Using Classroom Artifacts to Measure Instructional Practices in Middle School Mathematics: A Two-State Field Test

CSE Report 662

Hilda Borko, University of Colorado
Brian M. Stecher, RAND

December 2005

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

Project 1.1 Comparative Analysis of Current Assessment and Accountability Systems, Strand 1: The Impact of State Accountability on Classroom Practice/Hilda Borko and Brain M. Stecher, Project Directors

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

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

USING CLASSROOM ARTIFACTS TO MEASURE INSTRUCTIONAL PRACTICES IN MIDDLE SCHOOL MATHEMATICS: A TWO-STATE FIELD TEST

Brian M. Stecher, Alice C. Wood, Mary Lou Gilbert
RAND

Hilda Borko, Karin L. Kuffner, Suzanne C. Arnold, Elizabeth H. Dorman
University of Colorado, Boulder

Abstract

The purpose of this research is to determine whether we can use classroom artifacts as the basis for making valid judgments about the presence of reform-oriented teaching practices in middle-school mathematics classes. Our approach compares ratings based on collections of artifacts assembled by teachers according to our directions (the "Scoop Notebook") with judgments based on direct classroom observation of these teachers, direct observation supplemented by artifacts, and transcripts of discourse recorded during classroom observations. Eleven dimensions of reform-oriented practice were identified for use in this investigation, and each was rated on a dimension-specific five-point scale.

Data to answer questions about the reliability and validity of judgments based on the Scoop Notebook are drawn from a field study of 36 middle-school mathematics teachers in two states conducted in Spring 2003. Notebooks were rated independently on each of 11 dimensions by at least three raters who had no prior knowledge of the classroom. In addition, each teacher was observed on two or three occasions during the Scoop period by a researcher who rated each lesson on the same 11 dimensions. At a later time, the observer also reviewed the Scoop Notebook and assigned a "gold standard" rating reflecting all the information available from the Notebook and the classroom observations. For a subset of classrooms, the observed lessons were audiotaped and transcribed, and one researcher with no prior knowledge of the classrooms assigned ratings on the basis of an analysis of the lesson transcripts.

Results indicate that the notebooks could be rated with acceptable reliability and that the notebook scores provided a reasonable estimate of the scores obtained by direct observation and by observation supplemented with the review of artifacts. Notebook scores also differentiated between teachers known to be using reform curricula and those known to be using traditional curricula. However, the reliability and validity were not high enough to justify using the Scoop Notebook for making judgments about individual teachers.

Project Goals and Rationale

Information about classroom practice is central to efforts to improve education, for several reasons. First, teachers play a key role in determining the success of reform efforts. Their actions mediate the impact of educational reforms such as accountability systems, curriculum programs, and instructional approaches, on student achievement. As Fullan & Miles (1992) noted, "local implementation by everyday teachers, principals, parents, and students is the only way that change happens" (p. 752). Spillane (1999) made a similar argument: "While policy makers and reformers at all levels of the system are crucial if these reforms are to be enacted locally, teachers are the key agents when it comes to changing classroom practice. They are the final policy brokers" (p. 144). Thus, information on teachers' classroom practice is key to understanding why programs of reform succeed or fail.

Second, information about classroom practice provides evidence that is relevant to judgments about the validity of test scores and score gains. Better measures of classroom practice can help us to understand what happens under the broad heading of "teaching to the test" and can reveal specific classroom activities that may affect inferences from test scores to the broader domain they are supposed to represent (Borko & Elliott, 1999; Koretz, Stecher, Klein, & McCaffrey, 1994; Stecher, Barron, Chun, & Ross, 2000; Wolf & McIver, 1999).

Third, higher state standards demand more not only of students, but of teachers as well. Many of the standards call for core changes in classroom practices that teachers may not be prepared to incorporate into their classrooms (Firestone, Mayrowetz, & Fairman, 1998). To help teachers develop the capacity to prepare students to meet higher standards, it is important to have reliable measures of classroom practices that can inform improvements in teacher education and professional development programs.

For all these reasons, it is perhaps not surprising that policymakers are calling for more and better measures of instructional practices in schools—measures that will enable researchers and policymakers to capture instruction reliably and efficiently, across a large number of classrooms, over time, without causing an unreasonable burden on teachers, and in a way that can be linked to evidence of student achievement (Brewer & Stasz, 1996; Burstein et al., 1995; Mayer, 1999).

A number of educational researchers are developing new measures of instructional practices in schools as a way of addressing this need (e.g., Aschbacher, 1999; Ball & Rowan, 2004; Camburn & Barnes, 2004; Clare, 2000; Clare & Aschbacher, 2001; Clare, Valdes, Pascal, & Steinberg, 2001; Matsumura, Garnier, Pascal, & Valdes, 2002; Rowan, Camburn, & Correnti, 2004; Rowan, Harrison, & Hayes, 2004). Our project, entitled "The Impact of Accountability Systems on Classroom Practice," is one such effort. The central goal of this five-year research project, funded through the Center for Evaluation, Standards, and Student Testing (CRESST), is to develop an instrument that can provide indicators of reform-oriented instruction across a large number of classrooms without causing an unreasonable burden on teachers and researchers.

Our research investigates the feasibility, reliability, and validity of using artifacts to measure reform-oriented instructional practices. We focus on instructional artifacts because of their potential strength for representing what teachers and students actually do (rather than believe they should do) in the classroom. We consider two subject areas—middle school mathematics and science. In order to develop instruments that are widely applicable to reform-oriented instructional programs in mathematics and science, we identified characteristics of instruction that are broadly endorsed in the reform literature. These characteristics informed the development of guidelines for collecting instructional artifacts and rubrics for scoring the artifact collections.

We used a data collection tool called the "Scoop Notebook" to gather classroom artifacts and teacher reflections related to key features of classroom practice. We conducted pilot studies in five middle school science and eight middle school mathematics classrooms to provide initial information about the reliability, validity, and feasibility of artifact collections as measures of classroom practice (Borko, Stecher, Alonzo, Moncure, & McClam, 2003; Borko, Stecher, Alonzo, Moncure & McClam, 2005). The pilot studies yielded positive results, indicating that the Scoop Notebook and scoring guide have promise for providing accurate representations of what teachers and students do in classrooms, without the expense of classroom observations. Our analyses also suggested that the Scoop Notebook may capture some features of classroom practice more accurately than others, and they provided insights into ways the artifact collection and scoring procedures might be improved.

On the basis of our pilot study results, we made several revisions to the Scoop Notebook prior to using it in this study. These changes were designed to make the notebook and the data collection process more streamlined, straightforward, and easily understood. For example, we revised instructions for collecting three sets of materials—photographs, reflections, and student work—in an attempt to get information that is more detailed and to achieve greater consistency across participants in the materials they provide. We also added two dimensions to the scoring rubrics—to notebook completeness and rater confidence—in order to explore whether differences in the reliability and validity of ratings, by notebook, can be explained by differences in notebook completeness or rater confidence. Finally, we added a more extensive collection and analysis of classroom discourse, to explore whether discourse provides additional insights into instructional practice that are not captured by the Scoop Notebook.

This report presents results of our mathematics validation study, which addressed the following research questions:

1. To what extent do raters agree on the scores that they assign to the Scoop Notebook to characterize the various dimensions of instructional practice?
2. How much agreement is there among the scores assigned by the same observer on different occasions to characterize various dimensions of instructional practice?
3. To what extent do the scores assigned by raters based only on the Scoop Notebook agree with scores assigned by raters based on classroom observations? And to scores assigned based on observations and the Scoop Notebook ("gold standard" ratings)?
4. How do ratings based on transcripts of classroom discourse compare to ratings based on the Scoop Notebook or on gold standard ratings? What additional insights about instructional practices does an analysis of classroom discourse provide?
5. How much variation is there among ratings on the 11 dimensions of reform-oriented practice that were used in rating Scoop Notebooks and classroom observations?

Methods

Overview

Thirty-six middle school mathematics teachers from two states (California and Colorado) participated in the mathematics study in Spring 2003. Each teacher assembled a Scoop Notebook containing artifacts and reflections covering approximately one week of instruction in one class. Members of the research team observed each classroom for two to three days during the time in which the teacher collected artifacts in the Scoop Notebook. Classroom instruction was audiotaped in the classrooms of seven teachers, during the days class was observed.

Notebooks[1] were rated on 11 dimensions of classroom practice, by at least three raters who had no prior knowledge of the classroom (i.e., they had not observed the teacher). The researchers who observed in the classrooms rated each lesson on the same 11 dimensions, and then assigned an overall rating reflecting the two or three classroom observations. At a later time, the rater also reviewed the Scoop Notebook and assigned a "gold standard" rating reflecting all the information available from the notebook and the classroom observations. The "notebook-only" ratings were compared across raters to determine their consistency. The average notebook-only ratings were then compared to the observation ratings and the gold standard ratings to assess their similarity. In addition, the classroom audiotapes were transcribed and were subject to a detailed discourse analysis leading to an independent set of ratings on the same dimensions. These ratings were also compared to the ratings from notebooks and observations.

Participants

We received permission from school districts in the Los Angeles and Denver areas to contact middle school mathematics teachers and solicit volunteers for the study. In most cases, we visited mathematics department meetings to describe the study and recruit participants. In some cases, we recruited by placing one-page flyers in teachers' school mailboxes. Thirty-six teachers from 14 different middle schools in both urban and suburban areas agreed to participate in the study (see Table 1). Each teacher received a $200 honorarium for participating in the study. All teachers' names used in this study are pseudonyms.

[1] The object of study is the notebook, and we use that term in the paper. Since each notebook was created by a different teacher, "notebook" is synonymous with "teacher."

Table 1

Characteristics of Participating Schools in California and Colorado

	California				Colorado									
School	1	2	3	4	1	2	3	4	5	6	7	8	9	10
No. of Teachers	3	3	3	4	5	1	1	2	2	4	1	1	3	3
Location	S	S	U	U	U	U	U	U	U/S	S	S	S	U	U

Note: S = Suburban; U = Urban.

Data Collection: The "Scoop Notebook"

We designed the Scoop Notebook to incorporate a variety of methods for capturing aspects or "artifacts" of classroom practice: samples of student work, lesson plans, photographs, and teachers' responses to reflective questions. We asked teachers to collect artifacts from one of their classes for five consecutive days of instruction. For teachers whose instruction varies from day to day, our pilot study suggested that this period would be a sufficient length of time to capture a range of teaching practices. We specified that the teacher should begin the "Scoop" on a day that was a logical starting point from an instructional perspective (e.g., the beginning of a unit or series of lessons on a single topic), not necessarily the first day of the week. Teachers with block scheduling or other non-traditional scheduling were instructed to "scoop" for an amount of instructional time approximately equivalent to five days on a normal schedule. We asked teachers to select a class comprised of students who were fairly typical of their students and to pick a series of lessons that was fairly typical of instruction in their classroom.

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 in your classroom?" Because we were interested in all types of materials used to foster student learning, we asked teachers to "scoop" materials that they generated, as well as materials drawn from a textbook or other curricular resources. We packaged the Scoop Notebook as a three-ring binder, consisting of the following components:

- project overview

- directions for collecting a “Classroom Scoop”
- folders for assembling artifacts
- sticky notes for labeling artifacts
- calendar for describing “scooped” class sessions
- daily reminders and final checklist
- disposable camera
- photograph log
- consent forms
- pre-scoop, post-scoop, and daily reflection questions

Directions in the notebook asked 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 were encouraged to include any other instructional artifacts not specifically mentioned in the directions. For each instance of student-generated work, teachers were asked to collect examples of “high,” “average,” and “low” quality work. Because we were interested in teachers’ judgments about the quality of student work, we requested that their selections be based on the quality of the work rather than the ability of the students. We also asked that they make an independent selection of student work for each assignment, rather than tracking the same students throughout the artifact collection process.

In addition, the teachers were given disposable cameras and asked to take pictures of the classroom layout and equipment, transitory evidence of instruction (e.g., work written on the board during class), and materials that could not be included in the notebook (e.g., posters and 3-dimensional projects prepared by students). Teachers also kept a photograph log in which they identified each picture taken with the camera.

Each day teachers made an entry in the calendar, giving a brief description of the day’s lesson. Prior to the Scoop period they responded to pre-scoop reflection questions such as, “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?” During the Scoop, teachers answered daily reflection questions such as, “How well were your objectives/expectations for student learning met in today’s lesson?” After the Scoop period, they answered post-scoop reflection questions such as, “How well does this collection of artifacts, photographs, and reflections capture what it is like to learn mathematics in your classroom?" Appendix A provides a complete list of the three sets of reflection questions.

Additional Data Sources: Observations and Discourse

Members of the research team observed each classroom for two to three days in the spring of 2003, during which time the teacher collected artifacts in the Scoop Notebook. In most cases, a single researcher observed each teacher. In two or three cases, multiple researchers observed a given teacher, but usually on different occasions.

We also collected audiotapes of lessons in seven classrooms to explore the feasibility of obtaining classroom discourse data as part of the artifact collection process, as well as to determine what additional information discourse analysis provided. The researchers who observed in these classrooms also audiotaped the lessons. The audiotapes were transcribed to provide a record of classroom discourse.

Scoring the Notebooks, Observations and Discourse

Dimensions of reform practice. We developed a set of 11 dimensions of instructional practice in mathematics to use in analyzing the Scoop Notebook, classroom observations, and discourse. These dimensions, informed by documents such as the *Principles and Standards for School Mathematics* (NCTM, 2000), are listed below:[2]

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. Active teacher role in facilitating groups is not necessary.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities 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.

[2] The definitions of these dimensions were revised several times during our training and calibration activities (described in the next section of the report). These definitions are the final versions, which were used for notebook-only and gold standard ratings. The versions used for observation ratings are in Appendix B. We consider possible implications of using two different versions of the rating guide in the Discussion.

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, and generalization from specific instances to larger concepts or relationships. There are three aspects of cognitive depth: the lesson design, teacher enactment, and student performance. Thus, this dimension considers: extent to which lesson design focuses on central concepts or big ideas; extent to which teacher consistently and effectively promotes student conceptual understanding; and extent to which student performance demonstrates depth of understanding.

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 using the language of mathematics.

7. Explanation and Justification. The extent to which students are expected to explain and justify their reasoning and how they arrived at solutions to problems (both orally and in written assignments). The extent to which students' mathematical explanations and justifications incorporate conceptual, as well as computational and procedural arguments.

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 to support the learning 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.

11. Overall. How well the series of lessons reflect a model of instruction consistent with the NCTM Standards. This dimension takes into account both the curriculum and the instructional practice.

In addition to these 11 dimensions, researchers also rated notebooks on the following two criteria:

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.

A five-point scale was used to rate each dimension. Our scoring guide includes written descriptions and examples for the “high” (5), “medium” (3), and “low” (1) ratings for each dimension. Figure 1 shows the page from the scoring guide corresponding to the Problem Solving dimension.

Problem Solving. Extent to which instructional activities enable students to identify, apply and adapt a variety of strategies to solve problems. Extent to which problems that students solve are complex and allow for multiple solutions. [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?"

Figure 1. Scoring rubric for problem solving dimension

Scoring Procedures

Prior to scoring both the observations and the Scoop Notebooks, the research team engaged in extensive discussions of the scoring rubrics to ensure that all raters had similar understandings of the dimensions and scoring levels, and we revised the rubrics to clarify any discrepancies. To train researchers to use the scoring rubric to rate classroom observations, all raters watched a video of a middle-school mathematics class, rated the class using the dimensions, and then discussed the results for calibration purposes. The discussion was held via conference call, after researchers watched and rated the video independently in California and Colorado.

During the study, each classroom was rated on all 11 dimensions by the researcher (or researchers) who observed in that classroom. They rated each lesson immediately after observing. In addition, at the conclusion of the visits, the observer completed an "overall" summary rating on each dimension based on everything seen during the two or three observations. The overall summary observation ratings were not numerical averages of the individual observation ratings for a given dimension, but separate, qualitative judgments based on the total classroom experience regarding that dimension.

In addition, the researcher who observed in the classroom completed a gold standard rating on each dimension, taking into account both the observational data and the Scoop Notebook.

To train researchers to use the scoring rubric to rate notebooks, all researchers convened for two days in California. All members of the research team independently rated the same notebooks using the rating form. We then met as a group and discussed the results for calibration purposes. After the meeting, the remaining notebooks were assigned to team members, and each notebook was rated on all dimensions by three reviewers, none of whom was familiar with the teacher or the class in which materials were scooped. Notebooks were assigned to raters at random with the exception that observers were not assigned to rate notebooks from classrooms they observed. These notebook-only ratings were based solely on information in the Scoop Notebook.

Finally, for the seven classrooms in which the lessons were audiotaped, one researcher completed a set of ratings based on a discourse analysis of the transcribed lessons, and a second set of ratings taking into account both the discourse data and the Scoop Notebook. The analysis of lesson transcripts was conducted by a researcher with experience in discourse analysis and no prior knowledge of the seven classrooms. As a first step in analysis, she identified and coded information relevant to each dimension of instructional practice on the set of transcripts for each of the seven classrooms. For example, an interchange in which the teacher asked students to explain their reasoning was coded as “Explanation and Justification;” when student-to-student communication occurred, she coded the interchange as “Mathematical Discourse Community.” In addition, she coded the transcripts for univocal and dialogic discourse patterns, because of the relevance of these patterns to several of the dimensions (Lotman, 1988; Scott, 1998; Wertsch, 1991; Wertsch & Toma, 1995). Types of discourse such as fill-in-the-blank and discrete-answer questions, Initiation-Reply-Evaluation (IRE; Mehan, 1979) sequences, and direct explanations were coded as univocal. Types of discourse such as asking for explanations or justifications, posing questions that required higher-order thinking skills, and challenging another person’s thinking were coded as dialogic.

After coding the transcripts, the researcher listened to the audiotapes to confirm the codes. She then completed two different sets of ratings. First, she rated each classroom on the 11 dimensions of instructional practice, based on the discourse analyses only. Second, she rated each classroom based on the discourse analyses and the contents of the Scoop Notebook. She conducted several comparisons between these ratings and notebook-only and gold standard ratings. She used these comparisons and the qualitative analysis of discourse patterns to determine the additional insights about instructional practices that classroom discourse information can provide.

Results and Discussion

We conducted several different analyses to assess the accuracy of ratings, and to compare the ratings based on different sources. Because we had incomplete data from a small number of classrooms, these analyses were conducted using data from 30 classrooms.

Range of Notebook Ratings

All dimensions were rated on a five-point scale, but each scale was defined in terms of dimension-specific criteria. As a result, a score of 3 on one dimension should not be interpreted to be equivalent to a score of 3 on another dimension. Table 2 shows the mean and standard deviation of the average of the three ratings assigned to each of the 30 notebooks. The overall average rating was about 3 with a standard deviation just under 1, suggesting that the bulk of the ratings were concentrated in the middle of the five-point scales, but the full scales were used in many cases. On average, raters judged the notebooks to be closest to the highest score point on the dimension Structure of

Lessons (mean rating 4.23), and there was the least variation for this dimension, as a result. Notebooks achieved the lowest rating, on average, on the Connections/Application dimension (mean rating 2.61), but there was considerable variation across notebooks on this dimension.

Table 2

Mean and Standard Deviation of Average of Notebook Ratings, by Dimension

Dimension	Mean	Standard Deviation
Assessment	3.27	0.63
Cognitive Depth	3.01	1.03
Connections/Applications	2.61	1.06
Discourse Community	2.82	1.01
Explanation & Justification	2.74	1.02
Grouping	3.25	1.32
Mathematical Tools	2.87	0.89
Multiple Representations	3.26	0.77
Problem Solving	3.07	0.89
Structure of Lessons	4.23	0.58
Overall	3.05	0.86

Accuracy of Notebook Ratings

Because notebooks were rated by multiple individuals, we can obtain direct estimates of the accuracy of the notebook rating process. We can also isolate the sources of inaccuracies—whether individual raters were more or less lenient overall, or whether there was a more complex interaction between raters and notebooks (i.e., some raters were higher than others on one group of notebooks and lower than others on another group).

As a first step in these analyses, we computed the percent of exact agreement among raters on each dimension of each notebook and the percentage of agreement within one scale point.[3] We then averaged the percentages for each dimension across all notebooks to yield an overall indication of the accuracy of dimension-level ratings (see

[3] Values for each notebook were based on ratings by three raters, except Coyner and Mason, which were based on seven raters. Agreement within one was computed by looking at all scores from all possible pairs of raters and computing the percentage of these pairs for which the two ratings were the same or adjacent.

Table 3). Exact agreement ranged from 21% on Mathematical Tools to 44% on Connections/Applications. Using a more relaxed criterion, over 70% of ratings agreed within one scale point on every dimension, and over 80% were within one point on six dimensions.

Table 3

Percent Agreement in Notebook Ratings, by Dimension

Dimension	% Exact Agreement	% Within 1
Assessment	41.2	76.1
Cognitive Depth	25.2	74.3
Connections/Applications	44.3	76.6
Discourse Community	31.4	77.2
Explanation & Justification	41.2	80.1
Grouping	39.4	82.7
Mathematical Tools	21.1	71.1
Multiple Representations	27.4	82.3
Problem Solving	37.2	81.7
Structure of Lessons	36.8	87.8
Overall	37.1	88.9

Similarly, for each classroom, the agreement percentages for each dimension were averaged across all dimensions to yield an overall indication of the accuracy of notebook-level ratings (see Table 4). The results were similar to those reported for dimension-level averages, although there was more variability across classrooms than across dimensions. Exact agreement ranged from 12% for Bondarenko to 54% for Logan; agreement within one point ranged from 61% for Martin to 97% for Fischer.

Table 4

Percent Agreement in Notebook Ratings, by Classroom

Notebook	% Exact Agreement	% Within 1
Alschuler	33.2	84.9
Bondarenko	12.0	66.7
Carter	30.1	81.8
Coyner	51.1	95.2
D'Amico	27.1	75.7
Fischer	42.2	97.0
Foley	27.0	87.9
Foster	39.2	94.0
Gibson	27.1	66.6
Hall	12.1	73.0
Kirkwood	21.1	63.7
Klein	42.3	69.6
Kretke	36.2	84.9
Lewis	30.2	85.0
Loeb	39.2	81.8
Logan	54.4	90.9
Lowe	36.2	84.9
Martin	24.1	60.5
Mason	39.4	79.7
Matsumura	36.2	81.8
Merrow	36.2	78.8
Peterson	27.0	81.8
Price	21.0	78.8
Reynolds	45.2	87.8
Saliba	54.5	84.9
Shephard	42.2	90.9
Sleeve	27.1	57.5
Sze	60.5	90.9
Wirtz	36.2	84.9
Zinc	33.2	84.9

It is difficult to say, at first glance, whether the levels of agreement reported in Tables 3 and 4 should be considered low or high. Clearly, none of the values represent perfect agreement (100%). Another standard that can be used for judging the quality of these results is agreement by chance. What percent exact agreement and agreement within one point would be achieved if ratings on the five-point scale were assigned at random? If three raters had assigned ratings on a five-point scale at random, then the corresponding predicted value for the percent of exact agreement would be 4%, and the corresponding predicted value for agreement within one point would be 20%. Thus, agreement among raters, while far from perfect, is also far greater than if left to chance.

Generalizability theory offers a more sophisticated way to judge the accuracy of the notebook ratings. For each dimension, analysis of variance is used to estimate the percent of variance that is attributable to notebooks (i.e., teachers), raters, and residual error (including the interaction between notebooks and raters). These variance component estimates are used to compute a generalizability coefficient, which can be interpreted directly as a measure of accuracy. In addition, the variance component estimates can also be used as a design tool to predict the accuracy of ratings that would be obtained in future investigations using different numbers of raters.

For the purpose of this analysis, we treated the 11 dimensions as a fixed facet and, for each dimension, computed variance components for raters, notebooks and residual using SAS Proc VARCOMP. The results of these analyses are shown in Table 5.

Table 5

Estimated Scoop Rating Variance Components, by Dimension

Dimension	Notebook	Rater	Residual
Assessment	.184	.020	.579
Cognitive Depth	.750	.119	.681
Connections	.900	.120	.474
Discourse	.729	.224	.519
Explanation & Justification	.786	.131	.488
Grouping	1.55	.031	.564
Mathematical Tools	.599	.328	.714
Multiple Representations	.458	.103	.506
Problem Solving	.550	.176	.602
Structure of Lessons	.187	.024	.439
Overall	.546	.071	.404

The results are easier to interpret if they are converted to percentages of total variance. Table 6 shows the percent of variance attributable to each source. In general, these results are encouraging. Most variance was attributable to differences between notebooks, not to differences between raters or to error. Furthermore, raters accounted for the smallest percent of variance for all dimensions. It was generally not the case that one rater was consistently more strict or lenient than another. However, Table 6 also shows large differences among dimensions in the distribution of variance. Some dimensions, such as Connections and Grouping, were rated very consistently, i.e., 60-70% of the variance was due to differences between notebooks, while only 20-30% was unexplained. Other dimensions, such as Assessment and Structure of Lessons, were not rated very consistently, i.e., more than one-half of the variance is unexplained. This unexplained residual variance includes both unsystematic differences (error) and interactions between raters and notebooks, i.e., a situation where Rater A thought more highly of Notebook 1 than Rater B, but Rater B thought more highly of Notebook 2 than Rater A.

Table 6

Percentage of Scoop Rating Variance Attributed to Each Component

Dimension	Notebook	Rater	Residual
Assessment	23.5%	2.6%	73.9%
Cognitive Depth	48.4%	7.7%	43.9%
Connections	60.2%	8.0%	31.7%
Discourse	49.5%	15.2%	35.3%
Explanation & Justification	55.9%	9.3%	34.7%
Grouping	72.3%	1.4%	26.3%
Mathematical Tools	36.5%	20.0%	43.5%
Multiple Representations	42.9%	9.7%	47.4%
Problem Solving	41.4%	13.3%	45.3%
Structure of Lessons	28.8%	3.7%	67.5%
Overall	53.5%	7.0%	39.6%

A simpler way to judge the quality of the ratings is to use the variance components to estimate a generalizability coefficient (which can be interpreted like a reliability coefficient). We can estimate the level of generalizability that would be attained in a future study if we were to use two, three or four raters. In addition, we can predict the generalizability for absolute decisions (i.e., assigning a specific score on the five-point scale) and for relative decisions (i.e., ranking teachers from low to high on the dimension). These results are shown in Table 7. Using three raters, we could achieve a generalizability coefficient above 0.7 for absolute decisions and 0.8 for relative decisions on most of the dimensions.[4] The three dimensions that clearly would not be rated with acceptable generalizability are Assessment, Structure of Lessons, and Mathematical Tools.

[4] The differences in generalizability coefficients between absolute and relative decisions are small because the variance component for raters is small relative to the other components.

Table 7

Generalizability Coefficients for Scoop Rating for Absolute and Relative Decisions Using Two, Three or Four Raters, by Dimension

	Absolute Decisions			Relative Decisions		
Dimension	Two Raters	Three Raters	Four Raters	Two Raters	Three Raters	Four Raters
Assessment	0.38	0.48	0.55	0.39	0.49	0.49
Cognitive Depth	0.65	0.74	0.79	0.69	0.77	0.77
Connections	0.75	0.82	0.86	0.79	0.85	0.85
Discourse	0.66	0.75	0.80	0.74	0.81	0.81
Explanation & Justification	0.72	0.79	0.84	0.76	0.83	0.83
Grouping	0.84	0.89	0.91	0.85	0.89	0.89
Mathematical Tools	0.53	0.63	0.70	0.63	0.72	0.72
Multiple Representations	0.60	0.69	0.75	0.64	0.73	0.73
Problem Solving	0.59	0.68	0.74	0.65	0.73	0.73
Structure of Lessons	0.45	0.55	0.62	0.46	0.56	0.56
Overall	0.70	0.78	0.82	0.73	0.80	0.80

Although generalizability was low for these three dimensions, the values appear to reflect high levels of agreement, not high levels of disagreement. This apparent paradox occurs because there was very little variation in scores on these dimensions. For example, the average rating for Structure of Lessons was 4.23, and the standard deviation was only 0.58. This means that most notebooks were rated at the top of the scale. There was less variability overall, and the resulting generalizability coefficient was artificially low. Indeed, the agreement indicators in Table 3 show that raters were as consistent in rating Structure of Lessons as with any other dimension.

Similarly, there is very little variation in scores for Assessment. The standard deviation of average Assessment ratings in Table 2 is only 0.63, the second lowest of any dimension. In this case, the average score was not at the top of the scale, but ratings were concentrated tightly around the middle of the scale. Again, measures of agreement show that raters were fairly consistent in assigning scores for this dimension, as well.

The story is different for Mathematical Tools. The low generalizability coefficient is not an artifact of highly concentrated scores. Instead, it appears to be the case that raters were not consistent in applying the scoring guide. This situation may have occurred because the dimension was defined in terms of two components—opportunity to use tools and enabling students to represent abstract ideas. Raters may have found it difficult to weight these distinct elements when assigning a rating to the notebooks.

Accuracy of Classroom Observation Ratings

We cannot estimate the accuracy of observation ratings as well as we estimated the accuracy of notebook ratings because we did not have multiple observers in each classroom. Each classroom was observed on two or three occasions during the Scoop period by one of eight observers, and each observer attended between seven and 15 lessons. Only four classrooms were observed by more than one observer, and this occurred because of scheduling difficulties. As a result, we cannot estimate the inaccuracy in observation ratings due to differences between individual raters, which is potentially the largest source of error. (We plan to address this limitation in future studies of the Scoop Notebook.)

Since each classroom was observed on multiple occasions, we can examine the consistency of ratings for a given classroom over time. For this analysis, occasion was defined sequentially, i.e., as the first, second or third observation of a given classroom. Thus, all first observations are treated as comparable in terms of the occasion facet, although first observations occurred on different dates in each classroom. The same is true of second and third observations. As a result, a significant effect associated with occasion would indicate differences from one day to another within a classroom, but not differences associated with a particular date in the year or lesson in the curriculum. As in the case of notebook ratings, we conducted the analyses separately for each of the 11 dimensions. Analyses of variance were conducted with classroom, rater, and occasion as factors. There were no significant effects associated with the occasion factor in any of the 11 analyses. Thus, classroom observation ratings did not vary systematically over time on any dimension. This finding does not indicate that there were no changes in practice from day to day, just that day-to-day changes were not consistent from classroom to classroom. This result is not surprising since we would not expect to find similar day-to-day variations in practice across classrooms. The lack of occasion effects does not provide much new insight into the accuracy of classroom observations.

Comparing Notebooks and Observations

Next, we examined the correspondence between notebook ratings and observation ratings to determine how well the notebooks could substitute for direct classroom observation. For each notebook, we computed the average rating across all raters on each dimension. The generalizability analyses reported above suggest that these average scores, which were based on three readers in most cases, should be reliable for most dimensions. These notebook-based ratings on each dimension were compared to the summary observation ratings obtained for each teacher.

For each dimension, we computed a Pearson correlation between the average notebook rating and the summary observation rating. The analyses were conducted using 27 teachers because we did not have summary observation ratings for three teachers. Table 8 shows the dimension-level correlations between ratings based on notebooks and classroom observations. For most dimensions, the correlations were 0.7 or higher, suggesting that the notebooks rank teachers in roughly the same order as observations. Assessment and Structure of Lessons had the lowest correlations, as would be expected based on the low generalizability of the notebook ratings on these two dimensions.

The low correlation between notebook and observation ratings for Assessment might also be due to the fact that Assessment was defined to include both formal and informal measures, and these two types of measures were not equally easy to identify when observing in a classroom and when reviewing the notebook. For example, when reviewing the notebooks it was quite easy to determine whether the teacher used formal assessments—there were direct references to quizzes, and tests and copies were often provided by the teachers—but difficult to make judgments about informal assessment. Observers faced just the opposite problem—informal assessment was quite apparent in the questions the teacher asked and the manner in which he or she monitored student work, but on any given day there might not be an actual test or quiz. This difference may also have contributed to difference in ratings for the Assessment dimension.

The low correlation between notebook and observation ratings for Structure of Lessons might also reflect the fact that we revised the description of this dimension between the scoring of observations and notebooks, based on our calibration activities. Specifically, we omitted the criterion that activities in the series of lessons lead toward deeper conceptual understanding. This difference may also have contributed to the low correspondence between notebook ratings and observation ratings.

Table 8
Correlation Between Average Scoop Rating and Summary Observation Rating, by Dimension

Dimension	Correlation
Assessment	0.37
Cognitive Depth	0.70
Connections	0.65
Discourse Community	0.69
Explanation & Justification	0.75
Grouping	0.86
Mathematical Tools	0.70
Multiple Representations	0.78
Problem Solving	0.66
Structure of Lessons	0.38
Overall	0.78

We also computed two summary correlations to indicate the overall correspondence between notebook ratings and observation ratings. First, we considered each of the 11 dimensions for each of the 27 teachers as a separate piece of information (297 in all). We computed the correlation between the 297 notebook ratings and 297 observation ratings. This correlation was 0.65. Second, we computed the average notebook rating and average observation rating for each classroom by averaging across the 11 dimensions. The correlation between the overall classroom average based on notebooks and the overall classroom average based on observations was 0.85. This second correlation is probably the more meaningful one to use as an indicator of the extent to which the artifacts portray the reform-oriented practices of individual teachers. The high value suggests that combined judgments of reform-oriented practice based on Scoop Notebooks are similar to combined judgments based on classroom observations.

Comparing Notebooks and Gold Standard Ratings

To investigate the second component of the third research question, we compared ratings assigned to the classroom on the basis of the Scoop Notebook to gold standard ratings, which were assigned on the basis of direct classroom observation supplemented with the information contained in the notebook. Table 9 shows the correlations between average notebook ratings and gold standard ratings on the 11 dimensions. For ten of the dimensions, the correlation between average notebook ratings and gold standard ratings is higher than the correlation between average notebook ratings and overall observation ratings reported in Table 8. This pattern makes sense because the gold standard ratings were based on the observations as well as the additional information contained in the Scoop Notebook.

Table 9
Correlation Between Average Notebook Rating and Gold Standard Rating, by Dimension

Dimension	Correlation
Assessment	0.43
Cognitive Depth	0.81
Connections	0.78
Discourse Community	0.79
Explanation & Justification	0.80
Grouping	0.87
Mathematical Tools	0.82
Multiple Representations	0.70
Problem Solving	0.74
Structure of Lessons	0.45
Overall	0.81

In addition, we computed two summary correlations to measure the overall match between average notebook scores and gold standard scores. The correlation based on 297 separate pieces of information (i.e., taking each dimension for each classroom separately) was 0.77. A comparison of the average gold standard score for each classroom across the 11 dimensions with the average notebook score for that classroom yielded a correlation of 0.89. The high correlation indicates that when summarized at the classroom level, notebook scores rank teachers almost the same as gold standard scores. Thus, combined judgments of a teacher's reform oriented practice based on the Scoop Notebook are similar to combined judgments based on the notebook plus classroom observations.

Notebook ratings and gold standard ratings were also compared in terms of absolute agreement. Two measures of agreement were computed—the percentage of ratings that were within 0.33 units on the five-point rating scale and the percentage of ratings that were within 0.67 units on the five-point rating scale. There was moderate agreement between the gold standard ratings and the average of the notebook ratings. Considering the complete set of 330 possible rating comparisons (30 notebooks scored on 11 dimensions), the difference between average notebook-only ratings and gold standard ratings was within 0.33 points for 130 comparisons (39%) and within 0.67 points for 193 comparisons (58%). If we relax the standard to 1.0 rating point, which is the criterion used in the earlier reliability analysis, the number of close comparisons increases to 265 (80%). It seems reasonable to characterize these results as demonstrating moderate agreement in absolute terms between the two methods for describing classroom practice.

Analyses of Classroom Discourse

This section presents analyses conducted to address the fourth research question. We compare ratings based on transcripts of classroom discourse to notebook-only ratings and gold standard ratings in seven classrooms. We then consider what insights can be gained by taking into account both classroom discourse and information in the Scoop Notebook.

Comparing discourse-only and notebook-only ratings. We compared ratings based on the discourse analyses alone to the average notebook-only ratings for each dimension, for each of the seven classrooms in which discourse was audiotaped. Three levels of agreement were calculated—within 0.33, within 0.67, and within 1.0 unit on the five-point rating scale. Percent of agreement at each of these three levels was then calculated for each dimension of instructional practice, across all seven classrooms. Table 10 presents a summary of these comparisons.

Table 10

Discourse-Only Ratings, Average Notebook-Only Ratings, and Percent Agreement, by Dimension (Averaged Across Seven Classrooms)

Dimension	Discourse Only Rating	Average NB Only Rating	% Within 0.33	% Within 0.67	% Within 1.0
Assessment	2.71	3.43	57.1	57.1	85.7
Cognitive Depth	3.29	2.71	42.8	57.1	71.4
Connections/Applications	3.00	2.38	28.5	57.1	71.4
Discourse Comm.	3.29	2.67	28.5	42.8	100.0
Explanation & Justification	3.29	2.48	42.8	42.8	57.1
Grouping	2.29	3.14	0.00	42.8	57.1
Mathematical Tools	2.29	2.57	71.4	71.4	85.7
Multiple Representations	3.14	3.09	28.5	28.5	85.7
Problem Solving	2.71	2.72	42.8	57.1	71.4
Structure of Lessons	4.71	4.38	42.8	57.1	57.1
Overall	3.00	2.86	57.1	57.1	100.0
TOTAL AVERAGES	3.07	2.95	40.2	51.9	76.6

Considering the complete set of 77 possible comparisons (seven notebooks scored on 11 dimensions), the difference between average notebook-only ratings and discourse-only ratings was within 0.33 points for 31 comparisons (40.2%) and within 0.67 points for 40 comparisons (51.9%). If we relax the standard to 1.0 scale point, the number of matches increases to 61 (79.2%). We interpret these results as evidence of moderate agreement between the two methods for rating classroom practice, i.e., analysis of discourse and analysis of the Scoop Notebook content.

If the Scoop Notebooks and the audiotapes/transcripts alone accurately reflect teaching practice in similar ways, then we would expect there to be a high correlation between notebook-only ratings and discourse-only ratings. Across all dimensions, the correlation between the average discourse-only ratings and the average notebook-only ratings was 0.61. This moderate correlation provides additional evidence that there are some differences between raters' judgments about instructional practices based on the contents of the Scoop Notebook and judgments based on an analysis of transcripts of classroom lessons.

Comparing discourse-plus-notebook ratings and gold standard ratings. Next, we examined the correspondence between ratings based on discourse-plus-notebooks and ratings based on observations-plus-notebooks (i.e., gold standard ratings) for each dimension of instructional practice, across the 7 teachers. Because there was only one discourse-plus-notebook rater and one gold standard rater for each classroom, we computed the percent of exact agreement between raters on each dimension and the percent of agreement within one scale point. The percentages for each dimension were averaged to yield an overall indication of correspondence of dimension-level ratings (see Table 11).

Exact agreement occurred in 45.4% of the comparisons, with a range from 14.3% (Grouping) to 71.4% (Structure of Lessons). Agreement within one scale point occurred in 92.2% of the comparisons, with the two raters agreeing within one point for all seven classrooms (100% agreement) on seven of the 11 dimensions. These relatively high levels of agreement between gold standard ratings and discourse-plus-notebook ratings suggest that judgments about instruction made on the basis of the artifacts in the Scoop Notebook and observational data are similar to judgments formed by listening to audiotapes of classroom interaction, studying transcripts, and reviewing the Scoop Notebook.

If judgments based on the notebook and classroom discourse are similar to judgments based on the notebook and classroom observations, then we would expect there to be a high correlation between these two sets of ratings. Across all dimensions, the correlation between the average discourse-plus-notebook ratings and the average gold standard ratings was 0.96. This high correlation provides additional evidence that these two combinations of data provide similar representations of instructional practices.

Table 11

Average Ratings and Percent Agreement for Discourse + Notebook Ratings and Gold Standard Ratings, by Dimension

Dimension	Discourse + Notebook Rating	Gold Standard Rating	% Exact Agreement	% Agreement Within 1.0
Assessment	3.57	3.86	42.8	100.0
Cognitive Depth	3.29	3.14	28.5	100.0
Connections/Applications	3.00	2.71	57.1	85.7
Discourse Comm.	3.29	2.86	28.5	100.0
Explanation & Justification	3.29	3.14	57.1	100.0
Grouping	3.14	2.89	14.3	71.4
Mathematical Tools	2.29	2.14	28.5	71.4
Multiple Representations	3.43	3.14	57.1	85.7
Problem Solving	2.89	2.71	57.1	100.0
Structure of Lessons	4.71	4.71	71.4	100.0
Overall	3.14	3.29	57.1	100.0
AVERAGE	3.28	3.14	45.4	92.2

Additional insights provided by classroom discourse analyses. Discourse analyses, together with reflections recorded by the researchers as they conducted the various ratings (notebook-only, gold standard, discourse-only, discourse-plus-notebook) indicate that the analysis of classroom discourse provided additional insights regarding five dimensions: Mathematical Discourse Community, Explanation and Justification, Cognitive Depth, Connections/Applications, and Assessment. For these dimensions the transcripts provided information that was not present in the notebooks (and thus there was lower agreement between discourse-only and notebook-only ratings). Consequently, there may be a benefit to having access to transcripts of classroom discourse, in addition to the contents of the Scoop Notebook, when rating these five dimensions of instructional practice.

Notebook raters consistently reported that Mathematical Discourse was the most difficult dimension to rate on the basis of evidence in the Scoop Notebooks. Analysis of the transcripts and audiotapes provided much more detail than the collection of notebook artifacts in three main areas. First, student-to-student communication was revealed more naturally and thoroughly in tapes and transcripts than in artifacts. Despite the fact that the audiotaped recordings did not capture all that the students were saying, the content or nature of interchanges could often be inferred from muffled comments and from the response of the teacher or of other students. Also, the tapes did clearly reveal when the teacher was speaking versus when the students were speaking; it was therefore possible to get a sense of how teacher-directed or how student-centered a particular classroom was. Second, the tapes and transcripts demonstrated the ways that the teacher solicited, explored, and attended to student thinking, something that is very difficult to capture in the notebook. Third, the tapes and transcripts enabled the researcher to identify the ways in which the teacher modeled use of appropriate mathematical language and encouraged such language use by students.

When scoring the Explanation and Justification dimension, raters were asked to evaluate the quality and depth of students' explanations, as well as how often teachers expect students to provide such explanations and justifications. The Scoop Notebook was limited to print-based assignments (e.g., worksheets), samples of written explanations and justifications by students, and teachers' reflections on the lessons and student work. Thus, it contained no direct evidence of oral explanations and justifications. In contrast, the transcripts and audiotapes provided direct evidence about the quality and depth of students' oral explanations and justifications. In addition, an analysis of the teacher's questions provided information about the teacher's expectations for students to explain and justify their answers.

Raters were asked to evaluate three aspects of cognitive depth: the lesson design, teacher enactment, and student performance. Through lesson plans, instructional materials, samples of student work, and teacher reflections, the notebook artifacts often adequately demonstrated the extent to which lesson design focuses on central concepts or big ideas. However, transcripts and audiotapes of classroom discourse provided much more information than the notebook regarding the extent to which the teacher consistently and effectively promotes student conceptual understanding (rather than focusing on the correctness of answers). For example, analyses of these data sources addressed the extent to which a teacher's questions were univocal and convergent, thus affording very limited opportunities for students to elaborate on their thinking or provide explanations that demonstrated the depth of their understanding. They also revealed contrasting cases, in which teachers asked questions that were dialogic and divergent, and students had many more opportunities to demonstrate their understanding by explaining their thinking and providing evidence to support their conjectures. Transcripts and audiotapes also provided access to what the students were saying, the kinds of questions they asked, and how they responded to various teacher questions, thus making it easier to judge the level of student conceptual understanding.

The transcripts and audiotapes also contained information to supplement the collection of artifacts when rating the Connections/Applications dimension. The researcher conducting the discourse analyses noted that although teachers often provided rich and relevant examples orally during class discussion or lecture—examples that help students apply mathematics to real world contexts—they sometimes neglected to write about these oral examples when completing their reflections. Also, students sometimes made oral contributions about connections between mathematics and their own experiences that were not captured in the notebook.

The transcripts and audiotapes provided much more detail about the teacher's use of informal, formative assessment strategies (especially oral questioning) than was contained within the collection of notebook artifacts. These data sources also provided valuable information about the teacher's oral feedback to students. Analysis of the discourse, however, does not provide a complete picture of the range of assessment strategies used in a classroom; a more complete picture of a teacher's assessment strategies is provided by considering classroom discourse in conjunction with the contents of the Scoop Notebook.

Dimensionality

The 11 dimensions used in this research represent aspects of practice that are important from a conceptual and a pedagogical point of view, but they may not all be distinct from a statistical perspective. To answer the fifth research question, we used correlations and factor analyses to examine the extent to which the dimensions were providing unique information when used for rating the notebooks and when used for rating classroom observations.

Eight different individuals rated the Scoop Notebooks, with each person rating between 10 and 12 notebooks. In total, there were 98 complete notebook ratings. Table 12 shows the correlations among the 11 dimensions based on these 98 notebook ratings. More than one-half of the off-diagonal correlations (28 of 55) are above 0.5, and seven are above 0.7, which indicates considerable overlap among the dimensions.

Table 12

Correlation among Dimensions, Based on Individual Notebook Ratings

	Assess	Cog Depth	Con-nect	Dis-course	Ex-plain	Group-ing	Multi Rep	Math Tools	Prob Solv	Struc -ture	Over -all
Assess	1.00										
Cog Depth	0.30	1.00									
Connect	0.43	0.21	1.00								
Discourse	0.48	0.09	0.60	1.00							
Explain	0.64	0.34	0.68	0.50	1.00						
Grouping	0.66	0.26	0.44	0.30	0.71	1.00					
Multie Rep	0.67	0.20	0.58	0.38	0.78	0.74	1.00				
Math Tool	0.68	0.40	0.68	0.51	0.84	0.73	0.82	1.00			
Prob Solv	0.38	0.28	0.38	0.29	0.45	0.42	0.48	0.61	1.00		
Structure	0.61	0.15	0.46	0.43	0.54	0.54	0.54	0.60	0.24	1.00	
Overall	0.56	0.18	0.56	0.48	0.69	0.51	0.66	0.76	0.41	0.49	1.00

We conducted a factor analysis of the same data set and found that a single factor explained 56% of the variance. Table 13 contains the factor loadings for each dimension on this dominant factor, showing loadings above 0.7 for all dimensions except Structure of Lessons, Cognitive Depth, Discourse and Problem Solving. In a four-factor solution, Structure of Lessons, Cognitive Depth, and Discourse each has its highest loading on one of the additional factors. This pattern suggests that Structure of Lessons, Cognitive Depth, Discourse and Problem Solving may be capturing features of practice that are distinctive from the other dimensions, although they are not distinctive enough as we have defined them to stand entirely on their own. We repeated both analyses using the average of each notebook's ratings rather than the separate ratings from each researcher, and the results were essentially the same. Overall, these results indicate that the 11 dimensions seem to be capturing a single underlying factor associated with generalized reform-oriented practice. They may also be capturing something distinctive about Structure of Lessons, Cognitive Depth, Discourse and Problem Solving, but we do not have enough evidence to confirm this.

Table 13

Factor Loadings for Individual Notebook Rating Dimensions

Dimension	Factor Loading
Assessment	0.77
Cognitive Depth	0.34
Connections	0.71
Discourse	0.58
Explanation & Justification	0.89
Grouping	0.77
Mathematical Tools	0.95
Multiple Representations	0.86
Problem Solving	0.55
Structure of Lessons	0.65
Overall	0.76

We conducted a similar analysis of dimensionality using the 79 classroom observations conducted by eight observers. Table 14 shows the correlations among the dimensions based on the classroom observations. Forty-eight of the 55 off-diagonal correlations are above 0.5, and 17 are above 0.7. As in the case of the notebook ratings, a factor analysis yielded a single dominant factor, which explained 67% of the variance.

Table 14

Correlation among Dimensions, Based on Individual Observation Ratings

	Assess	Cog Depth	Con-nect	Dis-course	Ex-plain	Grou-ping	Multi Rep	Math Tools	Prob Solv	Struc-ture	Over-all
Assess	1.00										
Cog Depth	0.56	1.00									
Connect	0.51	0.69	1.00								
Discourse	0.56	0.65	0.62	1.00							
Explain	0.53	0.85	0.76	0.59	1.00						
Grouping	0.53	0.70	0.58	0.53	0.67	1.00					
Multie Rep	0.60	0.80	0.68	0.55	0.78	0.79	1.00				
Math Tool	0.64	0.88	0.76	0.67	0.87	0.79	0.86	1.00			
Prob Solv	0.39	0.71	0.63	0.45	0.69	0.65	0.70	0.71	1.00		
Structure	0.43	0.51	0.44	0.26	0.52	0.42	0.54	0.59	0.43	1.00	
Overall	0.58	0.74	0.69	0.62	0.74	0.65	0.70	0.84	0.60	0.57	1.00

The factor loadings on the dominant factor are shown in Table 15. All dimensions except Structure of Lessons, Discourse and Assessment load very highly on this single factor. In a four factor solution, each of these dimensions has its highest loading on one of the other factors. We repeated the analyses using the average of the ratings on each classroom, and the results were essentially the same. This pattern confirms findings based on the notebooks—a single factor (reform-orientation practice) is enough to capture the bulk of the variance in observations from the 11 dimensions, but a few other distinctive aspects of classroom practice may be present. Structure of Lessons and Discourse stand out in both the notebook and observational analyses.

Table 15

Factor Loadings for Individual Observation Rating Dimensions

Dimension	Factor Loading
Assessment	0.66
Cognitive Depth	0.90
Connections	0.80
Discourse	0.69
Explanation & Justification	0.89
Grouping	0.80
Mathematical Tools	0.97
Multiple Representations	0.90
Problem Solving	0.75
Structure of Lessons	0.58
Overall	0.85

These findings suggest that it would be possible to achieve a similar degree of correspondence with far fewer dimensions. However, the elimination or combining of dimensions might have negative consequences for other reasons, e.g., understanding specific features of classroom practice for professional development purposes.

Comparing Colorado and California

An additional analysis compared the ratings of California notebooks and Colorado notebooks. We knew from independent information that the mathematics textbooks and curricula used by the Colorado teachers embodied more of the reform principles than the textbooks and curricula used by the California teachers. If the notebooks are capturing the teachers' instructional approaches accurately, the state-to-state difference should be reflected in the pattern of ratings assigned to the Colorado and California notebooks. Consistently higher ratings for Colorado notebooks, particularly on the dimensions most closely associated with reform curricula, would be further evidence of the validity of the artifact notebooks. (As noted above, notebooks were assigned to raters at random, so each rater read some notebooks from California and some from Colorado.) As a preliminary step we compared rater agreement between the two states and found it was similar. In California, exact agreement among raters and agreement within one point were 38.6% and 82.6%, respectively. In Colorado, exact agreement and agreement within one point were 32.8% and 80.1%, respectively. Thus, the notebooks from the two states were rated with roughly equal reliability.

Table 16 shows the average notebook-only and gold standard ratings for each notebook by state. In general, Colorado notebooks had higher ratings (3.42) than California notebooks (2.48). Colorado classrooms also had higher gold standard ratings (3.47) than California classrooms (2.30). Furthermore, only two of the California classrooms had average notebook-only ratings *above* 3.0 while only four of the Colorado classrooms had average notebook-only ratings *below* 3.0. The gold standard ratings tell a similar story. Colorado classrooms were rated higher than California classrooms by a margin of about one point on the five-point rating scale. These results are consistent with our knowledge of the curriculum materials used in the two sets of classrooms.

Table 16

Average Notebook and Gold Standard Ratings, by Classroom: Colorado and California Teachers

Notebook	Average	Gold Standard
Colorado Teachers		
Carter	3.42	3.91
Fischer	2.85	3.18
Foley	3.30	1.64
Foster	4.21	5.00
Gibson	3.21	2.36
Hall	3.70	4.00
Kirkwood	3.88	4.82
Klein	3.15	4.09
Kretke	4.15	4.00
Lewis	3.12	3.09
Logan	1.91	1.27
Lowe	3.09	2.91
Martin	3.88	4.73
Mason*	3.36	3.36
Peterson	2.82	1.82
Price	3.42	3.82
Reynolds	4.12	4.55
Shephard	4.06	4.36
Sleeve	3.94	3.36
Zinc	2.94	3.18
CO AVERAGE	3.42	3.47
California Teachers		
Alschuler	2.06	2.55
Bondarenko	3.09	2.91
Coyner*	2.14	2.27
D'Amico	2.33	1.73
Loeb	2.27	1.64
Matsumura	2.42	2.09
Merrow	2.91	2.91
Saliba	1.55	1.73
Sze	2.09	1.55
Wirtz	3.91	3.64
CA AVERAGE	2.48	2.30

Note: Each notebook average was based on three raters except those marked by * which were based on seven raters.

Table 17, which presents a comparison of the ratings of Colorado and California classrooms summarized by dimension, reveals some interesting patterns. The dimensions along which the two groups of classrooms are most alike are Structure of Lessons and Assessment. These results are consistent with what we know about the curriculum emphases in the two locations. For example, the textbooks used by teachers in both states are well structured, in that the topics are connected from lesson to lesson and the units are arranged in a logical manner. This structure would be sufficient to earn a high score on the Structure of Lessons dimension. Dimension does not capture a feature that is unique to reform-oriented mathematics teaching; neither does the assessment dimension. Teachers in both locales are being encouraged to employ a variety of formal and informal assessment strategies and to use data obtained from assessments for instructional planning and student feedback. To some extent the No Child Left Behind Act has made data-based classroom planning a consistent theme across the country.

Colorado classrooms scored much higher than California classrooms on the other dimensions, and this is consistent with what we know about the curricula being used by the teachers in our study. For example, the greatest difference between the two groups of classrooms was on Cognitive Depth. This dimension is at the heart of the reform curricula. Teachers are encouraged to push students beyond computational fluency to an understanding of the general principles of mathematics. We saw much more of this emphasis among the Colorado classrooms in our study than among those in California. It is important to clarify that the classrooms in our study were not selected to be representative of either state; thus one cannot generalize from this small set of classrooms to the states as a whole. That is not our purpose. We are merely noting that among the 30 teachers we studied, those using reform curricula (who happened to be located in a couple of Colorado districts) were engaged in more reform-oriented instructional practices than those using more traditional curricula (who happen to be located in a couple of California districts.) The differences provide additional evidence that the notebooks are revealing valid differences in practice.

Table 17

Average Notebook-Only and Gold Standard Ratings, by Dimension: Colorado vs. California Classrooms

Dimensions	COLORADO		CALIFORNIA	
	Average	Gold Standard	Average	Gold Standard
Assessment	3.33	3.70	3.13	2.80
Cognitive Depth	3.44	3.40	2.15	1.90
Connections/Applications	2.98	3,05	1.86	2.10
Discourse Community	3.18	3.30	2.09	2.00
Explanation & Justification	3.07	3.35	2.09	1.90
Grouping	3.70	3.45	2.34	1.70
Mathematical Tools	3.17	3.30	2.26	2.00
Multiple Representations	3.51	3.55	2.76	2.50
Problem Solving	3.42	3.40	2.37	2.40
Structure of Lessons	4.42	4.20	3.84	3.90
Overall	3.40	3.50	2.37	2.10
OVERALL	3.42	3.47	2.48	2.30

Note: Each notebook average was based on three raters, except two which were based on seven raters.

Conclusions

Current Analyses

The field study of the Scoop Notebook in middle-school mathematics classrooms reported in this paper reinforced and extended several of the conclusions from our pilot study (Borko, et al, 2003, 2005). Teachers were on the whole interested, supportive and cooperative. Researchers generally met with enthusiasm when presenting the study to the mathematics departments at various California and Colorado middle schools, and teachers responded positively to our invitation to participate in the project. They endorsed the underlying premise that much could be learned from looking at actual classroom work products. Many teachers also appreciated the inclusion of student generated work alongside their reflective comments.

The teachers demonstrated a willingness to participate in the study and put effort into completing the Scoop Notebooks. They were able to follow our artifact collection instructions fairly faithfully. Overall the Notebooks were returned in a timely manner, were reasonably complete, and included some very descriptive classroom photographs. It is significant to note that, as in any community-based research, participants' cooperation was central to the success of the project.

The study was designed to answer five questions about the reliability and validity of the Scoop Notebooks. The first question was whether researchers trained to use a scoring rubric developed for the project could make consistent judgments about instructional practice based on the Scoop Notebooks. On this point, the results are moderately positive. Agreement among raters was reasonably high for a non-standardized collection of materials. For example, on average, teachers in Vermont agreed on the ratings of individual entries in student mathematics portfolios 56% of the time in Grade 4 and 57% of the time in Grade 8. Their agreement on the total score assigned across the whole portfolio was of similar magnitude; 55% of total scores changed at least one quartile from the first to the second reader (Koretz, et al., 1994). Generalizability analyses showed that most dimensions could be rated with reasonable accuracy using three raters. The exceptions were Assessment and Structure of Lessons. As noted in the Results section, small standard deviations for average ratings across notebooks and average levels of agreement between raters suggest that the low generalizability for these two dimensions was due to low variability rather than high levels of disagreement.

We could not answer the second question about the consistency of ratings over time with any certainty. Our design allowed us to confirm a rather trivial point that first observations are not more alike than second observations (when first and second observation days are selected without any systematic plan). However, we could not test how much ratings vary over time or across instructional units within a given classroom.

We conducted several comparisons of ratings based on different data sources to address the third question. We found moderate similarity between scores based on the notebooks and scores based on direct classroom observation. The correlations between the overall classroom average score based on the two methods was 0.85. Dimension-level correlations were lower, but were above 0.70 for half of the dimensions. The lowest correlations were found for Structure of Lessons and Assessment. These correlations are not surprising, given the low generalizability of notebook ratings on these two dimensions. Differences in access to information about formal and informal assessments when observing a classroom versus examining artifacts of practice, and revisions to the Structure of Lessons dimension between the scoring of observations and notebooks, may also help to explain these results.

As we expected, there was even greater similarity between scores based on notebooks and gold standard scores. The correlation between the classroom average score based on notebooks and the average gold standard score was 0.89. Dimension-level correlations were also slightly higher than when comparing notebook scores to observation scores, but the lowest two dimensions remained Structure of Lessons and Assessment. Another way to think of these correlations is that when summarized across dimensions, notebook-only scores rank teachers almost the same as observation or as gold standard scores. Considering all of the analyses, it seems reasonable to conclude that judgments based only on the Scoop Notebook provide a reasonably accurate picture of classroom practice during the data collection period. At the same time, there is reason to consider refining the definitions of Structure of Lessons and Assessment, as these two dimensions posed the greatest problems with respect to both reliability and validity of ratings.

The analyses that incorporated audiotapes and transcripts of lessons suggest that classroom discourse provides additional insights into the nature of instructional practice, particularly for the dimensions of Mathematical Discourse Community, Explanation and Justification, Cognitive Depth, Connections/Applications, and Assessment. As it is currently designed, the Scoop Notebook by itself is not able to completely capture the nature of discourse-based classroom interactions that are represented in these dimensions. At the same time, discourse alone is not sufficient to provide an accurate portrayal of instructional practice. The level of agreement between discourse-plus-notebook ratings and gold standard ratings is relatively high. This suggests that these two combinations of data provide similar portrayals of instructional practice. These analyses indicate that it is worthwhile to consider adding discourse data to the Scoop Notebook. However, the costs associated with the collection and analysis of discourse data must be balanced against the potential value of these results, and the added cost may make it difficult to incorporate these data into large-scale research programs.

The fifth question concerned variation among dimensions when rating notebooks and observations. On the one hand, dimensions posed different challenges to raters; judgments about some dimensions could be made with high levels of consistency, while judgments about others could not. This suggests that dimensions represented distinctly different aspects of practice. On the other hand, a single factor captured most of the variation in ratings among classrooms. It may be the case that the dimensions represent different features of classrooms, but that reform-oriented practice, where it exists, typically includes all the features that were measured by the Scoop

Notebooks. For example, teachers who emphasize Cognitive Depth also emphasize Mathematical Discourse, at least to some extent.

As discussed above, some dimensions and instructional practices continue to present a greater challenge than others in terms of reliability and validity. In addition, some teachers and classrooms presented greater difficulties than others. There does not appear to be a relationship between consistency in rater judgments across classrooms and either average ratings (high vs. low), notebook completeness, or rater confidence. We plan to explore other possible explanations, such as individual differences among raters, in future studies.

Future Directions

Results of this field study suggest that the Scoop Notebook is useful for describing instructional practice in broad terms. For example, ratings of instructional practice based on the notebooks matched known differences between reform and traditional curricula used in Colorado and California, respectively. The Scoop Notebook might thus be useful for providing an indication of programmatic changes in instruction over time that occur as a result of overall program reform efforts. We do not think, however, that the evidence is strong enough to support use of the notebooks for making judgments about individual teachers.

The Scoop Notebook may also prove to be a useful tool in professional development programs. Teachers may find it helpful for describing and reflecting on their own instructional practices, and for collaborating with colleagues and researchers to better understand the nature of mathematics learning and teaching in their classrooms. As one example, the notebook could help them to trace changes in their instructional practices over time or across instructional units. Similarly, notebooks can offer a focal point for collaborative lesson study. Unsolicited comments offered by a number of participating teachers provide support for the use of the Scoop Notebook in these ways.

In closing, we should point out that there are important questions this study did not address. For example, teachers' instructional practices' may vary naturally over time and content, and these variations may have an impact on ratings of some or all of the dimensions. To explore this issue, future studies could incorporate multiple data collection points in the same classroom over time. Future studies should also use multiple classroom observers so they can determine the accuracy of observational ratings. We did not have the resources to send more than one observer to a classroom, so we were unable to determine how accurate these ratings were. This study offers some encouragement about the usefulness of the Scoop Notebooks as indicators of instructional practice in mathematics, but there is certainly more to be learned about their quality and functionality.

References

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

Ball, D. L. & Rowan, B. (2004). Introduction: measuring instruction. *The Elementary School Journal, 105,* 3-10.

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

Borko, H., Stecher, B.M., Alonzo, A., Moncure, S., & McClam, S. (2003). *Artifact packages for measuring instructional practice***.** (CSE Technical Report No. 615). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Borko, H., Stecher, B.M., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact packages for measuring instructional practice: A pilot study. *Educational Assessment, 10,* 73-104.

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

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

Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105,* 49-74.

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

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

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

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

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

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

Koretz, D., Stecher, B. M., Klein, S., McCaffrey, D. and Diebert, E. (1993). *Can portfolios assess student performance and influence instruction: The 1991-92 Vermont experience*. CSE Technical Report 371. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Lotman, Y.M. (1988). Text within a text. *Soviet Psychology, 26* (3), 32-51.

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

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

Mehan, H. (1979). *Learning lessons*. Cambridge, MA: Harvard University Press.

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

Rowan, B., Harrison, D.M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal, 105,* 103-128.

Scott, P. (1998). Teacher talk and meaning making in science classrooms: A Vygotskian analysis and review. *Studies in Science Education, 32*, 45-80.

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.

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

Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge: Harvard University Press.

Wertsch, J. V., & Toma, C. (1995). Discourse and learning in the classroom: A sociocultural approach. In L. P. Steffe & J. Gale (Eds.), *Constructivism in education* (pp. 159-174). Hillsdale, NJ: Erlbaum.

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

Appendix A:
Reflection Questions in the Scoop Notebook

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 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.
   - 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.)

   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 lecture/whole class discussion format. Finally, the students work independently (or sometimes in partners) on practice problems. 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 math content and problems we are working on. Usually I have a problem on the board when the students arrive to class. We discuss it briefly as a whole class to

be sure they all understand the problem, then they begin to work on it in groups. I walk around to help answer questions and facilitate the group discussions. When they are done solving the problem in their groups, each group takes a turn presenting/defending their solution to the class. I wrote that this varies, because sometimes we will work on a few problems in a period and other times the groups work together on a single problem for the period, and we don't take turns presenting solutions until the next day.

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.

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 math content, lesson goals/objectives, instructional strategies, student activities)

   For example,
   - We are in the middle of a unit on representations of data. This week, we will start out by exploring the strengths and weaknesses of various data representations (graphs, tables, etc) using examples from newspapers and other media sources. We will take data sets from these sources and represent them in different forms to see how the various representations change our impressions of the data.
   - This week, we are working on developing students' understandings of ratios and proportions and their ability to solve problems that use ratios and proportions. We will begin the week by reviewing what students have previously learned about both topics. Then we will learn how to simplify ratios and how to convert from proportions to percentages. Students will solve a number of word problems, involving ratios and proportions.

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 to the success of our project. 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 that a line on a graph represents a relationship between two variables (x and y). Today I used a comparison between a positive and a negative slope of a line graph. The objective of the lesson was for students to create two graphs: one with a positive and one with a negative slope. For each graph, the students needed to identify two variables that have a positive (or negative) relationship and create logical data to fit that relationship. I wanted the students to choose practical examples from their own life. (I gave the example of amount of ice cream eaten every day and pounds gained.) They then needed to graph the data correctly. I was also checking that they could draw the graphs correctly, because we worked on this in the last unit I taught.
   - Today's lesson had two different objectives. During this unit, we will be working on problems using fractions and percents. The first objective today was to begin using problem-solving strategies such as creating a representation of the problem (by drawing or using manipulatives) and then using this representation to create a solution. The second objective was for students to develop the ability to communicate mathematically. I wanted the students to work on their mathematical communication abilities by both working in groups and also writing in journals. Mathematical communication 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 problem of the day that was written on the board (see photo of board #4). Then we reviewed the homework, and I answered student's questions (see copy of homework assignment). The majority of class was spent doing a whole-class review of multiplication of mixed numbers (see handout and photo of board at end of class #5). Towards the end of the lesson, students worked individually on a set of problems in the textbook, which they are expected to complete for homework (page 113 odd-numbered problems).

- At the beginning of class, students turned in their group projects (see samples of student projects) comparing various representations of data on the number of M&Ms of each color in a bag of M&Ms. They spent about 10 minutes writing 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, followed by an introduction to the next activity with M&Ms—determining the proportion of M&Ms of each color in each bag of candy and comparing the proportions across groups (see handout).

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 line graphs and bar graphs, but I discovered that a number of students in the class didn't have a firm grasp of line graphs. So, we backtracked and created a line graph on the board. We plotted the students' grades on the last chapter test against the number of hours of television students watched last week.
   - My lesson for the day was addition of polynomials. But at the beginning of class, one of the students asked how come NBC News had declared the winner of last night's election just 1 hour after the polls had closed, with less than 1% of the votes counted. Since we had done a unit on statistics just a couple of weeks ago, I thought it would be a good opportunity to discuss statistical sampling. The students became very interested and we talked about actually doing a survey of the school and analyzing it. Don't know if we'll have time to fit it in.

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 be able to use the procedure for converting from decimals to percents. However, I was surprised that a couple of students struggled with problems converting percentages less than 10 % to decimals. I realized that they were struggling through…
   - My expectations for group work 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. They had worked out an approach to counting M&Ms by color and creating a table and graph to record their data. They also did a good job of

allocating tasks to each of the group members. I realized that my expectations were met by…

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 mathematics 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?*

Appendix B:
Dimensions of Reform Practice Used for Observation Ratings

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

2. Structure of Lessons. Extent to which the series of lessons is organized to be conceptually coherent such that activities build on one another in a logical manner leading toward deeper conceptual understanding.

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

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

5. Cognitive Depth. Extent to which the series of lessons promotes command of the central concepts or "big ideas" of the discipline and instruction generalizes from specific instances to larger concepts or relationships. Extent to which teacher listens to students and responds in ways that scaffold student understanding toward this larger understanding.

6. Mathematical Discourse Community. 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. Extent to which the teacher and students "talk mathematics," and students are expected to communicate their mathematical thinking clearly to their peers and teacher using the language of mathematics.

7. Explanation and Justification. Extent to which students are expected to explain and justify their reasoning and how they arrived at solutions to problems (both orally and in written assignments). The extent to which students' mathematical explanations and justifications incorporate conceptual, as well as computational and procedural arguments.

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

9. Assessment. The extent to which the series of lessons includes a variety of formal and informal assessment strategies to support the learning of important mathematical ideas and furnish useful information to both teachers and students (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. 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 reflect a model of instruction consistent with the NCTM Standards. This dimension takes into account both the curriculum and the instructional practice.