and lend themselves to multiple solution strategies. Sometimes problems have multiple solutions OR sometimes students are 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. The teacher reinforces

the idea that there are many different strategies for solving the problem and a variety of answers because the students used different estimates of car length to solve the problem.

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: Problem solving occurs occasionally and is a central component of some of the class’ mathematical activity. For the most part, students work on problems that incorporate one or two mathematical topics and require multiple steps. Some problems lend themselves to multiple solution strategies. Rarely if ever do problems have multiple solutions AND rarely are students asked to formulate problems.

Example: During a unit on measurement, the teacher presents problems such as: “A car is exactly 3.5 meters long. If you lined this car up bumper to bumper with other cars of the same size, about how many car lengths would equal the size of a blue whale?” After solving the problem in groups, the teacher asks the groups to show how they got their answer. She highlights the fact that they came up with several different and creative strategies for solving the problem.

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, or not at all. 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: During a unit on measurement, the teacher presents problems such as: “A car is exactly 3.5 meters long. If you lined this car up bumper to bumper with four other cars of the same size, how long would the cars be all together?” Before the students begin to solve the problem, the teacher uses a diagram to model the strategy for solving the problem. After the students solve the problem in groups, the teacher makes sure they all got the correct answer.

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?”

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important mathematical ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making).

[Often the observer will need to make inferences about how the teacher is using the information, or plans to use the information, especially on a daily observation.]

High: Assessment takes multiple forms, occurs throughout the unit, and taps a variety of mathematical thinking processes. Assessment is used to provide substantive feedback to students about their mathematical understanding, and to inform 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 lesson and engages the class in a discussion of the various representations focusing on several specific examples from the students' work. As a follow-up, she gives students a homework assignment or 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. There is a limited variety of assessment strategies, only some indication that assessment drives instructional decision-making, or limited evidence of substantive feedback to students.

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 to do the assignment. In response, she talks aloud as she constructs a correct version of the graph on the overhead. She tells the students they will have a quiz the next day, in which they will be asked to explain the meaning of a graph which she provides for them. She then begins a new lesson on another aspect of graphing linear equations.

Low: Assessment has few of the features mentioned above. There is little indication of a variety of formal and informal assessment strategies. There is little evidence that assessment drives instructional decision-making or is used to provide substantive feedback to students.

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. The teacher goes over two or three problems that were missed by the greatest number of students.

10. 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. The extent to which series of lessons helps students apply mathematics to real world contexts and to problems in other disciplines.

NOTE: The experiences may be teacher-generated or student-generated, but they should relate to the students' actual life situations.

High: Students are regularly asked to make connections between the math they are learning in class and their own experiences, the world around them, and other academic disciplines. Students learn to apply classroom math in contexts that are relevant to their own lives. Explicit connections are made between the mathematics and the students' personal experiences.

Example: In a lesson on percentages, students are engaged in a discussion about 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. The teacher poses a final question asking students when each of the differing representations would be used and why.

Example: In a lesson on representing data with tables and graphs, the students discuss ways in which they have displayed data in lab exercises in their science class. Several students bring in examples of data tables from their science class and they work in groups to display these data using graphs.

Medium: Students have some opportunities to connect math to their own experience, the world around them, and other disciplines, and to apply the mathematics they are learning to real-world settings. However, these opportunities occur only occasionally, or the examples are potentially relevant to the students' own lives but not explicitly connected to their experiences.

Example: In a lesson on computing percentages, the teacher shares a newspaper article about the fact that the income tax has risen. The 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 income 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 their own experience, the world around them, and other disciplines, or to apply the mathematics they learn to the world around them. When connections/applications are made, they are through happenstance and are 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.

11. Overall. How well the series of lessons reflects a model of instruction consistent with dimensions previously described. This dimension takes into account both the curriculum and the instructional practices.

NOTE: The rating on this dimension is implicitly a weighted average of the ratings on the first ten dimensions, with greater weight being given to Problem Solving, Cognitive depth, Mathematics Discourse Community, and Explanation/Justification.

[For Gold Standard and Notebook Ratings add:

12. Notebook Completeness. The extent to which the notebook contains all the materials we asked teachers to assemble.

High: The notebook contains clear examples of almost all of the requested materials, including:

- Summary of content for the Scoop period
- Information about each day's lesson (content, instructional activities, materials used, student work in class, grouping of students for class work, homework assigned, and projects worked on)
- Complete pre-Scoop reflections (context of teaching situation, typical lesson, assessing student learning, overall plans for Scoop period)

- Complete post-Scoop reflections (fit of Scoop lessons in long-term goals, how representative the Scoop lessons are of own teaching, how well the Scoop notebook represents teaching, suggestions for additions to Scoop notebook)
- Complete daily reflections (objectives/expectations, lesson plan and changes, meeting objectives/expectations, effect of today's lesson on tomorrow's plan)
- Sufficient number of pictures and a completed photo log
- Examples of student work for three different assignments; including a range of work from low to high with completed teacher reflections on the work
- Example of a student assessment task with a corresponding reflection

Medium: The notebook contains clear examples of many of the requested materials, but some materials are not clear or are missing altogether.

Low: The notebook contains clear examples of a few of the requested materials, but most materials are not clear or are missing altogether.

13. Confidence. The degree of confidence you have in your ratings of the notebook across all dimensions.

High: I was able to rate the notebook on almost all dimensions with certainty. For each dimension, the evidence in the notebook matched well with one of the levels on the rating scale.

Medium: I was able to rate the notebook on many dimensions with certainty, but on some dimensions it was difficult to determine what rating to assign based on the evidence in the notebook.

Low: I was able to rate the notebook with certainty on at most one or two dimensions. On most dimensions I had difficulty determining what rating to assign based on the evidence in the notebook.]

Mathematics SCOOP Rating Reporting Form

Teacher: _____ Date: _____

Rater: _____

| 1. Grouping | (Circle one) 1 2 3 4 5 |
|----------------------|------------------------|
| <i>Justification</i> | |

| 2. Structure of Lessons | (Circle one) 1 2 3 4 5 |
|-------------------------|------------------------|
| <i>Justification</i> | |

Note: Scoring sheets continue in this manner for the rest of the dimensions.

APPENDIX B
Science SCOOP Rating Guide
CRESST Artifact Project

The Rating Guide is designed to be used for three types of ratings: observation, Scoop Notebook, and Gold Standard. In general, the language applies equally well in all cases. However, in some cases, we modified language or added examples to reflect differences between ratings of one type or another. [Comments that apply just to observation, notebooks or gold standard ratings are noted in brackets.]

The Rating Guide consists of three parts: a quick reference guide; a description of all the rating levels with examples; and a reporting form for recording ratings and justifications/evidence.

For all dimensions (unless otherwise specified)...

- *Rate each dimension based on the highest level you found in the notebook or observed during the lesson. (Guiding principle: “When in doubt, be nice.” i.e., give the higher of the 2 ratings.)*
- *The rating should take into account teacher, students, and materials that are used.*
- *Remember, a rating of “5” does not mean perfection; it means that the lesson or series of lessons meets the description of a 5.*
- *One characteristic (limitation) of the Scoop rating scale is that there are many different ways a classroom can be a “medium” on each dimension.*
- *A rating of “medium” may be based on the frequency of multiple features of a dimension (e.g. assessment) and/or different levels of enactment by teachers and students (e.g. explanation/justification). In particular:*
 - *frequent occurrence of some features and limited occurrence of others*
 - *medium occurrence of all features*
 - *medium levels of enactment by both teacher and students*
 - *high level of enactment by one and low level by the other*

Specific Notes for Observation Ratings:

1. Take notes during the observation of each lesson.
2. Complete an observation rating each day and then an “summary rating” at the end of the series of observations (after all days of observation).
3. The summary rating completed at the end of the series of observations is a holistic rating (rather than mathematical average).
4. It is sometimes difficult to rate a dimension based on the observation of one lesson, especially when the dimension description includes a “series of lessons.”

Specific Notes for Gold Standard Ratings:

1. Use your rating forms and notes from the class observations, as well as the Scoop Notebook, in order to determine the rating for each dimension. In other words, use “everything that you know” to determine the gold standard ratings.
2. When explaining your rating of each dimension, describe the evidence you used from both your observations (overall rating) and the notebook (student work, reflections, pictures, lesson plan, etc.) to determine your rating.

3. Two areas are included in the Scoop Notebook and gold standard ratings (but not the observation ratings): notebook completeness and confidence.

Quick Reference Guide for CRESST Science SCOOP Ratings

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on scientific tasks that are directly related to the scientific goals of the lesson, and to enable students to together to complete these activities. Active teacher role in facilitating groups is not necessary.

[The focus for observation of a single lesson is on the nature of the activities, and how integral the group activity is to the substance of the lesson (and not necessarily the amount of time spent in groups).]

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related scientifically and build on one another in a logical manner.

NOTE: The focus of this dimension is on design, rather than enactment. [Ratings of observations should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.]

3. Use of Scientific Resources. The extent to which a variety of scientific resources (e.g., computer software, internet resources, video materials, laboratory equipment and supplies, scientific tools, print materials,) permeate the learning environment and are integral to the series of lessons. These resources could be handled by the teacher and/or the students, but the lesson is meant to engage all students. By variety we mean different types of resources OR variety within a type of scientific resource.

4. "Hands-On". The extent to which students participate in activities that allow them to physically engage with scientific phenomena by handling materials and scientific equipment.

NOTE: The emphasis is on direct observation and interaction with scientific equipment and physical objects, to address the substance of the science lesson. Acting out a scientific phenomenon does count. Computers don't unless use involves equipment such as probes.

5. Inquiry. The extent to which the series of lessons involves the students actively engaged in posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.

NOTE: There is a "high bar" on this one. The focus is on the enactment of the lesson and student engagement. A key question is whether the unit/activity is designed so that all phases of inquiry are part of the unit, not whether we observe all phases during the Scoop days. To be true to the intent of this dimension, we should make inferences about the features of inquiry that are incorporated into the entire investigation.

6. Cognitive Depth. Cognitive depth refers to a focus on the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and connections and relationships among science concepts. This dimension considers two aspects of cognitive depth: lesson design and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently promotes cognitive depth.

7. Scientific Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their scientific ideas honestly and openly. The extent to which the teacher and students “talk science,” and students are expected to communicate their scientific thinking clearly to their peers and teacher, both orally and in writing, using the language of science. NOTE: There is a “high bar” on this one, because there is an expectation for student active role in promoting discourse; this should not be only the teacher’s role. This is in contrast to Explanation/Justification. The rating does take into account whether discourse focuses on science content but not the cognitive depth of that content.

8. Explanation/Justification. The extent to which the teacher expects and students provide explanations/justifications, both orally and on written assignments. NOTE: This one is different from “cognitive depth” because it is not dependent on “big ideas” in the discipline.

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important scientific ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making). [Often the observer will need to make inferences about how the teacher is using the information, or plans to use the information, especially on a daily observation.]

10. Connections/Applications. The extent to which the series of lessons helps students: connect science to their own experience and the world around them; apply science to real world contexts; or understand the role of science in society (e.g., how science can be used to inform social policy). NOTE: The experiences may be teacher-generated or student-generated, but they should relate to the students’ actual life situations or social issue relevant to their lives.

11. Overall. How well the series of lessons reflects a model of instruction consistent with dimensions previously described. This dimension takes into account both the curriculum and the instructional practices. NOTE: The rating on this dimension is implicitly a weighted average of the ratings on the first ten dimensions, with greater weight being given to Inquiry, Cognitive Depth, Scientific Discourse Community, and Explanation/Justification to the extent that the rater felt he/she could rate these dimensions accurately.

For notebooks and gold standard ratings add:

12. Notebook Completeness. The extent to which the notebook contains all the materials we asked teachers to assemble.

13. Confidence. The degree of confidence the rater has in his/her ratings of the notebook across all dimensions.

Description of CRESST Science SCOOP Rating Levels

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on scientific tasks that are directly related to the scientific goals of the lesson, and to enable students to together to complete these activities. Active teacher role in facilitating groups is not necessary.
[The focus for observation of a single lesson is on the nature of the activities, and how integral the group activity is to the substance of the lesson (and not necessarily the amount of time spent in groups).]

High: Teacher designs activities to be done in groups that are directly related to the scientific goals of the lesson. The majority of students work on these activities in 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: Teacher designs activities to be done in groups, but some students work independently on the activities, without interacting with other students OR students occasionally work in groups doing activities that are directly related to the scientific goals of the lesson OR students regularly work in groups doing rote or routine activities (e.g. checking each other's homework for accuracy and completeness, quiz each other on scientific terminology).

Example: In a unit on the solar system, each day 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: Students do not work in groups OR group activities do not involve science.

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. They form into groups to practice for the upcoming state test by bubbling in answer sheets.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related scientifically and build on one another in a logical manner.

NOTE: The focus of this dimension is on design, rather than enactment. [Ratings of observations should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.]

High: The series of lessons is conceptually coherent throughout; activities are related scientifically and build on one another in a logical manner.

Example: A unit of instruction on air pressure begins by engaging students through a provocative event in which they experience the 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 activity includes opportunities for students to explore and raise questions about their experiences with the orange juice. The teacher then involves students in a logical sequence of experiments and class discussions about air pressure. Lessons culminate in conclusions or generalizations made through evidence gained during students' exploration of the effects 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 series of lessons is conceptually coherent to some extent, but some activities appear to not be related to one another, OR some activities do not appear to follow in a logical order.

Example: A unit of instruction on air pressure begins with the teacher explaining air pressure and its effect on our lives. The next day the teacher hands back a test on force and motion and the class discusses the results. Following that, the teacher involves students in a series of disjointed 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: The series of lessons does not appear to be logically organized and connected.

Example: In a unit on air pressure, students see a video on scuba diving one day, review homework from a previous unit on force and motion the next day, listen to a lecture on the ideal gas law the third day, practice identifying scientific apparatus in preparation for the state test on the next day, and participate in the orange juice/straw experiment described above on the final day.

3. Use of Scientific Resources. The extent to which a variety of scientific resources (e.g., computer software, internet resources, video materials, laboratory equipment and supplies, scientific tools, print materials,) permeate the learning environment and are integral to the series of lessons. These resources could be handled by the teacher and/or the students, but the lesson is meant to engage all students. By variety we mean different types of resources OR variety within a type of scientific resource.

High: The use of a variety of scientific resources forms a regular and integral part of instruction throughout the lesson/series of lessons. The lesson is meant to engage all students (e.g. teacher demonstration in which all students are watching).

NOTE: there are at least two categories of resources – scientific lab resources and print-based resources. Variety could be variety within each category or across the two categories.

Example: On the first day of an ecosystem unit, the students work in pairs in the computer lab on a predator/prey simulation activity from a science cd-rom. The next day, the teacher leads a discussion on ecosystems and uses clips from a video throughout the lesson. The following day, the students are assigned an ecosystem to research in the library. After gathering their information, the students create posters about each of their ecosystems.

Example: As an introduction to a unit on Newton’s Laws, the teacher begins with a free fall demonstration using a variety of objects (e.g. bowling ball, tennis ball, feather,) and a stopwatch. The students all watch the demonstration and individually write predictions, observations, and explanations. The next day the teacher shows the students how to access data from the NASA website and asks them to use the data to discover the rate of falling objects. On the following day, a professional skydiver comes to talk about her own experience with free falling.

Medium: The series of lessons has some but not all of the features mentioned above. A limited variety of resources are used, OR a variety of resources are used, but only occasionally, OR some but not all students have access.

Example: Throughout the Scoop timeframe, the class is divided into groups of students, each assigned to a different ecosystem. Their task is to create a poster that represents the interactions of the organisms in their ecosystem. For three days, the groups work on their posters drawing information from their textbook and a science cd-rom.

Example: As an introduction to a unit on Newton’s Laws, the teacher begins with a free fall demonstration using a variety of objects (e.g. bowling ball, tennis ball, feather,) and a stopwatch. The students watch the demonstration and individually write predictions, observations, and explanations. The next day the teacher lectures and uses video clips about free fall. For the remaining time in the Scoop, the students work on questions from the textbook.

Low: Scientific resources are rarely used in class other than textbooks and worksheets.

Example: Throughout the Scoop timeframe, the class is divided into groups of students, each assigned to a different ecosystem. Their task is to create a poster that represents the interactions of the organisms in their ecosystem. The students use their science textbooks as resources.

Example: As an introduction to a unit on Newton’s Laws, the teacher conducts a lesson using power point and the students copy notes from the presentation. To conclude the lesson, the students work on questions from the textbook.

4. “Hands-On”. The extent to which students participate in activities that allow them to physically engage with scientific phenomena by handling materials and scientific equipment.

NOTE: The emphasis is on direct observation and interaction with scientific equipment and physical objects, to address the substance of the science lesson. Acting out a scientific phenomenon does count. Computers don't unless use involves equipment such as probes.

High: During a series of lessons, all students have regular opportunities to work with materials and scientific equipment.

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. The next day students cut out pictures of different types of plate boundaries, assemble them on a separate sheet of paper, and label and define each one. Later in the unit, students perform a lab on convection currents using a variety of laboratory equipment (e.g. beakers, hot plate, food coloring) to further their understanding of the mechanics of plate movement.

Medium: During a series of lessons, some of the students work regularly with materials or scientific equipment OR all students work with materials or scientific equipment but only occasionally.

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. Later in the unit, students supplement their reading about faults by using wooden blocks to represent different fault types.

Low: There are no activities that require students to handle or work with materials or scientific equipment (other than pencil and paper).

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: During a series of lessons on plate tectonics, the students take notes while the teacher lectures. The students read the textbook to supplement the lectures.

5. Inquiry. The extent to which the series of lessons involves the students actively engaged in posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.

NOTE: There is a "high bar" on this one. The focus is on the enactment of the lesson and student engagement. A key question is whether the unit/activity is designed so

that all phases of inquiry are part of the unit, not whether we observe all phases during the Scoop days. To be true to the intent of this dimension, we should make inferences about the features of inquiry that are incorporated into the entire investigation.

High: Over a series of lessons, students are engaged in all features of inquiry including posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.

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. 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. The groups collect the data, analyze them, and present their results

Medium: The series of lessons has some but not all of the features mentioned above. Students are occasionally engaged in designing investigations and finding answers to scientific questions OR engagement occurs regularly but does not include all components of the inquiry process.

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 show how the temperature of different cities has changed over the past 100 years. They select cities, gather data from the library, graph the information and report what they found.

EXAMPLE OF A "2" RATING:

Example: Students follow a carefully scripted lab manual 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 conclude with answering factual questions in the lab manual.

NOTE: Another situation that would receive a lower rating is one in which the teacher does one thing well and richly (e.g., have students pose questions), but doesn't carry it through, and the rater sees no evidence that the class is on a trajectory to carry it through.

Low: During a series of lessons, students are rarely or never engaged in scientific inquiry.

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.

6. Cognitive Depth. Cognitive depth refers to a focus on the central concepts or “big ideas” of the discipline, generalization from specific instances to larger concepts, and connections and relationships among science concepts. This dimension considers two aspects of cognitive depth: lesson design and teacher enactment. That is, it considers the extent to which lesson design focuses on achieving cognitive depth and the extent to which the teacher consistently promotes cognitive depth.

NOTE: If you are unfamiliar with the content area for the unit studied during the Scoop period, refer to the state or national Science standards to better understand the “big ideas.”

High: Lessons focus on central concepts or “big ideas” and promote generalization from specific instances to larger concepts or relationships. The teacher consistently promotes student conceptual understanding. The teacher regularly attempts to engage students in discussions or activities that address central scientific ideas and principles.

Example: The teacher designs a series of lessons in which students are asked to use their understandings of the relative motions of the Earth, sun, and moon and how light is reflected between these celestial bodies to demonstrate and explain the phases of the moon. Students work in groups to develop a kinesthetic model and verbal explanation of their understanding of this concept, and then present their ideas to their classmates and teacher. After the group demonstrations, the teacher facilitates a discussion in which students compare and contrast the different groups’ portrayals of the concept.

Medium: The series of lessons has some but not all of the features mentioned above. Lessons may focus on mastery of isolated concepts, but not on connections among them, (e.g. lessons may require students to explain or describe the concept but not to use it or apply it). OR, the teacher sometimes attempts to engage students in discussions about connections between scientific concepts and sometimes responds to students in ways that promote student conceptual understanding.

Example: During a class discussion, the teacher asks students to explain the phases of the moon. They respond with a description of the experiment from the day before on reflection of light. They also describe that light from the sun reflects off the moon, however they do not discuss the relationship between the reflection of light and the location of the sun, Earth, and moon as the key to the phases of the moon.

Low: The series of lessons focuses on discrete pieces of scientific information, e.g., disconnected vocabulary, definitions, formulas, and procedural steps. These are elements of science that can be memorized without requiring an understanding of the larger concepts. The teacher rarely attempts to engage

students in instructional activities that demonstrate the connectedness of scientific concepts and principles. The teacher's interactions with students focus on correctness of their answers rather than on conceptual understanding.

Example: Over a series of lessons, the students learn the orbit of the moon and Earth, and the names for each phase of the moon. As a culminating activity, students complete a fill-in-the-blank worksheet of the phases of the moon.

7. Scientific Discourse Community. The extent to which the classroom social norms foster a sense of community in which students feel free to express their scientific ideas honestly and openly. The extent to which the teacher and students "talk science," and students are expected to communicate their scientific thinking clearly to their peers and teacher, both orally and in writing, using the language of science. NOTE: There is a "high bar" on this one, because there is an expectation for student active role in promoting discourse, this should not be only the teacher's role. This is in contrast to Explanation/Justification. The rating does take into account whether discourse focuses on science content but not the cognitive depth of that content.

[The kind of indirect evidence we might find in the notebook includes:

- teacher reflections, such as:
 - I had students compare their solution strategies with one another;
 - I consciously try to get students to voice their ideas;
 - I walked around and listened to students' conversations
 - I encourage students to ask each other questions when they present their solutions to the class.
- peer reflections on student written work
- lesson plans showing discussion of scientific topics]

High: Students consistently are encouraged to express their scientific reasoning to other students and the teacher, and they are supported by the teacher and other students in their efforts to do so. Students' ideas are solicited, explored, and attended to throughout the lesson, and students consistently use appropriate scientific language. Emphasis is placed on making scientific reasoning public, raising questions and challenging ideas presented by classmates.

Example: Students work in groups, investigating plant growth. 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. Following the group work, students present their findings to the class. Classmates actively engage in a critique of each presentation by raising questions, challenging assumptions, and verbally reflecting on their reactions to the findings presented. The teacher asks probing questions, and pushes the scientific thinking of both presenters and peers. These discourse patterns appear to be the norm.

Example: In a class discussion on the behavior of gases, the teacher asks students to share their thinking about why the diameter of a balloon increases when placed in hot water and decreases when placed in cold water. The teacher uses wait time to allow students to formulate their thinking. When students share their ideas, the teacher listens carefully and

asks other students to reflect on, build on, or challenge the ideas presented by their classmates. The teacher may offer suggestions or alternative ways of thinking about the question when gaps in student thinking are evident, but does not engage in correcting students' ideas, or in giving the "real/right" answer.

Example: During a lesson on cell structure and function, the teacher asks students to work in pairs on a lab activity. Their task is to determine the effect of a salt solution on green plant cells. Prior to the activity, each pair creates a hypothesis statement. They prepare their microscope slide and write down observations; describing the effects of salt and identifying various cell structures, and discuss the lab directed questions challenging each other's scientific reasoning and formulating their conclusions together.

Medium: Students are expected to communicate about science in the classroom with other students and the teacher, but communication is typically teacher-initiated (e.g., the teacher attempts to foster student-to-student communication but students don't communicate with each other without teacher mediation) OR, student communication is directed to the teacher. The use of appropriate scientific language may or may not be consistent.

Example: Students work in groups, investigating plant growth. The teacher moves around the room listening to their discussions. When students stop her and ask questions, the teacher responds by providing suggestions or answers. Following the group work, 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 questions to elicit both procedural and conceptual understanding from the presenters. The teacher supplements students' answers with 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 on the behavior of gases, the teacher asks students to reflect on how air particles might be affecting the diameter of a balloon when it is moved from a 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?"

Example: During a lesson on cell structure and function, the teacher has the students sitting in groups of four, sharing a microscope and prepared slides. Their task is to determine the effect of a salt solution on green plant cells. Prior to the activity, each student creates a hypothesis statement. Throughout the lab activity, the students ask questions to each other, but are not necessarily challenging each other's scientific reasoning.

Low: The teacher transmits knowledge to the students primarily through lecture or direct instruction. Those discussions that occur are typically characterized by IRE (initiation, response, evaluation) or "guess-what's-in-my-head" discourse patterns. Students rarely use appropriate scientific language. Student-to-student communication, when it occurs, is typically procedural and not about science.

Example: Following an investigation on plant growth, the 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 the behavior of gases, explaining that all things (including air) are made up of particles; those particles move more quickly and with greater energy when they are heated up and they move more slowly when they are cooled down. The teacher follows this lecture with a demonstration of how the diameter of a balloon decreases when moved from a bowl of hot water to a bowl of cold water. She then asks the class to use the information that they learned in her lecture to complete a worksheet on which they explain why the diameter of the balloon decreased.

Example: During a lesson on cell structure and function, the teacher has students individually work through a microscope lab activity on their own. The students are asked to state a hypothesis, follow the directions of the lab and complete concluding questions.

8. Explanation/Justification. The extent to which the teacher expects and students provide explanations/justifications, both orally and on written assignments.

NOTE: This one is different from "cognitive depth" because it is not dependent on "big ideas" in the discipline.

High: Teacher consistently asks students to explain/justify their scientific reasoning, both orally and on written assignments. Students' explanations show their use of concepts or scientific evidence to support their claims. NOTE: We need to see evidence not only of teacher expectations, but also of a variety of students giving explanations/justifications.

Example: Following a whole class discussion on plate boundaries, the teacher poses a question for students to begin in class and complete for homework. The teacher asks the students to explain how the geologic features found near Nepal were created. Using maps in the classroom one student indicates that there is a mountain range present in this region. The student compares a map of plate boundaries with a world map and points out that Nepal is located along a plate boundary. For homework, she uses data found from the Internet about the recent tectonic activity and is able to further her argument of converging plates with the data. The next day, she explains to the class, using her evidence from the maps and Internet search that two continental plate boundaries are converging to create mountains. In the discussion that follows, several other students also share their explanations and supporting evidence. It appears that this type of discussion following a homework assignment is the norm.

Example: Throughout a unit on plant anatomy and physiology, the teacher incorporates a series of experiments with plants. On the first day of the Scoop, the students are analyzing their data from the most recent plant experiment. The teacher asks each lab group to explain whether their data support their hypotheses and then to justify their conclusions. After writing these explanations and justifications in their lab reports, the teacher asks them to find textual evidence to support or refute their explanations. The following day, the groups take turns presenting their explanations and justifications to the class.

Medium: Teacher sometimes asks students to explain/justify their scientific reasoning and students sometimes provide explanations/justifications that use concepts and scientific evidence to support their claims OR teacher consistently

asks students to explain their scientific reasoning, but students rarely provide such explanations.

Example: Following a whole class discussion on plate boundaries, the teacher poses a question for students to begin in class and complete for homework. The teacher asks the students to explain how the geologic features found near Nepal were created. One student looks in her textbook and on the Internet to help answer the question. She compares a map of plate boundaries with a world map and notes that Nepal is located on a plate boundary. The next day she shows the maps to the class, and uses the information to explain that two continental plate boundaries are converging to create mountains. One other student also shares her explanation and evidence. The teacher poses similar questions at the end of some lessons and some students respond with similar concrete explanations.

Example: As one component of a unit on plant anatomy and physiology, the students perform a series of experiments with plants in which they collect and record data. At the conclusion of these experiments, the teacher asks each lab group to explain whether their data support their hypotheses and then to justify their conclusions. The teacher continues the following day with a lecture on plant growth, during which the students take notes. The next day there is a fill-in-the-blank and multiple choice quiz.

One possibility for a "2" rating:

Teacher sometimes asks students to explain/justify their scientific reasoning, but students rarely provide such explanations.

Low: Teacher rarely asks students to explain/justify their scientific reasoning, and students rarely provide explanations/justifications. When they do, they are typically concrete or copied from text or notes.

Example: A teacher uses a world map to show the class where the Himalayas are located and points out that they are along a plate boundary. She asks the students to explain how the mountains could have been created. A student responds by reading from the notes from the previous class: "Mountains are created by two converging continental plates."

Example: For a unit on plant anatomy and physiology, the teacher begins with an experiment. The students follow the procedures and use their data to answer multiple-choice questions at the end of the lab handout. The following day the teacher gives a lecture on plant growth. The students are given a worksheet to start in class, which has fill-in-the-blank questions. The teacher encourages the students to use their notes and text to find the answers.

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies that measure student understanding of important scientific ideas and furnish useful information to both teachers and students (e.g., to inform instructional decision-making).

[Often the observer will need to make inferences about how the teacher is using the information, or plans to use the information, especially on a daily observation.]

High: Assessment takes multiple forms, occurs throughout the unit, and includes measures of students' understanding of important scientific ideas. Assessment is used to provide feedback to students about their understanding of science (not just whether or not their answers are correct), and to inform instructional practice.

Example: The first assignment in the lesson on plate tectonics reveals that students did not learn the concepts well from the book, so the teacher adds an additional lesson to the unit. He sets up four different plate simulations, using a variety of materials. Students are divided into four groups and assigned one activity to work on. They present their activity and description of their observations to the full class. During this time, the teacher asks probing questions to “get at their conceptual understanding.” Students receive a group grade for their presentation. The class concludes with each student writing what they understand about each demonstration on plate tectonics and what they find confusing.

Example: The lesson on chemical changes begins with a lab activity, and students’ written lab observations are reviewed by the teacher who writes questions and gives suggestions for clarification. The next day, students use their textbook and library materials to prepare a short paper using information derived from their lab notebook and responding to the teacher’s comments. A test at the end of the unit asks factual and reasoning questions.

Medium: Assessment has some but not all of the features mentioned above. There is a limited variety of assessment strategies, limited focus on important scientific ideas, only some indication that assessment drives instructional decision-making or limited evidence of substantive feedback to students.

Example: In the lesson on plate tectonics, the students turn in a homework assignment that is graded by the teacher. The students work with a partner to make corrections (get the right answers). The teacher decides to postpone the test until the next day because he sees that the students need more time to work on the corrections with their partners.

Example: A week-long unit on chemical change involves three activities that are graded with teacher comments: a homework assignment, an in-class writing assignment, and an exam consisting of multiple choice items and one essay. Results count toward grades but are not otherwise used. There is no evidence that the students were asked to revise any of the work based on the teacher’s comments.

Low: Assessment has few of the features mentioned above. There is little indication of a variety of formal and informal assessment strategies. The assessments focus on a recall of facts rather than understanding of important scientific ideas. There is little evidence that assessment drives instructional decision-making or is used to provide substantive feedback to students.

Example: The class is studying plate tectonics and they take a multiple-choice test when the unit is completed.

Example: A series of lessons on chemical change ends with a worksheet scored by the teacher

10. Connections/Applications. The extent to which the series of lessons helps students: connect science to their own experience and the world around them; apply science to real world contexts; or understand the role of science in society (e.g., how science can be used to inform social policy).

NOTE: The experiences may be teacher-generated or student-generated, but they should relate to the students’ actual life situations or social issue relevant to their lives.

High: Teacher or students regularly make connections between the science they are learning in class and their own experiences and the world around them. Students learn to apply classroom science in contexts that are relevant to their own lives or to consider the role of science in society (for example, how science can be used to inform social policy).

Example: As a conclusion to an ecology unit, the students are asked to help the school address the problem of fish dying in the pond behind the school. The students divide into groups and pick individual topics to research (pond life, water chemistry, pond floor composition). After sharing their findings with each other, the class creates a summative report for the principal and school board that include recommendations for action.

Example: The class is learning about Newton's Laws of Motion. After learning about each law and doing simple demonstrations in the class, the teacher asks the students to work in groups to design and perform a demonstration of the law. They are required to collect and analyze data using one form of motion from their own lives (e.g., biking, riding a rollercoaster, skateboarding, skiing) and to comment about the safety of one activity from a scientific perspective.

Medium: Teacher or students sometimes make connections between the science they are learning in class and their own experiences, or the world around them. Students have some opportunities to learn to apply classroom science in contexts that are relevant to their own lives or to consider the role of science in society (for example, how science can be used to inform social policy). However, these opportunities occur only occasionally, or the examples are potentially relevant to the students' own lives or to the role of science in society, but these connections are not made explicit.

Example: As a conclusion to an ecology unit, the students work in groups, each studying a different lake, assigned by the teacher that has been identified as having an unstable ecosystem. They locate data on the water chemistry and fish life of the lake using library-based resources and write a report to share with the class.

Example: After completing a month-long unit on Newton's Laws, the teacher asks the students to work in groups to design and perform a demonstration of one of Newton's laws. They are required to collect and analyze data using one form of motion from their own lives (e.g., biking, riding a rollercoaster, skateboarding, skiing).

Low: Students are rarely asked to make connections between the science learned in the classroom and their own experience, the world around them, and other disciplines, or to apply the science they learn to social policy issues. When connections/applications are made, they are through happenstance, are not a planned effort on the part of the instructor and not elaborated upon by the teacher or integrated into the lesson.

Example: As a conclusion to an ecology unit, the students work in groups, each studying an ecosystem from the textbook (tundra, rainforest, and ocean). Each group writes a report and makes a poster to share with the class.

Example: During a unit on Newton's Laws, the teacher uses demonstrations and lab activities from the lab manual (i.e. ramps with rolling objects, pendulum).

11. Overall. How well the series of lessons reflects a model of instruction consistent with dimensions previously described. This dimension takes into account both the curriculum and the instructional practices.

NOTE: The rating on this dimension is implicitly a weighted average of the ratings on the first ten dimensions, with greater weight being given to Inquiry, Cognitive Depth, Scientific Discourse Community, and Explanation/Justification to the extent that the rater felt he/she could rate these dimensions accurately.

[For gold standard and notebook ratings add:

12. Notebook Completeness. The extent to which the notebook contains all the materials we asked teachers to assemble.

High: The notebook contains clear examples of almost all of the requested materials, including:

- Summary of content for the Scoop period
- Information about each day's lesson (content, instructional activities, materials used, student work in class, grouping of students for class work, homework assigned, and projects worked on)
- Complete pre-Scoop reflections (context of teaching situation, typical lesson, assessing student learning, overall plans for Scoop period)
- Complete post-Scoop reflections (fit of Scoop lessons in long-term goals, how representative the Scoop lessons are of own teaching, how well the Scoop notebook represents teaching, suggestions for additions to Scoop notebook)
- Complete daily reflections (objectives/expectations, lesson plan and changes, meeting objectives/expectations, effect of today's lesson on tomorrow's plan)
- Sufficient number of pictures and a completed photo log
- Examples of student work for three different assignments; including a range of work from low to high with completed teacher reflections on the work
- Example of a student assessment task with a corresponding reflection

Medium: The notebook contains clear examples of many of the requested materials, but some materials are not clear or are missing altogether.

Low: The notebook contains clear examples of a few of the requested materials, but most materials are not clear or are missing altogether.

13. Confidence. The degree of confidence the rater has in his/her ratings of the notebook across all dimensions.

High: I was able to rate the notebook on almost all dimensions with certainty. For each dimension, the evidence in the notebook matched well with one of the levels on the rating scale.

Medium: I was able to rate the notebook on many dimensions with certainty, but on some dimensions it was difficult to determine what rating to assign based on the evidence in the notebook.

Low: I was able to rate the notebook with certainty on at most one or two dimensions. On most dimensions I had difficulty determining what rating to assign based on the evidence in the notebook.]

Science SCOOP Rating Reporting Form

Rater: _____ Date: _____

Teacher: _____

| 1. Grouping | (Circle one) 1 2 3 4 5 |
|----------------------|------------------------|
| <i>Justification</i> | |

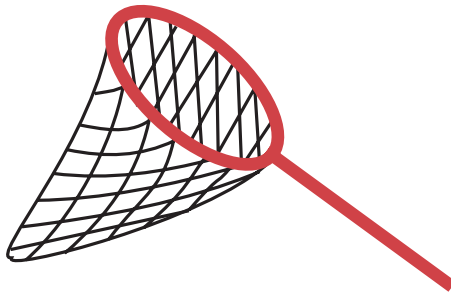
| 2. Structure of Lessons | (Circle one) 1 2 3 4 5 |
|-------------------------|------------------------|
| <i>Justification</i> | |

Note: Scoring sheets continue in this manner for the rest of the dimensions.

APPENDIX C

The Scoop Notebook (Science)

THE SCOOP PROJECT



**What Is It Like to Learn
Science in Your Classroom?**

**University of Colorado, Boulder
RAND, Santa Monica**



Table of Contents

Part 1: Project Overview

- ❑ What is it Like to Learn Science in Your Classroom?
- ❑ Scoop Checklist
- ❑ Final Checklist

Part 2: Directions for Collecting the Classroom Scoop

- ❑ Selecting a Class
- ❑ Selecting a Timeframe
- ❑ Taking Photographs
- ❑ Collecting Artifacts
 - Collecting Daily Instructional Materials
 - Labeling Daily Instructional Materials
 - Selecting Student Work
 - Labeling Student Work
 - Selecting a Formal Classroom Assessment Task and Examples of Student Performance on the Assessment
- ❑ Completing the Daily Calendar
- ❑ Responding to Reflection Questions

Part 3: Supporting Documents

- ❑ Pre-Scoop Reflection Questions
- ❑ Daily Reflection Questions
- ❑ Post-Scoop Reflection Questions
- ❑ Calendar
- ❑ Photograph Log
- ❑ Yellow and White Sticky Labels
- ❑ Pocket Folders for Daily Instructional Materials
- ❑ Pocket Folders for Student Work
- ❑ Pocket Folder for Formal Classroom Assessment Task (and Student Work, if applicable)



What Is It Like to Learn Science in Your Classroom?

Capturing What Takes Place in Your Science Class

We are trying to find ways to describe science instruction that capture the important features of each class. Probably the most accurate way to do this would be to observe every lesson, review every assignment, and examine every test. However, this method is too burdensome for both teachers and researchers. As a result, we are trying out an alternative way to collect information that will tell us about each science class.

A Scoop of Classroom Material

One way that scientists study unfamiliar territory (e.g., freshwater wetlands, Earth's crust) is to scoop up all the material they find in one place and take it to the laboratory for careful examination. Analysis of a typical scoop of material can tell a great deal about the area from which it was taken.

We would like you to do something similar in your classroom, i.e., scoop up a typical week's worth of material that we can use to learn about your class. The artifacts would include assignments, homework, tests, projects, problem solving activities, and anything else that is part of instruction during the week.

Some things you might include as part of your Scoop are:

- materials prepared for the class: e.g., worksheet assignments, overhead transparency masters, tests, formal classroom assessments
- materials generated during class: e.g., in-class notes, problems on the board
- materials produced by students: e.g., homework, in-class assignments, projects, journal entries, portfolio pieces
- photographs of the classroom taken during the week, to provide a visual image of the teaching and learning environment: e.g., the seating arrangement of your room each day, the white/chalkboard at different periods throughout the lesson, room arrangement and equipment for laboratory activities, student projects.

If you think of other things that could help us to understand your classroom, please include these as part of your Scoop.

Reflections on the Scoop

In addition, we want to know about your plans for the Scoop week, your reactions to each day's lesson, and your overall thoughts after the lessons are complete. We will ask you to respond to reflection questions:

- before the Scoop period begins,
- after each day's lesson, and
- at the end of the series of lessons in the Scoop timeframe.



Scoop Checklist

Please read this checklist before beginning the Scoop. Once you have started the Scoop, refer back to the checklist as necessary to make sure you have completed the daily activities. At the end of the Scoop period, review this checklist, as well as the Final Checklist, to be sure your notebook is complete.

Before the Scoop:

- Read through your Scoop Notebook and ask questions about anything that is unclear.
- Select a class and timeframe to use for the Scoop.
- Write or tape record a Pre-Scoop Reflection.

Each Day During the Scoop:

- Record your class plans daily in the Scoop calendar provided.
- Collect daily instructional materials, including handouts and worksheets used for instruction. Label each item with one of the yellow sticky labels provided, and complete the information on the label.
- Take photos of your classroom with the camera provided and record descriptions of the photos in the photograph log.
- Write down or photograph information and assignments from your chalk/whiteboard and overhead transparencies daily (if these transparencies cannot be copied).
- Select and copy samples of student work. Label each item with a white sticky label, and answer the questions on the label (*Note: Collect at least 3 assignments and 3 samples of student work per assignment over the course of the Scoop*).
- Write or tape record a Daily Reflection at the end of each day.

After the Scoop:

- Write or tape record a Post-Scoop Reflection.
- Make a copy of a recent formal class assessment and scoring rubric or answer key.
- Copy examples of three students' responses to the assessment (if available).
- Review the Final Checklist to be sure the Scoop Notebook is complete.



Final Checklist

Please be sure you have completed all of the items listed below before returning your notebook.

- Pre-Scoop reflections
- The Classroom Scoop materials for 5 days
- Assignments and examples of student work
- Formal classroom assessment task (with student work, if applicable) and scoring rubric
- Completed calendar of scooped classes
- Completed photograph log and used camera
- Daily reflections
- Post-Scoop reflections

THANK YOU FOR YOUR PARTICIPATION!



Directions for Collecting the Classroom Scoop

Please scoop up a typical week's worth of materials from one of your classes. We can then use these scooped materials to learn about your class. You may feel that some of the things we ask you to collect are more reflective of your class than others, but please collect it all.

Selecting a Class

Please collect the Scoop from one of your classes. Choose a class that represents your typical practice.

Selecting a Timeframe

Please collect the Scoop for the equivalent of five consecutive days of instruction in your selected class. It is ideal for the Scoop to begin at the start of a new instructional unit or topic. You may start on any day of the week (i.e., the scooped days do not have to be Monday through Friday). You should start scooping at a logical point in your lesson plans. For example, if you are starting a new unit of instruction on Wednesday, then your Scoop should start on Wednesday.

The five days of instruction might not coincide with five consecutive school days.

If you teach on a block schedule, you should collect your Scoop for the equivalent of five days of instruction on a regular (non-block) schedule, assuming 50-minute periods. This will most likely be 3 days of instruction on the block schedule.

Even if you typically teach your class every day, your class schedule may be disrupted by assemblies, disaster drills, etc. Please do not include these days or other non-instruction days in your Scoop; instead, add replacement instructional days at the end.



Taking Photographs

Photographs are a good way to capture the learning environment in your classroom and to provide a sense of what your daily lessons look like. So, throughout the Scoop timeframe, please **take photographs regularly** with the disposable camera that we provided. We have not obtained permission for students' faces to be photographed, so please try to avoid taking pictures of students' faces.

We are interested in seeing photographs of:

- the classroom set-up (such as seating arrangement) every day
- bulletin boards
- contents of white/chalkboard at several points during the lesson
- student work on the board or on an overhead transparency (if it is not possible to include the actual overhead transparency or a photocopy)
- lesson activities
- instructional tools (e.g., lab equipment, manipulatives, calculators) used during the lesson. If the tools are being used by students, be sure to include only the students' hands in the picture and not their faces.
- students working in class (for example working with a partner or in a group)

Please take pictures of any other things that you feel will help us to better “see” and understand your classroom and teaching.

You may want to consider asking a responsible student in the class to take some of the pictures for you while the lesson is taking place and record them in the photograph log. It would be best if you still prompted the student when to take the picture.

We would like 4 to 5 pictures per day during the Scoop timeframe. Be sure to **complete the photograph log**. See the example below.

Please remember:

- ❑ Try to avoid taking pictures of students' faces.
- ❑ Use the flash when taking pictures. (*Exception: When taking a picture of a white board or overhead, do not use a flash because it creates too much glare. If you do use a flash, take the picture at an angle to the board. It is better to photocopy overhead transparencies than to take pictures of them being projected on the screen.*)
- ❑ Provide a brief description of each photograph that you take in the Photograph Log.



Photograph Log Example

| Photo # on camera | DATE | DESCRIPTION |
|-------------------------|---------|--|
| 27 | 11/2/06 | Layout of my room, looking from the perspective of a student |
| 26 | 11/2/06 | Warm-up activity on the board |
| 25 | 11/2/06 | Notes written on the board during class discussion |
| 24 | 11/2/06 | New room arrangement for laboratory activity |
| 23 | 11/3/06 | Laboratory equipment students will use today |
| 22 | 11/3/03 | A group of students doing the lab activity |
| 21 | 11/3/06 | New room arrangement for quiz review |

Collecting Artifacts

There are two kinds of artifacts that we are interested in collecting: instructional materials and student work. We would like you to scoop **all of your instructional materials**. As the teacher, you will have generated most of these materials; however, there are instances in which students may contribute to instructional materials by writing on an overhead transparency or creating an instructional document. Please include these instructional materials as well.

The second type of artifact to collect is student work. We would like you to scoop **only samples of student work**. Detailed instructions on selecting student work will follow.

Collecting Daily Instructional Materials

It may help to think of scooping materials and taking photographs each day at three different points:



1 BEFORE THE LESSON

Scoop all instructional materials you prepare for the class. For example:

- *written plans*
- *copies of Teacher's Edition's suggestions to teachers, if you are using this to guide your instruction*
- *handouts* (e.g., notes, worksheets, laboratory instructions, problem descriptions)
- *assignments* (e.g., directions for a project, pages to be read in a textbook)
- *overhead transparency masters*
- *tests or other forms of assessment (including rubric, if applicable)*

Photograph any changes to your classroom. For example:

- *the seating arrangement*
- *assignments, questions, problems, or instructions written on the chalk/whiteboard*
- *materials which have been added to the bulletin board*
- *instructional tools or lab equipment to be used during the lesson*

2 DURING THE LESSON

Scoop all instructional materials you generate during the class. For example:

- *notes or problems and questions written on an overhead transparency*
- *notes written to yourself about the lesson*
- *notes written to yourself about the students*

Photograph

- *the white/chalkboard throughout the lesson*
- *changes to the classroom set-up*
- *the set-up and use of instructional tools or lab equipment*
- *students working in groups to solve problems (without students' faces, please)*
- *students working in pairs on laboratory activity*



3 AFTER THE LESSON

Scoop any instructional materials that were not yet scooped before or during the lesson.

For example:

- *copies of overhead transparencies used in class*
- *photocopies of revised lesson plans or notes to self on lesson plan or Teacher's Edition suggestions*
- *copies of student-created instructional materials such as rubrics or review questions*

Photograph

- *any materials created during the lesson that cannot be scooped, such as a poster or project*
- *any changes to the classroom that could not be photographed with students in the room, such as the seating arrangement or set-up of instructional tools or materials*

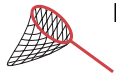
NOTE: Include any additional materials that will help us understand the lesson.

Labeling Daily Instructional Materials (Yellow Sticky Labels)

Use the **yellow sticky labels** to identify all instructional materials you place in the Pocket Folders for Daily Artifacts. Use these to let us know how each item fits into the timeframe of scooped materials. Indicate:

- *the date*
- *a brief description of the artifact.* For example:
 - directions for group project
 - rubric used for grading assessment
 - copy of overhead transparency used for the warm-up activity

Sample Yellow Sticky Label

| | |
|---|-------------------------|
|  | Instructional Materials |
| | Date: _____ |
| Description: | |



Selecting Student Work

We would like you to choose some student work that is indicative of the kind of work your students do. This work can be from individual students or from a group assignment.

Select at least three activities or assignments central to the unit you are teaching during the course of the Scoop. For each activity or assignment, pick three examples of student work that represent a range of quality (high, medium, and low).

Scoop examples of high, medium, and low quality student work for each of the three activities or assignments. For example:

- *worksheets*
- *in-class assignments*
- *journal entries*
- *portfolio pieces*
- *homework assignments*
- *projects*
- *reports or write-ups of problem-solving or laboratory activities*

Note: You do not have to provide sample work from the **same** students for each of the activities or assignments.

Make a photocopy of any student-generated materials for which this is possible. Be sure to cover the student's name before making the photocopy.

For student-generated materials that cannot be photocopied (e.g., a 3D model), please **take a picture** of the student's work. Be sure to cover the student's name before taking the picture, and do not include the student's face in the picture.

Labeling and Reflecting on Student Work (White Sticky Labels)

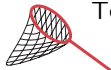
Use a **white sticky label** to identify each sample of student work you place into the Pocket Folder for Student Work.

On each white sticky label please tell us:

- *whether the student work is of high, average, or low quality*
- *why you rated it this way*
- *what this work tells you about the student's understanding of the material*



Sample White Sticky Label

| | |
|---|---|
|  | Teacher Reflection on Student Work |
| | Date: _____ |
| ➤ | Quality of this work? (<i>circle</i>) High Medium Low |
| ➤ | Why did you rate it this way? |
| ➤ | What does this work tell you about the student's understanding of the material? |

Selecting a Formal Classroom Assessment Task and Examples of Student Performance



Select a recent formal classroom assessment task (i.e., test, quiz, prompt or task description for a portfolio piece, paper, performance, final project, demonstration) that is representative of the assessments you use. **It's not necessary for it to have occurred during the timeframe of the Scoop.** Please attach a **yellow** sticky label and include it in the Pocket Folder for Formal Classroom Assessment Task. Also include a scoring rubric or answer key if available.

If you have copies of student responses to the assessment, please choose three examples that represent low, average, and high quality work and include them as well. Also include your written feedback to the students, if available. Please attach a **white** sticky label to each student response and answer the three questions. Be sure to remove the student's name or other identifying marks.

For student responses to an assessment task prompt that cannot be photocopied (e.g., a 3D model, a student performance), please take a picture of the student's work or include a copy of the student's rating sheet on the assessment. Be sure to cover the student's name before taking the picture or making a copy of the rating sheet, and do not include the student's face in the picture.

Please be sure to include a recent assessment, even if you do not have copies of student responses to include with it. Include feedback if available.



Completing the Daily Calendar of “Scooped” Classes

The Calendar of “Scooped” Classes is designed to give us a roadmap of the class sessions from which you will be scooping. The calendar asks for the following information:

□ **DATE**

□ **LENGTH OF SESSION**

For most teachers, this will be the number of minutes in the class period and it will not vary from day to day.

□ **TOPIC OF SESSION**

In this section, please add a descriptive title for the day’s activities. For example:

- Review the names and characteristics of the three groups of rocks (igneous, sedimentary, metamorphic)
- Group Laboratory Activity: observe the characteristics of a rock to decide if it is igneous, sedimentary, or metamorphic
- Use the identification tables in the textbook to identify each rock specimen
- Complete the discussion questions for the laboratory activity

□ **CURRICULUM MATERIALS USED**

List any materials used in the lesson, which were not specified above.

For example:

- If you are using a standard curriculum, indicate the title and page number(s) of the chapter, unit, section, or investigation.
- list of rock specimens utilized in the activity
- directions and handouts for the laboratory activity
- blank copies of identification charts to be completed by students

NOTE: For all of the materials, please be sure you have included a copy or taken a picture with the disposable camera.



Responding to the Reflection Questions

There are three sets of reflection questions:

1. The Pre-Scoop reflection questions should be answered **before** you begin the Scoop collection period.
2. The Daily Reflection questions should be answered **each day** as soon as possible after the scooped class. Having access to your immediate thoughts and reactions is crucial. Please make every effort to jot down your reflections right away after each Scoop class.
3. The Post-Scoop reflection questions should be answered **after** you complete the Scoop period and are ready to turn in your materials.

You may provide your responses in several different formats. Please choose one of the following that is most convenient for you:

- **Write** your answers on a separate sheet of paper.
- **Type** your answers on a computer and print them
- or*
- **Send** them to us on a disk or over email.
- **Audiotape** your answers.

If you choose to audiotape your responses, you do not need to use a new tape each day. However, please state your name and the date at the beginning of each set of responses.



Pre-Scoop Reflection Questions

To be answered once, before the Scoop period begins.

1. *What about the context of your teaching situation is important for us to know in order to understand the lessons you will include in the Scoop?*

This may include:

- characteristics of students
- features of the school and/or community
- description of the curriculum and /or textbook you are using and the students' past experience with it
- anything else you may find pertinent to our understanding of your teaching environment

For example, in the past teachers have told us they wanted us to know about features of their teaching situation such as:

- Many of the students in the class are second-language learners.
- The school just had a large turnover of staff.
- This is the students' first experience with an activity-based curriculum.
- This is the students' first experience with block scheduling.
- Students in this cohort have a reputation for having difficulty working together.

2. *What does a typical lesson look like in your classroom? If it varies day to day, then please describe the various possibilities.*

This may include:

- daily "routine" activities, such as checking homework at the start of class
- the format of the lesson (lecture, discussion, group work, etc.)
- description of a typical week if you have different lesson formats for each day (for example, introduction lecture on Monday, hands-on activity on Tuesday, review questions on Wednesday, etc.)

For example,

- The students come in and start with a 5-minute warm-up question that is written on the board. We then check the homework as a group for about 10 minutes. For the next 20 minutes, I teach the new concept in a whole class lecture/discussion format. Finally, the students work individually (or sometimes in partners) on questions and problems that utilize the concepts taught. During the last few minutes of class, I explain the homework and they copy the assignment from the board.
- It really varies from day to day, depending on the kind of science content we are working on. Usually I have a warm-up review question on the board when the students arrive to class. We discuss it briefly as a whole class to be sure they all understand the



concept. I then sometimes have them do a hands-on extension activity in pairs or groups while I walk around to help answer questions and facilitate the group discussions. When they are done with the activity in their groups, each group takes a turn presenting/defending their findings to the class. Other times, they work individually on problems and questions. I also take the class to the library about 2 or 3 times a month in order to do research for projects or mini reports that I assign. When we are in the library, they typically work either individually or in pairs, depending on the assignment.

3. *How often do you assess student learning, and what strategies/tools do you use?*

This may include commercially-produced assessments, teacher-created assessments, and informal assessments (e.g., check student homework, listen to student discussions).

4. *What are your overall plans for the set of lessons that will be included in the Scoop?*

This may include:

- a description of what the students have been learning until the point when the Scoop begins
- an overview of the lessons you will be teaching during the Scoop (e.g., description of science content, lesson goals/objectives, instructional strategies, student activities, lab activities)

For example,

- We are in the middle of a unit on conservation of mass. This week, we will start out by reviewing the use of some laboratory equipment (such as spring scale, graduated cylinder, etc.). The students will then work through a series of lab activities in which they will “prove” the law of conservation of mass. We will end the week with reading and questions from the textbook.
- This week we are using the process of discovery to work on developing students’ understanding of why earthquakes and volcanoes occur along plate boundaries. We will begin the week by reviewing what students have previously learned about both topics. They will plot recent volcanic and earthquake activity on a map of the world. Then, they will compare these locations with a map of plate boundaries. There will be a written test at the end of the week.



Daily Reflection Questions

To be answered every Scoop day, after the class is over.

Having access to your immediate thoughts and reactions following the lesson is crucial. Please make every effort to jot down your reflections right away after each Scoop class.

1. *What were your objectives/expectations for student learning during this lesson?*

For example,

- My goal for this lesson (and all the lessons during the Scoop) was for students to understand relationships among organisms and their physical environment. The objective of the lesson was for students to use the information they gathered yesterday on an ecosystem or their choice (e.g. ocean, prairie, rainforest, desert, high tundra, deciduous forest) in order to draw a visual representation of the ecosystem. Through this lesson I was checking if they know ways in which organisms interact and depend on one another through food chains and food webs in an ecosystem.
- Today's lesson had two different objectives. During this unit, we will be working on using appropriate tools, technologies and measurement units to gather and organize data. The first objective today was to begin solving problems involving units of measurement and calculating unit conversions. The second objective was for students to develop an understanding of the structure of the universe. Measurement conversion is an objective we have been working on all year. I didn't expect students to get to a "mastery" level for either of these objectives, but rather be at a "novice" level and show some improvement.

2. *Describe the lesson in enough detail so we understand how the Scoop materials were used or generated.*

For example:

- Class started with a full-class review of food chains and food webs. The students have already researched and collected information about an ecosystem of their own choice. Today they used this information to create a visual food chain or food web showing the interrelationships of organisms in the selected ecosystem. I asked them to make sure their drawings included one major producer, consumer, or decomposer in the ecosystem and to organize their food chain or food web so that it shows that each organism is connected to other organisms in the food chain or food web. The lesson will continue tomorrow with each student completing his or her visual and hanging it up on the back bulletin board (see photos).
- At the beginning of class students met for about 20 minutes in their project groups (Driving through the Solar System Project). In groups, students worked together to calculate the amount of time it would take to travel within the solar system if



they were traveling at the speed of an automobile (i.e. 75 mph). Each group was given a different starting point and destination within the solar system (ex. Earth to Mars). They then wrote in their journals about their work on the project (see copy of sample journals). The rest of the class was taken up with sharing each group's results

3. *Thinking back to your original plans for the lesson, were there any changes in how the lesson actually unfolded?*

For example:

- I thought that we would have a chance to explore the differences between the various ecosystems, but I discovered that a number of students in the class didn't have a firm grasp of the relationship between organisms in an ecosystem. So, we backtracked and reviewed those principles through an example of a freshwater ecosystem.
- I didn't originally plan on the students writing in their journals. However, when the students were working in their groups, the discussions were so rich. They were really "talking science," listening, and responding to each other. Since the discussions were so good, I decided I wanted them to individually reflect and write their thoughts in their journals.

4. *How well were your objectives/expectations for student learning met in today's lesson? How do you know?*

For example:

- Based on yesterday's class I assumed that everybody would have a good understanding of the relationship between organisms in an ecosystem. After we discussed the concept again with the freshwater ecosystem example, I think the students showed a greater understanding. I used the students' participation in the class discussion and the development of their visuals to know that they have met the objectives.
- My expectations for group work and unit measurements and conversions were met. Although some groups struggled to cooperate towards the beginning of class and had a hard time getting started, most seemed engaged with the task by the end of the lesson. Their group projects and presentations to the class provided evidence of their understanding of the science concepts and their ability to work in groups. When I read their individual journal entries I was able to see that most students did "get it" but a few are still struggling.

5. *Will today's class session affect your plans for tomorrow (or later in the unit)? If so, how?*

6. *Is there anything else you would like us to know about this lesson that you feel was not captured by the Scoop?*



Post-Scoop Reflection Questions

To be answered at the end of the Scoop timeframe

When answering these questions, please consider the entire set of lessons and all the materials you have gathered for the Scoop notebook.

1. *How does this series of lessons fit in with your long-term goals for this group of students?*
2. *How representative of your typical instruction was this series of lessons (with respect to content, instructional strategies and student activities)? What aspects were typical? What aspects were not typical?*
3. *How well does this collection of artifacts, photographs, and reflections capture what it is like to learn science in your classroom? How “true-to-life” is the picture of your teaching portrayed by the Scoop?*
4. *If you were preparing this notebook to help someone understand your teaching, what else would you want the notebook to include? Why?*

Please refer to the Final Checklist to be sure you have included everything in the notebook.



Daily Calendar

| | DAY #1 | DAY #2 | DAY #3 | DAY #4 | DAY #5 |
|----------------------------------|--------|--------|--------|--------|--------|
| Date | | | | | |
| Length of session | | | | | |
| Topic of session | | | | | |
| Curriculum materials used | | | | | |



| | | | | | |
|--|--|--|--|--|--|
| | | | | | |
|--|--|--|--|--|--|

Photograph Log

| Photo # on camera | DATE | DESCRIPTION |
|-------------------|------|-------------|
| 26 | | |
| 25 | | |
| 24 | | |
| 23 | | |
| 22 | | |
| 21 | | |
| 20 | | |
| 19 | | |
| 18 | | |
| 17 | | |
| 16 | | |
| 15 | | |
| 14 | | |

| | | |
|--|--|--|
| | | |
|--|--|--|

| | | |
|----|--|--|
| 13 | | |
| 12 | | |
| 11 | | |
| 10 | | |
| 9 | | |
| 8 | | |
| 7 | | |
| 6 | | |
| 5 | | |
| 4 | | |
| 3 | | |
| 2 | | |
| 1 | | |