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

CSE Technical Report 690

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

July 2006

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 Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analysis of Current Assessment and Accountability Systems
Hilda Borko, University of Colorado, and Brian M. Stecher, RAND, Project Directors

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

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

Hilda Borko, CRESST/University of Colorado
Brian Stecher, CRESST/RAND

Abstract

This report presents findings from two investigations of the use of classroom artifacts to measure the presence of reform-oriented teaching practices in middle-school science classes. It complements previous research on the use of artifacts to describe reform-oriented teaching practices in mathematics. In both studies, ratings based on collections of artifacts assembled by teachers following directions in the "Scoop Notebook" are compared to judgments based on other sources of information, including direct classroom observations and transcripts of discourse recorded during classroom observations. For this purpose, we developed descriptions of 11 dimensions of reform-oriented science instruction, and procedures for rating each on a dimension-specific five-point scale.

Two investigations were conducted. In 2004, data were collected from 39 middle-school science teachers in two states. Each teacher completed a Scoop Notebook, each was observed by a singe observer on two or three occasions, and eight of the teachers were also audio-taped, allowing us to create transcripts of classroom discourse. In 2005, 21 middle-school mathematics teachers participated in a similar study, in which each teacher was observed by a pair of observers, but no audio-taping occurred.

All data sources were rated independently on each of 11 dimensions. In addition, independent ratings were made using combinations of data sources. The person who observed in a classroom also reviewed the Scoop Notebook and assigned a "gold standard" rating reflecting all the information available from the Notebook and the classroom observations. Combined ratings were also assigned based on the transcripts and notebooks, and based on the observations and transcripts.

The results of these field studies suggest that the Scoop Notebook is a reasonable tool for describing instructional practice in broad terms. For example, it could be useful for providing an indication of changes in instruction over time that occur as a result of program reform efforts. There was a moderate degree of correspondence between judgments of classroom practice based on the Scoop Notebook and judgments based on direct classroom observation. Correspondence was particularly high for dimensions that did not exhibit great variation from one day to the next. Furthermore, judgments based on the Scoop Notebook corresponded moderately well to our "gold standard" ratings, which included all the information we had about practice.

Project Goals and Overview

Our long-term research program investigates the 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 actually do in classrooms (rather than what they believe they do). We use a data collection tool called the "Scoop Notebook" to gather classroom artifacts and teacher reflections related to key features of classroom practice. To date, we have studied the use of artifacts in two subject areas—middle school mathematics and science. We conducted pilot studies to provide initial information about the reliability, validity and feasibility of artifact collections as measures of classroom practice in these subjects (Borko, Stecher, Alonzo, Moncure, & McClam, 2005), and a field study to validate the Scoop Notebook in middle school mathematics classrooms (Stecher, Borko, Kuffner, Wood, Arnold, et al., 2005). Our notebook and scoring procedures were revised on the basis of results from each of these studies.

In this report, we present the results of two related studies conducted to validate the Scoop Notebook as a measure of reform-oriented instructional practice in middle school science classrooms. The report first describes the notebook, the 11 dimensions of instructional practice it measures, and the associated scoring rubrics. Next, the methodology employed for the two studies is presented, including study design and data collection procedures. Results from both studies are integrated in the next section which documents the reliability and validity of the Scoop Notebook for measuring reform-oriented practice in science. The analyses address two main research questions:

1. What is the reliability of raters' judgments of instructional practice based on the Scoop Notebook, transcripts of classroom discourse, and classroom observations?
2. What is the evidence to support conclusions about the validity of ratings based on the Scoop Notebook as a measure of reform-oriented instructional practice in science?

- To what extent do scores assigned on the basis of the Scoop Notebook agree with those assigned on the basis of transcripts or classroom observations, and with scores that use all available information about a classroom (i.e., "gold standard" scores based on observations and the Notebook).

- Are the patterns of relationships among the 11 dimensions of reform-oriented instruction consistent across notebooks and classroom observations?

Methods

Overview

We conducted two field studies in middle-school science classrooms in Colorado and California. The first study, conducted in 2003-04, investigated the reliability and validity of ratings of practice based on the Scoop Notebook and audiotape transcripts. The second study, conducted in the spring of 2004-05, focused on examining the reliability of ratings of practice based on classroom observations. Together, the studies provide complementary evidence regarding the validity of the Scoop Notebook as a tool for characterizing reform-oriented instructional practice in science.

Participants

For the 2003-04 field study, we contacted a diverse sample of middle schools in districts that had participated in a previous study of artifacts in mathematics classes (Stecher et al., 2005). In schools that agreed to participate in the new study, volunteers were sought through notices sent to all science teachers or announcements at meetings of the science department. Thirty-nine teachers participated in this study; 16 were from California, 23 from Colorado. For the 2004-05 study, we contacted districts and schools in California and Colorado that had participated in 2003-04 and in previous studies, and we recruited teachers in a similar manner. Twenty-one science teachers participated in this study—11 in California, and 10 in Colorado. Three or four of these teachers had participated in the 2003-04 study. One of the teachers had participated in an earlier pilot study. In both studies participating teachers received an honorarium of $200-$250 for collecting artifacts, completing reflections, assembling Scoop Notebooks, and being observed.[1]

Data Collection

The Scoop Notebook. As described in previous papers (Borko et al., 2005; Stecher et al., 2005), we developed a tool for the collection of data related to classroom instructional practices using an analogy to the approach of scientists studying unfamiliar territory (e.g., the Earth's crust, the ocean floor). Just as scientists

[1] Different rates were negotiated with districts in Colorado and California based on local practice.

may scoop a sample of materials to take to their laboratories for analysis, we planned to "scoop" materials from classrooms for *ex situ* analysis. Through the use of this tool we hoped to structure the collection of data to obtain information on instructional practices similar to what could be obtained through classroom observations, without the time and expense of such methods. We asked teachers to collect materials produced as part of their regular instruction and then place the materials in a notebook. Because of the usefulness of the analogy, we called our artifact collection package the "Scoop Notebook." When we described the Scoop Notebook to participating teachers, we framed the task in terms of the question: "What is it like to learn science in your classroom?"

For the 2003-04 and 2004-05 studies we asked teachers to collect artifacts from one of their classes for a period equivalent to five normal periods of instruction, following guidelines in the Scoop Notebook. Because we were interested in all types of materials used to foster student learning, we asked teachers to "scoop" materials or artifacts that they and their students generated, as well as materials drawn from textbooks or other curricular resources. The "scooped" artifacts included: instructional materials such as lesson plans, overhead transparencies, and grading rubrics; student work with corresponding teacher reflections; photographs of the classroom; and teacher reflections based on guiding questions posed throughout the Scoop period. 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., 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, and we asked them to 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). They 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 science in your classroom*?" Appendix A provides a complete list of the three sets of reflection questions.

Additional Data Sources: Classroom Observations and Transcripts

In addition to collecting Scoop Notebooks from teachers, 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. During these lessons, observers wrote open-ended field notes describing the lessons they observed. During

the 2003-04 study, observations were done individually, i.e., a single researcher observed each teacher for two or three days.[2] In the 2004-05 study, observations were done in pairs, i.e., the same two researchers observed each teacher on two or three occasions.

In the 2003-04 study, we also collected audiotapes of lessons in eleven 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 transcripts of classroom discourse provided. The researchers who observed in these classrooms audio-taped the lessons by having teachers wear a simple wireless microphone. The audiotapes were transcribed to provide a record of classroom discourse.

Scoring the Notebooks, Observations, and Discourse

In order to evaluate the extent to which teachers emphasized reform-oriented instructional practice in their science classrooms we developed 11 *dimensions* of practice, which were used as the basis for comparison of data collected through notebooks, classroom observations, and transcripts of audio-taped discourse. These dimensions were informed by documents such as the *National Science Education Standards* (National Research Council [NRC], 1996). In the following sections we describe these dimensions, the scoring guides and rating process, and the procedures followed for the training of raters and observers. Unless otherwise noted, the same dimensions, guides and procedures were used in 2003-04 and 2004-05.

Dimensions of Reform-Oriented Practice

The term "reform-oriented science" describes an approach to science teaching that encompasses both content ("what is taught") and pedagogy ("how it is taught"). Reform-oriented science includes practices associated with the idea of "science as process," i.e., having students focus on skills such as observation, measurement, and experimentation. In addition, in a reform-oriented science classroom students learn how to ask and pursue questions, construct and test explanations, form arguments, and communicate their ideas with others (NRC, 1996). Guided by the vision of a science classroom portrayed in the *National Science Education Standards*, as well as elements of standards-based science instruction defined by a panel of experts

[2] In three or four instances in California, a teacher was observed by two different researchers during the three-day observation period.

convened by the Mosaic-II project (Stecher et al., 2005), we identified 11 dimensions of "reform based" instruction in science. The initial versions of the dimensions were revised as a result of the pilot study (Borko et al., 2005) and they were modified slightly between the 2003-04 and the 2004-05 studies to add additional clarification.[3] The final dimension descriptions are as follows:

1. Grouping. The extent to which the teacher organizes the series of lessons to use groups to work on scientific tasks that are directly related to the scientific goals of the lesson and to enable students to work together to accomplish these activities. (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 are related scientifically and build on one another in a logical manner.

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

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

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

6. Cognitive Depth. Cognitive depth refers to a focus on the central ideas of the unit, generalization from specific instances to larger concepts and connections and relationships among science concepts. There are two aspects of cognitive depth: the lesson design and teacher enactment. Thus, this dimension considers extent to which

[3] All changes for the 2004-05 study were minor, except the change to Cognitive Depth described below.

lesson design focuses on cognitive depth and the extent to which teacher consistently promotes cognitive depth.[4]

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

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

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

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

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

Scoring Guides and Rating Process

Each dimension in the Scoop Notebook is rated on a five-point scale from low (1) to high (5). To facilitate the rating process, we developed a scoring guide containing an overall description of the dimension and specific descriptions of the low, medium and high anchor points. For each of these anchor levels, one or two classroom examples are provided, as well. The complete scoring guide used for rating the 2004-05 observations and notebooks is presented in Appendix B. Minor additions were made to the observation guides prior to rating the notebooks, so they would contain examples of the type of evidence found in the notebooks.

---

[4] In 2003-04 the last two sentences read: There are three aspects of cognitive depth: the lesson design, teacher enactment, and student performance. Thus, this dimension considers the extent to which lesson design focuses on cognitive depth; the extent to which the teacher consistently and effectively promotes cognitive depth; and the extent to which student performance demonstrates cognitive depth.

Each source of information (notebooks, transcripts, and observations) was rated on all eleven dimensions by the researcher (or researchers) assigned to do the rating. During the 2003-04 study, notebook readers were given a two-week period during which they checked out the notebooks and transcripts and completed their ratings independently. In addition to the 11 dimensions of instructional practice, researchers rated the Scoop Notebook on "completeness" (reflecting the extent to which the notebook contains all the materials we asked teachers to assemble), and assigned a "confidence" score to their overall set of ratings. In both cases, researchers used a one-to-five scale. In eight of the classrooms where we collected audiotapes, separate ratings were done on the basis of the transcripts alone, and on the basis of the transcripts combined with the notebooks. In both cases, researchers also assigned a confidence score to their ratings. (Notebooks collected in 2004-05 have not been rated because Notebook reliability was not the focus of that field study.)

During both studies, classroom observers took unstructured field notes while conducting the observations and then rated each lesson on the 11 dimensions after completing the observation. The field notes were useful as a reminder of events and as a source of evidence when writing justifications for ratings. In addition, at the conclusion of the scheduled visits to each classroom, observers completed a "summary" rating on each dimension based on everything seen during the two or three observations. The summary observation ratings were not numerical averages of the daily ratings for a given dimension, but separate, qualitative judgments regarding that dimension based on the total set of classroom observations.

At a later time, observers completed a "gold standard" rating for each dimension taking into account both their observations and the information collected by the teachers in the Scoop Notebook. Observers also assigned completeness and confidence scores to the gold standard ratings.

Training of Classroom Observers and Notebook Readers

Observations: In 2003-04 and 2004-05, training sessions were conducted in California and Colorado to prepare observers to use the scoring guide prior to the classroom visits. Videotapes of science lessons were used as the basis for training in order to standardize the training across the two sites. In addition, discussions of ratings were conducted by conference call so everyone could participate at the same time. The 2003-04 observer training meetings began with a review of the scoring guide. Each observer read the guide and the embedded examples, and the group

engaged in extensive discussions to ensure that all raters had similar understandings of the dimensions and scoring levels. The researchers then watched a videotape of a middle school science lesson, took free-form notes while watching the tape, rated the lesson using the guide, and provided descriptions of observed classroom occurrences that justified each rating. At the conclusion of this process, the individual ratings were posted and discussed extensively. This discussion was held via conference call among all observers in California and Colorado. The discussion continued until agreement was reached on the best ratings for the lesson. The discussion sometimes led to changes designed to clarify the scoring guide. These changes usually involved rephrasing descriptions in terms that were more easily understood, or adding examples to characterize the intermediate levels of the dimensions. The process was repeated with a second videotape. At that point, we were satisfied that observers were able to apply the rating guide in a consistent manner.

The 2004-05 training meetings were conducted in a similar manner. The initial meeting ended after rating and discussing one classroom videotape. Following the meeting observers individually viewed and rated three more tapes, and a second conference call was held to review these results. Ratings agreed within one point on almost all dimensions, and researchers came to consensus on all ratings by the conclusion of that call. As noted previously, during the rater training sessions for the 2004-05 study, small changes were again made in the dimension descriptions and scoring rubrics. As in the 2003-2004 study, these changes typically entailed further clarification of the descriptions and additions to the examples. The dimension descriptions and scoring guides reproduced in Appendix B are the versions used in the 2004-05 study. (The earlier versions are available from the principal authors upon request.)

Notebooks: For the 2003-04 study, all the researchers convened in a two-day meeting for training and calibration on using the scoring guide to rate notebooks and audio-taped transcripts. To familiarize readers with the scoring process and make effective use of the meeting time, readers independently rated three notebooks prior to the meeting. The meeting began with an extensive discussion of the scoring guide to ensure that all raters had similar understandings of the dimensions and scoring levels. Then ratings of the three notebooks were posted on a chalkboard, reviewed for differences, and discussed extensively. During these discussions, differences of opinion were resolved, uncertainty about the meaning of the scoring guide was clarified, and, where appropriate, the guide itself was changed. One type of

information that was frequently added to the scoring guide was descriptions of the evidence in the notebooks that would be relevant to rating certain dimensions. For example, the notebooks do not contain direct evidence of discourse, but it was possible to make inferences based on teachers' daily reflections, assignments, and student work contained in the notebook.

Once the ratings and discussion for these three notebooks was completed, the process was repeated with additional notebooks. After this second round of calibration, it appeared that the readers understood and were applying the scoring guide consistently. [5]

Study Design and Analysis Procedures

Reliability of Notebooks and Transcripts (2003-04). Each of 39 teachers who participated in the 2003-04 study completed a Scoop Notebook, and each classroom was observed for two or three days during the time the Scoop was being collected. In general, if the class period lasted 45-55 minutes, the researchers observed on three occasions; if the classroom was on a "block" schedule of 90-100 minutes, it was observed on two occasions. At the end of each lesson, the observer rated the lesson on all eleven dimensions. At the conclusion of the set of lessons, the observer also completed a "summary" observation on each dimension reflecting all that had occurred during the set of lessons. The reliability analyses used these summary ratings.

In eleven of the classrooms, the lessons that were observed were also audio-taped. The teacher wore a microphone and small transmitter, and the observer operated a compact receiver and tape recorder. Teachers were selected for audio-taping on the basis of convenience and their willingness to participate. The tapes were transcribed, and the written transcripts were included as a source of information for the analysis. Table 1 shows the distribution of the 39 teachers who participated in the 2003-04 study by state and type of evidence we collected.

[5] Transcripts were rated in the same manner as the notebooks, and there was not a separate training session for rating the transcripts.

Table 1

Participating Teachers, 2003-04

Type of Evidence	California	Colorado	Total
Scoop Notebook	16	23	39
Classroom Observation	16	23	39
Audio Tape of Discourse	4	7	11

Twenty-eight notebooks were used in the scoring reliability study. Twenty came from classrooms without transcripts, and eight were selected from classrooms that had been audio-taped. The notebooks and transcripts were assigned to eligible raters according to an incomplete, but balanced, design in which each reader rated 15 notebooks and/or transcripts, and each notebook or transcript was rated by two or three readers. The researcher who observed in a given classroom was not considered an eligible rater for the notebook from that classroom. Readers were paired with other readers an equal number of times, and the order in which notebooks and transcripts were read was different for each reader.

For the eight classrooms that were audio-taped, two or three researchers who had not observed in the class completed ratings based only on the transcripts. In addition, three other eligible researchers first rated the classroom based only on the Scoop Notebook and then completed an additional set of ratings based on information contained in both the notebook and the transcripts. This process provided transcript-only ratings for eight classrooms, and transcript-plus-notebook ratings for seven of these classrooms (due to a procedural mistake, only seven of eight were rated based on both information sources). The data sources and number of raters are summarized in Table 2.

In addition, the observers read the notebooks from the classrooms they observed and assigned "gold standard" ratings based on the observations and the notebook taken together. These gold standard ratings reflected all the information we obtained about teaching practice in a given classroom.

Table 2

Sources of Information about Reform-Oriented Classroom Practice, 2003-04

Source of Information	Classrooms	Raters per Classroom
Observation	28	1
Notebook	28	3
Observation + Notebook (GS)	28	1
Transcript	8	2 or 3
Notebook + Transcript	7	3

Note: Nine people acted as observers, and eight people acted as notebook raters in this study.

The primary question that guided the reliability analyses for this study was: How consistent are the ratings of notebooks and transcripts? Because there was only one observer, we could not compute a quantitative indicator of inter-rater reliability for observations. The 2004-05 study (described below) addressed this question.

Two approaches were used to estimate reliability. First, we compared each pair of ratings directly and determined the level of agreement for each of the 11 dimensions. Two levels were tabulated, exact agreement and agreement within one point (on the five-point scale). Since three readers rated each notebook, there were three pairs of comparisons. For each notebook, we computed the fraction of those three pairs that were exact matches (expressed as a percentage) and the fraction that were within one score point (expressed as a percentage). We also computed the average of those percentages across all dimensions for each notebook, and the average across all notebooks for each dimension.

Second, we conducted Generalizability analyses of notebook and transcript ratings using a C*R design in which the object of measurement (classrooms) is crossed with one random facet (raters). The design was incomplete in that not all raters rated all notebooks. These analyses were conducted separately for each dimension because the dimensions were selected purposefully from a universe that we considered limited. SAS PROC VARCOMP was used to estimate variance components, and then both generalizability coefficients (relative) and dependability coefficients (absolute) were estimated for designs using two, three or four raters. Both estimates of consistency were computed because we can envision the notebooks

being used in situations where ranking of teachers is all that is needed and in other situations where absolute interpretations are desired.

Reliability of Classroom Observations (2004-05)

Each of 21 teachers who participated in the 2004-05 study collected materials and completed a Scoop Notebook for a period of five consecutive days of instruction (or the equivalent, for teachers with block scheduling). During this time two researchers observed each classroom for three 45-55 minute periods or two 90- to 100-minute periods. Observers were assigned to classrooms in pairs according to a design that minimized the number of times any two raters were paired together as observers. Each observer visited four or five classrooms, and no two observers were paired together for more than two classrooms.

Each day during the study, observers assigned one rating for each of the eleven dimensions of instructional practice immediately after the lesson. In addition, at the conclusion of the visits, observers completed a "summary" rating on each dimension. As in the previous study, the summary ratings were not numerical averages of the daily ratings for each dimension, but holistic judgments based on the totality of the classroom observations. Finally, at a later time raters reviewed the Scoop Notebooks compiled by teachers they observed and assigned a "gold standard" rating on each dimension, taking into account both the observational data and the materials and information in the Scoop Notebook. Table 3 summarizes the sources of information used in the reliability and validity analyses for this study.

As in the prior study, analysis of the reliability of ratings based on classroom observations was approached in two ways. We first investigated the levels of inter-rater agreement (both exact and within one point) across occasions (classroom observations or visits), and for the summary and gold standard ratings. In addition, we investigated the reliability of ratings of classroom observations through generalizability (g) studies. We analyzed each dimension separately with a two-random-facet design in which the object of measurement (teachers or classrooms) is crossed with raters and occasions. The design is incomplete, because only two observers rate each classroom. SAS PROC VARCOMP was used to estimate variance components for these designs as well as the relative (generalizability) and absolute (dependability) coefficients that would be obtained under hypothetical scenarios with varying numbers of raters and occasions.

Table 3

Sources of Information about Reform-Oriented Classroom Practice, 2004-05

Source of Information	Classrooms	Raters per Classroom
Observation	21	2
Observation + Notebook (GS)	21	2

Note: Nine raters participated in this study.

Validity of Notebook Ratings of Reform-Oriented Instructional Practices (2003-04 and 2004-05)

To investigate the validity of ratings of reform-oriented instructional practices based on the Scoop Notebook we first analyzed the factorial structure (i.e., the *dimensionality*) of data collected through Scoop Notebooks and compared it to the structure of the classroom observation data. Second, we investigated the degree of correspondence between ratings based on the notebook and ratings based on other sources of information. The ratings used in the validity analyses include:

1. *Notebook Ratings:* In 2003-04 each notebook was rated independently by three members of our research team who did not observe in the classroom.
2. *Observation Ratings:* Observers rated each of the 11 dimensions following each lesson. In the 2004 study there was only one rater observing each classroom; in 2005 there were two raters per classroom. In both studies, the researchers who conducted the classroom observations also assigned a "summary" observation rating for each dimension taking into account all of the observed lessons.
3. *Average Transcript Ratings:* For eight of the classrooms in the 2003-04 study, three researchers rated classroom practice based on transcripts of audiotapes of that classroom. We use the average ratings on each dimension across the three raters.
4. *Average Notebook + Transcript Ratings:* For seven of the classrooms in the 2003-04 study, the three researchers who rated the Notebook completed a second set of ratings, taking into account both the Notebook and the transcripts of classroom discourse.
5. *Gold Standard Ratings:* In both studies observers assigned an additional rating taking into account both their observations and the Scoop Notebook. These are termed "gold standard" ratings because they are based on all the information available about the teacher's approach to science instruction.

For each *validity comparison*, we conducted two types of analyses. The first type of analysis considered the level of agreement between the two ratings being

compared. We report level of agreement for each dimension as the percent of classrooms in which the two ratings fell within either 0.5 unit or 1.0 units on the 5-point rating scale. The second type of analysis considered the correlation between the two ratings. We computed a Pearson correlation between the two ratings for each dimension, across all teachers. These correlations indicate the strength of the linear relationship between the two sources of information. We also computed Pearson correlations between ratings for each teacher across dimensions to obtain a general sense of the linear relationship betwen overall judgments of a classroom obtained using both sources of information.

Results and Discussion

In this section we present the results of the two studies that investigated the reliability and validity of ratings of reform-oriented instructional practice in science classrooms based on the Scoop Notebook. Data are drawn from the 2003-04 study and the 2004-05 study as appropriate to the question being asked.

Assessment of reliability is the first step in determining whether notebooks, transcripts, or observations provide consistent judgments about practice. If these sources of information cannot be rated reliably, it is unlikely that they can prove to be a valid indicator of practice. We begin by presenting results that examine the reliability of scores based on each of the data sources: notebooks, transcripts, transcripts plus notebooks, observations, and observations plus notebooks (gold standard ratings).

We then address the issue of validity of ratings of reform-oriented instructional practices based on the Scoop Notebook. First, we investigate and compare the factorial structure (i.e., the *dimensionality*) of data collected through notebooks and observations; then, we investigate the degree of correspondence between ratings based on the notebook and other sources of information.

Reliability of Notebook Ratings

Based on the 2003-04 study, we found moderate agreement among notebook readers on most of the dimensions (see Table 4). When using a five-point scale there is a small likelihood that all three raters will agree on the basis of chance alone (1%), but a significant likelihood they will agree within one rating point due to chance

alone (23%).[6] The results in Table 4 are considerably above these chance thresholds. For example, on average, across the 28 notebooks the three raters agreed exactly on the score they assigned on the Assessment dimension 38% of the time. Similarly, on average, 76% of the pairs of ratings agreed within one point on their scores for Assessment.

Six dimensions had exact agreement at or above 40% and eight had within one agreement at or above 80%. It is interesting to note that the differences in agreement among the dimensions were not as large as in our field study of the mathematics Scoop Notebook and scoring guide (Borko et al., 2005). It may be that science artifacts reveal the dimensions more clearly than mathematics artifacts. Alternatively, it may be that lessons learned from the mathematics study led to improvements in the scoring guide for the science study. Both explanations receive some endorsement from the members of the research team who participated in both studies.

Table 4

Average Ratings and Percent Agreement in Notebook Ratings (3 Raters) by Dimension (28 Classrooms), 2003-04

Dimension	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Assessment	3.24	38	76
Cognitive Depth	2.89	40	82
Connections/Applications	2.82	22	85
Discourse Community	2.61	43	91
Explanation/Justification	2.54	33	88
Grouping	3.60	40	81
Hands-On	3.29	42	77
Inquiry	2.41	38	88
Scientific Resources	3.58	37	75
Structure of Lessons	4.26	47	82
Overall	2.88	40	91

[6] These estimates assume that readers are equally likely to use all five rating levels. If that is not the case (there is some evidence that readers avoid extreme values), then these values underestimate the true chance probability.

The level of agreement did not seem to be a function of the average score on the dimension. That is, readers seemed to do equally well rating notebooks in cases where we found relatively more of a dimension (e.g., Structure of Lessons) and where we found relatively less of a dimension (e.g., Discourse Community).

We found larger variation in agreement across classrooms than we did across notebooks (see Table 5). For example, readers of the notebooks from teachers Shepard and Martin were much less alike in their ratings than readers of the notebooks from Kretke and Saunders. We hypothesized that this difference was due to differences in the completeness of the notebooks. However, further analysis did not bear out this hypothesis. We split the sample of notebooks roughly in half based on completeness (those notebooks rated above 4 on completeness and those rated 4 or below on completeness). There was no consistent pattern in terms of either exact agreement or agreement within one between the two sets of notebooks. For each dimensions, we also computed the correlation between the completeness rating and the percent agreement among raters on the dimension. None of these correlations were significant for exact agreement, and only one of 11 was significant for agreement within one. Lack of relevant information in the notebook does not appear to be a major contributor to lack of agreement among readers. (However, the less complete notebooks received lower scores than the more complete notebooks on almost all dimensions.)

In addition, rater agreement for notebooks was not a function of the average rating for the notebook (which can be construed as an overall indicator of the presence or absence of reform-oriented practices). Some notebooks with low average ratings had high levels of agreement (e.g., Milton), while others had lower levels of agreement (e.g., Sleeve).

In addition to examining inter-rater agreement, we used Generalizability theory to determine how much of the variation in notebook ratings is associated with variation among teachers (classrooms) and raters, and how much is due to unsystematic or random measurement error. This is a C*R design, with separate sources of variation for classrooms (C), raters (R) and residual error (C*R, e). The first step in a Generalizability analysis is to partition the variance among notebooks, raters, and residual/error. Table 6 shows the percent of variance due to each facet for each dimension. High generalizability is obtained when most of the variance is associated with classrooms (*true* variance), and little is associated with unmeasured facets or measurement error (residual).

Table 5

Average Ratings and Percent Agreement in Notebook Ratings (3 Raters) by Classroom (11 Dimensions), 2003-04

Classroom	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Atkinson	2.91	21	82
Baker	3.30	48	85
Beck	2.58	30	85
Bennett	3.15	42	79
Cook	2.52	27	76
Coolidge	3.97	54	91
Douglas	3.64	57	94
Elder	3.21	24	73
Garman	3.15	27	79
Good	3.03	30	79
Jones	2.40	30	63
Kretke	3.82	63	100
Lapp	2.61	24	79
Lesner	3.88	54	91
Martin	3.88	24	76
Milton	1.82	48	97
Newman	2.00	36	91
Reginald	2.51	24	82
Saunders	4.12	73	97
Schmidt	4.21	48	97
Shaker	2.79	30	79
Shafer	3.73	60	85
Shepard	3.88	15	70
Sleeve	2.36	36	79
Solaris	2.21	39	82
Taylor	3.88	30	73
Vonne	2.21	33	88
Walters	3.06	39	82

Table 6

Percent of Variance in Notebook Ratings Attributed to Each Facet by Dimension (28 Classrooms, 3 Raters Per Notebook), 2003-04

Dimension	Classroom	Rater	Residual
Assessment	43.3%	15.3%	41.4%
Cognitive Depth	53.2%	10.4%	36.5%
Connections/Applications	52.6%	5.5%	41.9%
Discourse Community	61.0%	4.7%	34.3%
Explanation/Justification	43.0%	8.9%	48.1%
Grouping	61.0%	-	38.2%
Hands-On	59.6%	-	38.4%
Inquiry	49.7%	8.4%	42.0%
Scientific Resources	51.9%	8.2%	39.9%
Structure of Lessons	28.9%	9.9%	61.1%
Overall	57.1%	8.2%	34.7%

Note: Components that account for less than 3% of the variance are set to zero for ease of interpretation.

The estimated variance components can be used to estimate reliability coefficients for relative decisions (generalizability) and absolute decisions (dependability) for scenarios using different numbers of raters. Relative decisions are based on ranking classrooms along the dimension (i.e., putting them in order) but not considering their absolute score. Absolute decisions involve classifying a teacher against fixed external criteria in terms of *absolute* score on the five-point scale, rather than standing relative to the other teachers in the sample. These coefficients provide a more direct assessment of reliability than agreement indices or variance components alone. Table 7 shows the coefficients that would be obtained if we were to use two, three, or four raters for each notebook. The coefficients are independent of the scale of the rating and can be interpreted as traditional 0 to 1 reliability coefficients.

Table 7

Generalizability and Dependability Coefficients for Notebook Ratings Using Two, Three, and Four Raters by Dimension (28 Classrooms, 3 Raters Per Notebook), 2003-04

Dimension	Relative Decisions			Absolute Decisions		
	Two Raters	Three Raters	Four Raters	Two Raters	Three Raters	Four Raters
Assessment	0.68	0.76	0.81	0.60	0.70	0.75
Cognitive Depth	0.74	0.81	0.85	0.69	0.77	0.82
Connections/Applications	0.71	0.79	0.83	0.69	0.77	0.82
Discourse Community	0.78	0.84	0.88	0.76	0.82	0.86
Explanation/Justification	0.64	0.73	0.78	0.60	0.69	0.75
Grouping	0.76	0.83	0.86	0.76	0.82	0.86
Hands-On	0.76	0.82	0.86	0.75	0.82	0.86
Inquiry	0.70	0.78	0.83	0.66	0.75	0.80
Scientific Resources	0.72	0.80	0.84	0.68	0.76	0.81
Structure of Lessons	0.49	0.59	0.65	0.45	0.55	0.62
Overall	0.77	0.83	0.87	0.73	0.80	0.84

The generalizability analyses confirm that some dimensions are rated with greater accuracy than others. Structure of Lessons stands out as the dimension that is rated with the lowest consistency on the basis of the notebooks; the g-coefficients are notably lower than for any other dimension (note that this dimension is also the one with the largest proportion of residual variance, as shown in Table 6). Using three raters, all dimensions except Structure of Lessons (and perhaps Explanation/Justification and Assessment) can be rated with a reasonable level of consistency if relative decisions are sought (i.e., close to 0.80 or above). The low generalizability of Structure of Lessons may be due to the fact that most classrooms score very high on this dimension (the average rating is over 4.2) and thus there is not much variance among teachers to detect.

For absolute decisions, two other dimensions (Assessment and Explanation/Justification) also fall below a 0.80 threshold. Assessment was difficult to rate, in part, because the definition included "informal assessment" (e.g., judgments made by teachers on the basis of classroom questions and answers), which was hard to infer from the notebooks. Some teachers addressed informal

assessment through comments in their reflections, but raters may have differed in the weight that they gave to these comments. Explanation/Justification also may have been difficult to rate on the basis of the notebooks because it is manifest, in a large part, in teachers' and students' verbal behavior, which is not apparent in most notebooks. The aspects that are evident in the notebook–for example "why" and "justify your answer" questions in written assignments--are easy to overlook when rating the notebooks.

Reliability of Transcript Ratings [7]

Ratings based on transcripts alone were more reliable than ratings based on notebooks for some dimensions and less reliable for others (see Table 8). Relatively higher agreement was achieved for Explanation/Justification and Discourse Community. This makes sense since both dimensions relate to teacher and student verbal behaviors. Comments from readers indicate that it was easier to rate dimensions that depended on conversation on the basis of the transcripts than on the basis of the notebooks. However, transcripts revealed much less about other dimensions, making them more difficult to rate. Particularly difficult were Assessment, Connections, Hands-on, Scientific Resources, and Structure of Lessons. The best evidence for these dimensions comes not from what teachers say but from the activities in which students and teachers engage. This pattern of findings suggests that a combination of notebooks and transcripts may achieve higher reliability than either source alone. However, data presented in the next section show that this is not necessarily the case.

As in the case of notebook ratings, agreement on transcript ratings was not a function of the average score assigned. Dimensions where ratings were higher overall, such as Scientific Resources, were rated with comparable accuracy to dimensions where ratings were lower overall, such as Assessment.

Finally, it is interesting to note that the average ratings across classrooms based on notebooks and transcripts were similar—within 0.5 points—on all 11 dimensions (see Tables 4 and 8). The match between ratings from different sources will be explored in the section on validity.

[7] Because of the smaller sample size (n = 8) we can place less confidence in the reliability estimates based on transcript ratings. The results should be interpreted as exploratory only.

Table 8

Average Ratings and Percent Agreement in Transcript Ratings (2 or 3 Raters) by Dimension (8 Classrooms), 2003-04

Dimension	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Assessment	2.75	21	58
Cognitive Depth	2.77	8	92
Connections/Applications	3.06	33	75
Discourse Community	2.75	46	79
Explanation/Justification	2.23	58	96
Grouping	3.52	17	79
Hands-On	2.75	17	63
Inquiry	2.27	4	92
Scientific Resources	3.21	21	58
Structure of Lessons	3.88	8	75
Overall	2.52	33	96

There was considerable variation in the level of agreement in transcript ratings summarized by classroom (see Table 9). Exact agreement ranged from 9% (Sleeve) to 36% (Jones and Martin); agreement within 1 ranged from 64% to 100%. As was the case with notebook ratings, level of agreement was not associated with average rating. In some cases (e.g., Jones, Taylor) the notebooks and transcripts had comparable levels of agreement—a difference of 10 points or less. However, for other classrooms (e.g., Lesner, Sleeve) the difference between the level of agreement obtained with notebooks and transcripts was considerable—20 or more points.

We also conducted generalizability analyses for transcript ratings, although there were too few classrooms to obtain good estimates. The results should be treated as suggestive, at best. They indicated that acceptable levels of generalizability can be obtained with three raters for most dimensions, except Assessment, Connections, Hands-On, and Structure of Lessons (see Table 10). This pattern is fairly consistent with the results obtained in the previous analysis of rater agreement.

Table 9

Average Ratings and Percent Agreement in Transcript Ratings (2 or 3 Raters) by Classroom (11 Dimensions), 2003-04

Classroom	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Baker	3.24	24	73
Cook	3.18	18	73
Jones	1.86	36	100
Lesner	2.91	18	64
Martin	4.09	36	91
Milton	2.73	24	82
Sleeve	2.23	9	64
Taylor	2.82	27	82

Table 10

Generalizability and Dependability Coefficients for Transcript Ratings Using Two, Three, or Four Raters by Dimension (8 Classrooms, 2 or 3 Raters Per Transcript), 2003-04[8]

Dimension	Relative Decisions			Absolute Decisions		
	Two Raters	Three Raters	Four Raters	Two Raters	Three Raters	Four Raters
Assessment	0.39	0.49	0.56	0.33	0.42	0.49
Cognitive Depth	0.61	0.70	0.76	0.56	0.65	0.71
Connections/Applications	0.51	0.61	0.67	0.36	0.46	0.53
Discourse Community	0.91	0.94	0.96	0.80	0.86	0.89
Explanation/Justification	0.72	0.80	0.84	0.72	0.80	0.84
Grouping	0.80	0.85	0.89	0.76	0.82	0.86
Hands-On	0.00	0.00	0.00	0.00	0.00	0.00
Inquiry	0.75	0.82	0.86	0.50	0.60	0.67
Scientific Resources	0.72	0.80	0.84	0.43	0.53	0.60
Structure of Lessons	0.00	0.00	0.00	0.00	0.00	0.00
Overall	0.72	0.80	0.84	0.72	0.80	0.84

[8] Negative variance estimates for teachers were set to zero, and the absence of true score variance for teachers leads to zero reliability coefficients.

Reliability of Notebook Plus Transcript Ratings

Combining information from notebooks and transcripts does not yield results that are more reliable than notebooks alone. Table 11 shows the level of rater agreement by dimensions for the combination of these two data sources, and Table 12 shows level of agreement by classroom. As revealed by a comparison of the results in Tables 4 and 11, the level of agreement was lower on some dimensions for ratings that took into account both notebooks and transcripts, than for ratings based on notebooks alone. Comparing only those seven classrooms included in Table 12 with the same classrooms in Table 8 shows similar levels of agreement, indicating that agreement was not improved by the considering the information in the notebooks in addition to the transcripts.

Table 11

Average Ratings and Percent Agreement in Notebook Plus Transcript Ratings (3 Raters) by Dimension (7 Classrooms), 2003-04

Dimension	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Assessment	2.76	28	57
Cognitive Depth	2.71	28	81
Connections/Applications	3.09	33	71
Discourse Community	2.72	28	95
Explanation/Justification	2.43	43	86
Grouping	3.48	57	76
Hands-On	3.05	33	86
Inquiry	2.29	19	71
Scientific Resources	3.33	24	52
Structure of Lessons	4.09	38	57
Overall	2.81	28	81

Table 12

Average Ratings and Percent Agreement in Notebook Plus Transcript Ratings (3 Raters) by Classroom (11 Dimensions), 2003-04

Classroom	Average Rating	Exact Agreement (%)	Within 1 Agreement (%)
Baker	3.49	36	82
Cook	2.70	18	73
Jones	2.24	42	60
Lesner	3.76	51	88
Martin	4.06	24	64
Milton	2.48	36	79
Sleeve	2.12	21	73

This lack of improvement may be due, in part, to the small number of classrooms that were included in the notebook-plus-transcript analyses. It may also be due to the fact that the two sources of information may provide conflicting points of view on a dimension. For example, the transcript reveals that the teacher asks students to explain their answers, while the assignments included in the notebook do not call on students to provide any explanations. Readers may resolve such conflicting information differently.

Reliability of Classroom Observations

One common problem when trying to measure teacher practice is unmeasured variance due to *occasions*—differences in teacher practice from one day to another (see Shavelson, Webb, & Burstein, 1986). This was not a concern when rating notebooks because there is only one rating per teacher for the entire period; day-to-day variance in teacher practice is *averaged over the whole Scoop period*. For classroom observations, day-to-day variation is a concern. The 2003-04 study used a single rater across multiple occasions and as a result, fluctuations in teacher practice from one occasion to another cannot be separated from inconsistencies in rater and from unexplained error. There is essentially only one facet that can be measured, which includes all these sources of variation. The 2004-05 study employed multiple observations and multiple observers per classroom to address this problem, and that study provides the best estimates of the reliability of classroom observations. In this section, we first present analyses of levels of agreement for summary ratings only;

we then present generalizability analyses that take into account all potential sources of variation. Table 13 presents descriptive statistics and levels of agreement between two observers on summary ratings after three visits to the classroom on each of 11 dimensions of reform-oriented instructional practice. Exact agreement was moderate and comparable to that observed in the 2003-04 study of notebook ratings. Agreement was consistently greater than the levels expected by chance (25%). Agreement within one point was also similar to that observed with notebook ratings. While there is variation across dimensions in terms of levels of exact and within-1 agreement, this variation is not dependent on the average level of a dimension observed by raters—there was no correlation between average ratings and exact or within-1 agreement.

Table 13

Mean Summary Observation Rating and Inter-Rater Agreement for 11 Dimensions of Reform-Oriented Instructional Practice in Science (Across Teachers), 2004-05

Dimension	Summary Rating			
	Average Rating	Std. Dev.	% Agreement	
			Exact	Within 1
Assessment	2.90	0.98	28.6	76.2
Cognitive Depth	2.60	1.08	38.1	90.5
Connection/Applications	2.88	1.23	38.1	76.2
Discourse Community	2.67	1.07	33.3	81.0
Explanation/Justification	2.43	1.11	28.6	85.7
Grouping	3.43	1.40	47.6	85.7
Hands-On	2.71	1.52	57.1	95.2
Inquiry	2.07	0.84	61.9	95.2
Scientific Resource	3.45	1.31	47.6	90.5
Structure of Lesson	3.74	1.11	33.3	85.7
Overall	2.71	0.89	57.1	95.2

As with notebooks, the range of variation in levels of agreement for observations across classrooms (i.e., agreement for each teacher averaging over all 11 dimensions) was greater than across dimensions (i.e., agreement for each dimension averaging over all teachers). Table 14 shows that exact agreement for summary

ratings was as low as 9% in some classrooms, and as high as 73% in others, while the within-1 point agreement levels ranged from 45% to 100%.

For the Generalizability studies of classroom observation ratings, we considered each observation separately and thus employed a design with two random facets (a C*R*O design, for classrooms, raters, and occasions) that partitions the variance in ratings into seven separate components: teachers or classrooms (C), raters (R), occasions (O), the two-way interactions between these three components (C*R, C*O, R*O), and a final component which combines the interaction between the three, other unspecified sources of variance, and measurement error (C*R*O+e). Classrooms (or teachers) are the object of measurement and therefore variance attributed to this component is considered *true score variance*; generalizability increases as the proportion of this variance relative to the total variance in the model increases.

Table 14

Average Summary Observation Ratings and Percent Agreement for 21 Teachers (2 Raters) by Classroom (11 Dimensions), 2004-05

Teacher	Summary Rating		
	Average Rating	% Agreement	
		Exact	Within 1
Alleman	2.64	45.5	100.0
Carmiano	3.73	18.2	90.9
Koper	2.18	63.6	100.0
Marek	2.41	54.5	100.0
Lcpp	2.41	36.4	63.6
Richardson	3.86	36.4	100.0
Schmidt	4.27	45.5	81.8
Shafer	2.55	54.5	90.9
Said	3.23	72.7	100.0
Storm	3.68	72.7	100.0
Judd	2.64	36.4	90.9
Kellogg	3.41	45.5	90.9
Jones	3.36	27.3	81.8
Vaughan	3.27	63.6	100.0
Sampson	2.95	18.2	81.8
Davis	2.09	36.4	90.9
Kemble	2.23	27.3	90.9
Valley	1.55	63.6	100.0
Foster	2.14	36.4	45.5
Kennedy	2.91	36.4	72.7

Simmons	2.82	9.1	54.5

Table 15 presents the estimated variance components for each dimension of the generalizability design described above. Across dimensions, classroom variance (i.e., *true score* variance) accounts for 17-52% of the total variance in the model. *True* variance is highest for Scientific Resources and the Overall dimensions and lowest for Connections/Applications and Discourse Community.

Table 15

Percent of Variance in Classroom Observation Ratings Attributed to each Facet by Dimension (21 Classrooms, 2 Raters, 3 Occasions), 2004-05

Dimension	C	R	O	C*R	C*O	R*O	CRO,e
Assessment	25.5%	19.8%	-	12.4%	13.0%	-	29.2%
Cognitive Depth	30.2%	-	-	11.5%	16.2%	-	42.0%
Connections/Applications	16.7%	8.6%	-	12.8%	28.5%	-	25.9%
Discourse Community	17.3%	-	-	49.9%	7.4%	-	22.6%
Explanation/Justification	33.2%	17.4%	-	19.7%	3.3%	-	26.1%
Grouping	24.2%	-	-	5.9%	52.0%	-	17.4%
Hands-On	47.8%	-	-	-	45.3%	-	-
Inquiry	47.9%	10.0%	-	6.3%	10.5%	-	25.1%
Scientific Resources	52.1%	-	-	-	26.3%	-	12.2%
Structure of Lessons	26.6%	10.5%	-	12.1%	16.6%	-	29.2%
Overall	51.9%	6.9%	-	10.8%	13.5%	-	16 .6%

Note: Components that account for less than 3% of the variance are set to zero for ease of interpretation.

Variance due to Raters is considerable only for two dimensions: Explanation/Justification and Assessment, indicating that some raters tend to perceive consistently higher or lower levels of these dimensions than other raters. These are also the dimensions where exact agreement was lowest, which could indicate a need for further rater training or refinement of the dimensions.

A considerable amount of variance is attributed to interaction of Classrooms and Raters (C*R) for two dimensions: Discourse Community and Explanation/Justification. This variance component reflects inconsistencies in the

ratings of classrooms by different raters, averaged across observations. These inconsistencies are of particular concern, as they essentially reflect disagreement among raters with respect to their relative ranking of the same classrooms (averaging over occasions), and thus may indicate inconsistencies in the ways raters understand the dimensions for particular kinds of teachers or classroom situations. Additional rater training or fine-tuning of the definitions of these dimensions may be needed to clarify the patterns of behaviors characteristic of each level of these two dimensions.[9] In light of the agreement results presented previously, however, one optimistic (but feasible) interpretation would be that although the classroom by rater interaction is sometimes an important source of variation, it still reflects somewhat limited *disagreement*, as most pairs of ratings fall within an acceptable one-point range.

Occasions (or observations) are not a significant source of variance as a main effect—for all dimensions similar levels were observed across classroom observations, averaging over teachers and raters. The interaction of occasions and raters is also negligible—raters were equally consistent across time; they did not vary systematically in their ratings from one observation to the other, averaging across classrooms.

On the other hand, in this study there is a substantial Classroom by Occasion interaction for most of the dimensions. The nature of this interaction deserves special attention. Statistically, this component indicates that classrooms are rank-ordered differently across classroom observations, averaging over raters. Insofar as it reflects expected (and *true*) day-to-day variation in teacher practice in the classroom during a unit or series of lessons, C*O is not *measurement error* in the typical sense. However, this component does indicate that a considerable degree of uncertainty (and thus error) would be involved in generalizing from a single classroom observation to the *true* score for that classroom if it were observed an infinite number of times (Shavelson et al., 1986), and therefore is considered error variance when estimating reliability (generalizability) coefficients. This kind of uncertainty is particularly acute in the case of Grouping, Use of Scientific Resources, Hands-On Activities, and Connections and Applications—perhaps naturally, since hands-on activities can be heavily emphasized one day and not the next, while cognitive depth would ideally

[9] Variance components estimated separately by state indicate the C*R interaction was smaller in Colorado than in California. The agreement levels in Table 15 and discussions among the researchers indicate that despite the common training observers in the two states may have rated using somewhat different rulers for some dimensions (i.e. gave points based on different elements of the dimension).

remain more consistent across lessons. The effect of this interaction on the estimated reliability of ratings of these dimensions will become clear in the following section.

While the small proportion of true variance relative to all other variance in Table 16 could seem to imply low levels of reliability across dimensions, a direct reading of the variance components can be misleading without considering the final design that will be employed for the ratings. The variance components reflect the proportion of the variance of individual ratings due to the different sources in the model. However, the average rating for a teacher over multiple observers and occasions should intuitively contain less measurement error and be a more reliable indicator of the teacher *true* score than any one of those ratings taken separately.

Generalizability Theory enables researchers to estimate the reliability that would be obtained under alternative scoring designs on a 0 - 1 scale similar to that of other common reliability coefficients (e.g., Cronbach's alpha). Table 16 presents the estimated generalizability coefficients for relative and absolute decisions that would be obtained in alternative scenarios that employ five observations and two, three and four raters. These coefficients provide an estimate of the reliability that would be attained if observers visited the classrooms the same number of times as the number of days teachers collect evidence of their teaching practice for the Scoop Notebook.

In general, the results indicate that adequate reliability based on classroom observations can be attained for some dimensions. Assuming three raters, the average over five hypothetical independent observations would achieve reliability of 0.7 or above for relative decisions (i.e., rank-ordering classrooms) for eight of the 11 dimensions. For absolute decisions, five of the 11 dimensions achieve reliability of 0.7 or above.[10] The dimensions rated with the lowest absolute reliability are Scientific Discourse Community and Connections/Applications, while the most reliable are Use of Scientific Resources and Inquiry.

[10] These coefficients reflect the reliability of ratings based on classroom observations in a 0 to 1 scale representing the ratio of true variance to score variance. However, whether a particular coefficient constitutes *good enough* reliability is a substantive decision that depends crucially on the intended use of the score. Standard errors can also be used to provide a sense of the precision of the measures with respect to the rating scale used (Cronbach, Linn, Brennan, & Haertel, 1997).

Table 16

Generalizability and Dependability Coefficients for Classroom Observation Ratings by Dimension (5 Occasions), 2004-05

Dimension	Relative Decisions			Absolute Decisions		
	Two Raters	Three Raters	Four Raters	Two Raters	Three Raters	Four Raters
Assessment	0.69	0.75	0.78	0.54	0.63	0.68
Cognitive Depth	0.70	0.75	0.79	0.70	0.75	0.79
Connections/Applications	0.53	0.59	0.62	0.45	0.52	0.55
Discourse Community	0.38	0.47	0.53	0.37	0.46	0.52
Explanation and	0.72	0.79	0.83	0.60	0.69	0.75
Grouping	0.62	0.64	0.66	0.62	0.64	0.65
Hands-On	0.83	0.83	0.83	0.82	0.83	0.83
Inquiry	0.86	0.89	0.91	0.79	0.84	0.87
Scientific Resources	0.87	0.88	0.89	0.84	0.86	0.87
Structure of Lessons	0.68	0.74	0.77	0.59	0.66	0.71
Overall	0.84	0.87	0.89	0.80	0.84	0.87

The variance components in Table 15 indicate that differences in teacher practice from day to day are an important source of variance in the model, and the reliability coefficients in Table 16 therefore include this variance component as error variance. With the Scoop Notebook, on the other hand, although teachers collect information and materials during five days of instruction they then receive a single score for the entire period (for each dimension). Therefore, variance across occasions cannot be estimated with notebook ratings and thus does not enter in the estimation of reliability coefficients.

To provide a direct comparison of the reliability of observations to the reliability of notebook ratings, variance components and reliability coefficients were estimated based on the summary observation rating assigned by raters after three visits to the classroom. This results in a one-facet (C*R) design equivalent to the design used with notebook ratings. Table 17 presents the variance components: as with notebooks (see Table 5), an occasion facet is not included and variance is thus partitioned into three components: Classrooms (C), Raters (R), and a combination of interaction and error (C*R,E).

Table 17

Percent of Variance in Summary Observation Ratings Attributed to each Facet by Dimension (21 Classrooms, 2 Raters), 2004-05

Dimension	Classroom	Rater	Residual
Assessment	35.0%	30.9%	34.0%
Cognitive Depth	63.9%	-	32.6%
Connections/Applications	48.3%	9.7%	41.9%
Discourse Community	33.2%	8.8%	57.9%
Explanation and Justification	31.8%	24.4%	43.7%
Grouping	76.1%	-	23.8%
Hands-On	87.8%	-	12.1%
Inquiry	68.0%	8.3%	23.5%
Scientific Resources	76.8%	-	23.1%
Structure of Lessons	57.6%	21.9%	20.3%
Overall	64.8%	-	34.0%

Note: Components that account for less than 3% of the variance are set to zero.

By way of comparison, Table 18 presents the generalizability coefficients for absolute decisions (dependability coefficients) estimated from multiple classroom observations, summary observation ratings, and notebook ratings. In general, the reliability coefficients for summary observation ratings are higher than those for the average of multiple observation ratings. In interpreting this finding it should be taken into account that summary ratings are also based in multiple classroom visits and inherently "average over" measurement error associated with occasions. Moreover, this summary is not a simple mathematical average, but instead is likely to weigh pieces of evidence from different days and sources differently. Therefore, it should not be surprising that these summary scores are more reliable than the averages of scores for multiple observations, particularly in the case of dimensions where large differences occur from one day to another (e.g., Grouping, Hands-On).

The particular dimensions identified as problematic with classroom observation data vary depending on the ratings (and thus the design) used for the analyses. Averaging over three raters and multiple hypothetical observations (i.e., the C*R*O design) the dimensions with lowest reliability were Discourse Community, Connections/Applications, and Grouping. For the two latter dimensions, the reason

was the large variation in teacher practice from one observation to another (i.e., a large C*O interaction), while Discourse Community had a very large teacher by rater interaction. For summary observation ratings after three visits to the classroom, the dimensions with lowest reliability are now Explanation/Justification, Discourse Community, and Assessment—these results resemble those for notebook ratings in Table 7 more closely.

Table 18

Dependability Coefficients for Multiple and Summary Observation Ratings and Notebooks Ratings by Dimension, 2004-05

Dimension	Five Observations		Summary Observation		Notebook	
	2 Raters	3 Raters	2 Raters	3 Raters	2 Raters	3 Raters
Assessment	0.54	0.63	0.52	0.62	0.60	0.70
Cognitive Depth	0.70	0.75	0.78	0.84	0.69	0.77
Connections/Applications	0.45	0.52	0.65	0.74	0.69	0.77
Discourse Community	0.37	0.46	0.50	0.60	0.76	0.82
Explanation/Justification	0.60	0.69	0.48	0.58	0.60	0.69
Grouping	0.62	0.64	0.86	0.91	0.77	0.82
Hands-On	0.82	0.83	0.94	0.96	0.75	0.82
Inquiry	0.79	0.84	0.81	0.86	0.66	0.75
Scientific Resources	0.84	0.86	0.87	0.91	0.68	0.76
Structure of Lessons	0.59	0.66	0.73	0.80	0.45	0.55
Overall	0.80	0.84	0.79	0.85	0.73	0.80

As was mentioned previously, the most relevant comparison for the reliability of notebook ratings is that of summary observation ratings, because of the similarity of the designs (i.e., a single rating is assigned after considering information collected over multiple days). The results in Table 18 indicate that ratings of comparable reliability can be achieved with both sources of information. For four dimensions (Inquiry, Hands-on, Scientific Resources, and Structure of Lessons) the generalizability of summary observation ratings was higher than that of ratings based on the notebook, while for three other dimensions (Assessment, Discourse Community, and Explanation/Justification) the reverse was true. For the remaining

three dimensions Generalizability coefficients obtained with both data collection methods are within 0.1 point. Also, the reliability of the Overall dimension (a weighted average across the other 10 dimensions) is similar with notebooks compared to summary observation ratings.

Reliability of Gold Standard Ratings

Table 19 shows descriptive statistics and levels of agreement between observers on ratings of 11 dimensions of reform-oriented instructional practice that integrate all information from classroom visits and the notebook. Percent of exact agreement with these *gold standard* (GS) ratings was generally similar to summary observation ratings, although somewhat higher for agreement within 1. One interesting exception was Inquiry: exact agreement in this dimension was 38% for gold standard ratings, compared to 62% for classroom observations. One possible explanation is that for this dimension raters may have interpreted the materials in the notebook and the information they obtained during classroom observations as being inconsistent rather than complementary, and different raters may have resolved these inconsistencies in different ways.

Table 19

Mean Gold Standard Ratings and Inter-Rater Agreement for 11 Dimensions of Reform-Oriented Instructional Practice in Science (Across Teachers), 2004-05

Dimension	Gold Standard			
	Average Rating	Std. Dev.	% Agreement	
			Exact	Within 1
Assessment	3.24	0.98	38.1	85.7
Cognitive Depth	2.79	1.07	28.6	100.0
Connection/Applications	3.19	1.21	42.9	81.0
Discourse Community	2.74	0.94	38.1	81.0
Explanation/Justification	2.71	0.94	28.6	76.2
Grouping	3.55	1.25	47.6	90.5
Hands-On	2.90	1.59	47.6	81.0
Inquiry	2.21	0.87	38.1	100.0
Scientific Resource	3.81	1.38	57.1	85.7
Structure of Lesson	3.88	1.02	38.1	90.5
Overall	2.88	0.90	50.0	100.0

As with observation and notebook ratings, the range of variation in agreement of gold standard ratings across classrooms was greater than across dimensions. Table 20 shows a range of exact agreement for gold standard ratings that goes from a low of 0% to a high of 82%, while for within-1 agreement, the levels ranged from 64% to 100%.

Table 20

Average Gold Standard Ratings and Percent Agreement for 21 Teachers (2 Raters) by Classroom (11 Dimensions), 2004-05

Teacher	Gold Standard		
	Average Rating	% Agreement	
		Exact	Within 1
Alleman	2.68	36.4	100.0
Carmiano	4.00	36.4	90.9
Koper	2.23	36.4	100.0
Marek	2.23	54.5	81.8
Lcpp	2.50	45.5	72.7
Richardson	3.86	36.4	100.0
Schmidt	4.23	63.6	81.8
Shafer	2.73	45.5	100.0
Said	3.41	45.5	90.9
Storm	3.82	81.8	100.0
Judd	3.45	45.5	100.0
Kellogg	3.55	54.5	90.9
Jones	3.41	36.4	81.8
Vaughan	3.71	20.0	70.0
Sampson	3.14	27.3	90.9
Davis	2.50	9.1	90.9
Kemble	2.55	63.6	100.0
Valley	2.00	45.5	100.0
Foster	2.09	54.5	81.8
Kennedy	3.64	27.3	63.6
Simmons	3.05	0.0	63.6

Table 21 presents variance components and reliability coefficients for absolute decisions (dependability coefficients) for the gold standard ratings combining information from the materials collected with the notebook over five days, and classroom observations conducted during the same period of time. For each dimension raters assigned only one gold standard rating to each classroom so the design does not include an occasion facet and is thus equivalent to that of notebook ratings (Table 6) and summary observation ratings (Table 17). The Dependability coefficient for gold standard ratings was 0.7 or more for 9 of the 11 dimensions. Overall, the reliability coefficients for gold standard ratings tend to be similar to or higher than reliability coefficients for summary observation ratings (and Scoop Notebook ratings, see Table 18. The only exception is Explanation/Justification, where reliability decreases when raters incorporate information from the notebooks in their ratings. This could point to inconsistencies in rater interpretation of the contents of the notebook in relation to their own observations of the classroom behavior relevant to the Explanation/Justification dimension. For example, some raters may have given more weight to their observations with respect to this dimension, while others may have given considerably higher weight to materials and artifacts in the notebook as evidence of practices that they were not able to observe during their visits.

Table 21

Percent of Variance and Dependability Coefficients for Gold Standard Ratings by Dimension (21 Classrooms, 2 Raters), 2004-05

Dimension	Variance Components			Dependability Coefficient		
	Classroom	Rater	Residual	2 Raters	3 Raters	4 Raters
Assessment	54.9%	9.9%	35.1%	0.71	0.79	0.83
Cognitive Depth	69.9%	6.6%	23.4%	0.82	0.87	0.90
Connections/Applications	45.7%	-	54.2%	0.63	0.72	0.77
Discourse Community	39.0%	6.8%	54.1%	0.56	0.66	0.72
Explanation and Justification	20.3%	-	79.6%	0.34	0.43	0.50
Grouping	73.8%	-	24.6%	0.85	0.89	0.92
Hands-On	74.0%	-	25.0%	0.85	0.90	0.92
Inquiry	66.6%	21.7%	11.5%	0.80	0.86	0.89
Scientific Resources	78.0%	-	21.9%	0.88	0.91	0.93
Structure of Lessons	56.7%	10.6%	32.6%	0.72	0.80	0.84
Overall	73.0%	5.9%	21.0%	0.84	0.89	0.92

Note: Components that account for less than 3% of the variance are set to zero.

Validity of Ratings of the Scoop Notebook

We now address the question of whether the scores assigned to the Scoop Notebook (or, alternatively, the notebooks plus transcripts) are valid indicators of reform-oriented classroom practice. Our criterion measures of instructional practice are the observation ratings (or, alternatively, the gold standard ratings, i.e., observations plus notebooks). For this purpose, we first employ factor analytic techniques to investigate the extent to which a reduced number of constructs (i.e., dimensions) can be used to explain the 11 dimensions of reform-oriented practice developed for this study—and additionally whether different patterns are observed based on notebooks and observations. Then we investigate the degree of correspondence between ratings assigned on the basis of notebooks, transcripts, and observations, to determine whether the notebooks (or notebooks plus transcripts) can serve as a reasonable surrogate for more costly and time-consuming data sources, such as direct classroom observation.

Dimensionality of Notebook and Classroom Observation Ratings

Table 22 presents the results of factor analysis of notebook ratings from the 2003-04 study. Based on ratings from three readers, the first extracted factor explained 49% of the variance in the 10 dimensions of reform-oriented instructional practice (the *overall* dimension was omitted here so as to not to introduce an artificial element of unidimensionality). Factor loadings indicate that the first factor gives more weight to Inquiry, Cognitive Depth, and Scientific Discourse Community. While these results suggest an important proportion of shared variance among the 10 dimensions, they also indicate that additional factors are needed to more fully account for the patterns of relationship among the dimensions. This finding is consistent with a previous study of the Mathematics Scoop notebook, where a single factor accounted for 53% of the variance across 10 dimensions of reform instruction in Mathematics (Stecher et al., 2005).

A three-factor solution was attempted based on the results of exploratory analysis and a priori considerations about the relationship among the dimensions. The first extracted factor groups in six dimensions, with heaviest load on Inquiry, Cognitive Depth, Scientific Discourse Community, and Assessment. The second extracted factor is closely related to use of Scientific Resources and the extent of Hands-On experiences in the classroom. This factor can help illuminate the nature of the difference in results observed with mathematics and science notebooks. In

science classrooms, these features are distinct from other dimensions typically associated with reform-oriented practices such as cognitive depth and inquiry. Indeed, as noted in the National Science Education Standards, "Conducting hands-on science activity does not guarantee inquiry" (NRC, 1995, p. 23).

Table 22

Factor Loadings for Unidimensional and Three-Factor Solutions, Notebook Ratings, 2003-04

	1 Factor	3 Factors (Promax Rotation)		
		Factor 1	Factor 2	Factor 3
Assessment	0.80	0.82	0.34	0.29
Cognitive Depth	0.81	0.83	0.33	0.44
Connections/Applications	0.60	0.56	0.52	0.01
Discourse Community	0.81	0.87	0.26	0.18
Explanation/Justification	0.76	0.77	0.39	0.17
Grouping	0.76	0.77	0.43	0.12
Hands-On	0.58	0.40	0.91	0.05
Inquiry	0.82	0.85	0.35	0.25
Scientific Resources	0.59	0.37	0.92	0.30
Structure of Lessons	0.36	0.27	0.16	0.97

Connections/Applications is the only dimension that exhibits a pattern of cross-loading across factors (i.e., it relates equally strongly to the first and second extracted factors). This dimension includes both conceptual connections between science, the world, and society, and the application of science to real world contexts. As a result it could be conceived as relating both to Cognitive Depth or Inquiry (the first factor), and to laboratory experiments aimed at applying science, which often involve hands-on activities with scientific equipment. Finally, Structure of Lessons is not strongly related to the other two factors and is singled out in a third factor. This suggests that our conception of structure of lessons is distinct from the factor that emphasizes cognitive aspects of practice and the factor that emphasizes experimental science. To the extent that the dimension reflects well-ordered, connected lessons, then we might expect it to be equally likely in reform-oriented and traditional classrooms and classrooms that are different than the vision set forth in the *National Science Education Standards* (NRC, 1996).

For comparison, Table 23 presents the results of similar factorial analyses conducted on the data from classroom observations from the 2004-05 study. Based on ratings from two observers on three occasions the first extracted factor explained 39% of the variance in the 10 dimensions (as before, the O*verall* dimension was omitted from analysis). Using summary observation ratings instead, the proportion increases only slightly to 42%.[11] In both cases the factor loadings suggest that this factor gives more weight to Cognitive Depth, Scientific Discourse Community, and Explanation-Justification.

Table 23

Factor loadings for Unidimensional and Three-Factor Solutions, Classroom Observation Ratings, 2004-05

	1 Factor	3 Factors (Promax Rotation)		
		Factor 1	Factor 2	Factor 3
Assessment	0.59	0.75	0.04	0.11
Cognitive Depth	0.83	0.75	0.41	0.52
Connections/Applications	0.57	0.53	0.21	0.42
Discourse Community	0.79	0.84	0.34	0.20
Explanation/Justification	0.70	0.89	0.13	0.04
Grouping	0.37	0.00	0.70	0.32
Hands-On	0.45	0.06	0.84	0.21
Inquiry	0.64	0.50	0.72	-0.12
Scientific Resources	0.66	0.38	0.80	0.20
Structure of Lessons	0.47	0.23	0.24	0.93

In a three-factor solution the first factor groups Cognitive Depth, Discourse Community, Explanation/Justification, Assessment, and Connections/Applications. The second factor now includes Grouping, Scientific Resources, Hands-on, and Inquiry. As with notebooks Structure of Lessons loads on a third factor; however, in this case Connections/Applications is also moderately related to this third factor. In general, these results resemble those obtained with notebook ratings (see Table 21), suggesting a similar dimensional structure is revealed by the two methods of rating

[11] Consistent with this pattern, the first extracted factor accounts for 44% of the variance in gold standard ratings.

practice. This adds some support to our claim that the two procedures are providing evidence about the same constructs. On the other hand, it also raises questions about the number of independent dimensions of practice that are being captured by either method. Both analyses were conducted using a relatively small number of cases compared to the number of dimensions, so these results should be interpreted cautiously.

Comparison of Information from Multiple Sources

The second step in investigating the validity of ratings of reform-oriented instruction based on the Scoop Notebook involves determining the extent to which these yield judgments about instructional practice in a classroom similar to those we would obtain using other sources of information (i.e., classroom observations, audiotape transcripts). In this section we present analyses from both the 2003-04 and the 2004-05 studies which investigated the level of agreement and the correlation between pairs of ratings of the same classrooms based on different sources of information. First we compared ratings based on the notebook to summary observation and gold standard ratings to determine the extent to which the notebook produced similar ratings than these two methods. We then compared ratings based on the combination of notebooks and transcripts (n+t) to summary observation and gold standard ratings; these comparisons aimed to determine whether adding transcripts improved the effectiveness of notebooks as a measure of science instruction. Finally, we compared the summary observation ratings to gold standard ratings.

Comparing Notebook and Summary Observation Ratings.

We first investigate the extent to which inferences about instruction based on the Scoop Notebook are similar to those based on classroom observations. Table 24 provides information about the similarity of scores assigned by these two methods in the 2003-04 study.[12] While observation ratings were always integers, notebook ratings were averaged over multiple raters and were usually not integers. Thus, instead of "exact agreement" we selected two values for differences that represented high and moderate degrees of "closeness" as a standard of comparison—the percentage of classrooms for which ratings were within 0.5 units on the five-point rating scale and the percentage of classrooms for which ratings were within 1 unit.

[12] In the 2004-05 study independent notebook ratings were not obtained for each classroom.

Table 24

Percent Agreement and Correlations Between Average Notebook and Summary Observation Ratings by Dimension (28 Classrooms), 2003-04.

Dimension	Average Notebook Rating	Summary Observation Rating	Within 0.5 Agreement (%)	Within 1 Agreement (%)	Pearson Correlation
Assessment	3.24	3.04	29	75	0.54
Cognitive Depth	2.89	2.79	29	75	0.53
Connections/Applications	2.82	2.82	32	71	0.55
Discourse Community	2.61	2.61	36	75	0.64
Explanation/Justification	2.54	2.21	36	64	0.62
Grouping	3.60	3.43	57	75	0.61
Hands-On	3.29	3.00	50	75	0.76
Inquiry	2.41	2.14	36	86	0.69
Scientific Resources	3.58	3.46	50	71	0.55
Structure of Lessons	4.26	4.25	32	82	0.26
Overall	2.88	2.86	29	71	0.57

Agreement between average ratings of a classroom based on the Scoop Notebook and summary ratings based on classroom observations varied by dimension. For all dimensions except Explanation/Justification, ratings agreed within one point for over 70% of the classrooms (at least 20 of 28). For Inquiry, the average notebook ratings and summary observation ratings were within one point for 86% of the classrooms (24 of 28). Using a more strict criterion of agreement (within 0.5 point), consistency between ratings was low for several dimensions: Assessment, Cognitive Depth, Connections/Applications, Structure of Lessons, and Overall. On these five dimensions, ratings based on the notebook and ratings based on observations were within 0.5 point for only eight or nine of the 28 classrooms.

Table 24 also shows the correlations between ratings based on notebooks and classroom observations for each dimension in the 2003-04 study. Pearson correlations provide another way to analyze the correspondence of judgments made on the basis of different sources of information. These values indicate the extent to which there is a linear relationship among the ratings based on the two different data sources.

For all dimensions except Structure of Lessons correlations were 0.5 or higher, suggesting moderate agreement between rank orderings of classrooms based on

these two sources. The much lower correlation for Structure of Lessons would be expected based on the low generalizability of the notebook ratings on this dimension (see Table 7).

We computed two summary correlations to indicate the overall correspondence between notebook ratings and observation ratings. First we considered each of the eleven dimensions for each of the 28 teachers as a separate piece of information (308 in all). This correlation was 0.67. Second, we computed average notebook and observation ratings across the eleven dimensions for each classroom. We can think of these average ratings as representing the researcher's overall impression of the reform-oriented nature of a teacher's instructional practices. The correlation between average notebook and observation ratings across teachers was moderate to high at 0.71, suggesting that overall impressions of reform-oriented practice based on the Scoop Notebook are similar to overall impressions based on classroom observations.

Comparing Notebook and Gold Standard Ratings. The second comparison examines the correspondence between ratings of classrooms based solely on the Scoop Notebook and ratings based on the notebook supplemented with direct classroom observation (i.e., gold standard ratings). A high degree of correspondence between these two sets of ratings would be evidence that inferences about instruction based on the artifact notebook are similar to impressions formed when observing a class and reviewing classroom artifacts. Table 25 provides information about the correspondence of the scores assigned by these two methods in the 2003-04 study.

For all 11 dimensions, agreement was within one point for over 70% of the classrooms. For four dimensions – Discourse Community, Hands-on, Inquiry, and Scientific Resources, this level of agreement was achieved in over 80% of the classrooms. Using the stricter criterion of agreement within 0.5 point, ratings of Assessment, Cognitive Depth, Connections/Applications, Explanation/Justification, and Overall were consistent in only one third of the classrooms or less.

Table 25

Agreement and Correlation between Average Notebook and Gold Standard Ratings by Dimension (28 Classrooms), 2003-04

Dimension	Average Notebook Rating	Gold Standard Rating	Within 0.5 Agreement (%)	Within 1 Agreement (%)	Pearson Correlation
Assessment	3.24	3.11	21	71	0.54
Cognitive Depth	2.89	3.04	25	71	0.41
Connections/Applications	2.82	3.07	32	75	0.70
Discourse Community	2.61	2.64	39	82	0.70
Explanation/Justification	2.54	2.29	32	71	0.54
Grouping	3.60	3.57	61	75	0.67
Hands-On	3.29	3.21	43	89	0.85
Inquiry	2.41	2.39	43	86	0.62
Scientific Resources	3.58	3.68	61	82	0.59
Structure of Lessons	4.26	4.32	39	75	0.26
Overall	2.88	3.00	32	75	0.59

Correlations were 0.5 or higher for all dimensions except Cognitive Depth and Structure of Lessons. For three dimensions the correlations were 0.7 or higher, indicating reasonably high correspondence between rank orderings of classrooms when using ratings based on artifacts and gold standards. Again, the much lower correlation for Structure of Lessons would be expected based on the low generalizability of the notebook ratings on this dimension.

As before, we computed two summary correlations to indicate the overall correspondence between notebook ratings and gold standard ratings. First, considering each of the eleven dimensions for each of the 28 teachers as a separate piece of information the correlation was 0.68. Second, the correlation between the average notebook rating and average gold standard across dimensions for each classroom was 0.72. These correlations suggest that judgments of reform-oriented practice based on the Scoop Notebook are reasonably similar to judgments based on the combination of Scoop Notebook and classroom observations (gold standard judgments).

It is interesting to note that in most cases the level of agreement between notebooks and observations (Table 24) was not substantially different than that between notebooks and gold standard ratings (Table 25). Since gold standard ratings are based on observations and notebooks combined this suggests that, for most dimensions, notebooks may offer little additional information to raters beyond that collected during classroom observations (i.e., both sources provide a good deal of overlapping information). Alternatively, it could also mean that raters are more influenced by their own observations than by the materials in the notebook. The following set of results confirm this conclusion

Comparing Summary Observation and Gold Standards Ratings. Table 26 presents the levels of agreement and correlations between the average of summary observation and gold standard ratings (averaged across the two researchers in each classroom) in the 2004-05 study.[13] Agreement between summary observation and gold standards ratings was consistently high. Across dimensions, both ratings were within 0.5 points in more than 80% of classrooms; and agreement within one point was reached in 90% to 100% of classrooms. The correlation between the summary observation and gold standard ratings was also consistently high—for nine of 11 dimensions the correlation it was 0.90 or above—reflecting very similar relative standing of classrooms based on both measures.[14]

These very high levels of correspondence indicate that researchers' ratings of instructional practice based on a combination of observations and artifacts are very similar to their ratings made solely on the basis of their observations (without the added information that the artifacts provide). This confirms the earlier impressions either that the two sources contain overlapping information or that raters may be more persuaded by, or inclined to rely on, information collected through their own observations than on the information collected by the teacher in the notebooks.

[13] Unlike the 2003-04 study, in 2004-05 two raters observed each classroom so here we compare average Summary Observation and Gold Standard ratings.

[14] The results in the 2004 study (using a single summary observation and gold standard ratings per classroom) were very similar. The ratings were within one point in 93% of classrooms for one dimension, 96% for five dimensions, and 100% for the remaining five dimensions. In addition, correlation coefficients were above 0.9 for all but three dimensions–and all were above 0.8.

Table 26

Agreement and Correlation between Average Summary Observation and Gold Standard Ratings by Dimension (21 Classrooms), 2004-05

Dimension	Average Summary Rating	Average Gold Standard	Within 0.5 Agreement (%)	Within 1 Agreement (%)	Pearson Correlation
Assessment	2.90	3.24	71	100	0.82
Cognitive Depth	2.60	2.79	95	100	0.95
Connections/Applications	2.88	3.19	90	95	0.93
Discourse Community	2.67	2.74	86	100	0.90
Explanation and Justification	2.43	2.71	81	95	0.84
Grouping	3.43	3.55	90	100	0.96
Hands-On	2.71	2.90	81	95	0.95
Inquiry	2.07	2.21	100	100	0.96
Scientific Resources	3.45	3.81	81	90	0.92
Structure of Lessons	3.74	3.88	90	100	0.96
Overall	2.71	2.85	86	95	0.92

Comparing Notebook + Transcript and Observation Ratings. Another question the study sought to answer was whether the addition of transcript information to notebooks (n + t) improved their effectiveness as a measure of reform-oriented instruction. These analyses draw from data in the 2003-04 study, and they provide evidence about the practical value of supplementing artifact collection with classroom transcripts in studies of instructional practice. Across dimensions, the levels of agreement between notebooks-plus-transcript ratings and observation ratings presented in Table 27 are similar to those in Table 24 between notebooks and observations—50% to 75% for agreement within one- point, and 13% to 63% for agreement within 0.5 points. The correlations were 0.6 or higher for seven of the dimensions; for those seven dimensions classrooms were ranked fairly similarly using ratings based on the notebook and transcripts of classroom discourse, and ratings based on direct classroom observation. The lowest observed correlations were for Structure of Lessons (0.20), and Connections/Applications (0.33).

It is important to remember however, that the results presented in Table 27 are based on a subset of only seven classrooms for which notebook-plus-transcript ratings were available and thus should be considered tentative and exploratory.

Table 27

Agreement and Correlation between Average Notebook Plus Transcript Rating and Observation Ratings by Dimension (7 Classrooms), 2003-04

Dimension	Average Notebook + Transcript Rating	Average Observation Rating	Within 0.5 Agreement (%)	Within 1 Agreement (%)	Pearson Correlation
Assessment	2.89	2.69	25	63	0.69
Cognitive Depth	2.86	2.38	25	63	0.83
Connections/Applications	3.40	2.96	38	50	0.33
Discourse Community	2.59	2.46	13	63	0.88
Explanation/Justification	2.29	1.85	50	75	0.80
Grouping	3.60	3.17	63	75	0.38
Hands-On	2.95	2.54	25	63	0.60
Inquiry	2.15	1.60	25	63	0.87
Scientific Resources	3.46	2.90	25	38	0.39
Structure of Lessons	4.11	3.88	25	50	0.20
Overall	2.91	2.50	13	63	0.85

We computed two summary correlations to indicate the overall correspondence between notebook + transcript and observation ratings. First, considering the ratings for each dimension for each teacher as a separate piece of information the correlation between n+t and observation ratings was 0.67. Second, the correlation between the average n+t and observation ratings for each classroom was 0.75. Although the small sample size must be kept in mind, these values suggest that judgments of reform-oriented practice based on the combination of Scoop Notebook and transcripts are similar to judgments based on direct classroom observations.

Comparing Notebook + Transcript and Gold Standard Ratings. The final set of analyses presented in Table 28 addresses the question of whether ratings based on the Scoop Notebook supplemented by transcripts of lessons are similar to gold standard ratings (which combine information from the notebook and classroom observations). Agreement across dimensions was generally higher for this analysis than for the first two analyses that include transcripts–43% to 100% for agreement within 1 point, and 0% to 86% within 0.5 points. As before, however, these analyses

are based on a small sample of classrooms (n=7) and the results are therefore tentative.

Table 28

Agreement and Correlation between Average Notebook Plus Transcript Rating and Gold Standard Rating by Dimension (7 Classrooms), 2003-04

Dimension	Average Notebook + Transcript Rating	Gold Standard Rating	Within 0.5 Agreement (%)	Within 1 Agreement (%)	Pearson Correlation
Assessment	2.89	3.00	0	43	0.51
Cognitive Depth	2.86	3.00	43	86	0.80
Connections/Applications	3.40	3.71	14	86	0.77
Discourse Community	2.59	2.43	14	57	0.81
Explanation/Justification	2.29	2.14	14	86	0.91
Grouping	3.60	3.71	86	100	0.94
Hands-On	2.95	2.86	29	100	0.89
Inquiry	2.15	2.00	29	100	0.73
Scientific Resources	3.46	3.57	57	71	0.70
Structure of Lessons	4.11	4.14	29	57	0.21
Overall	2.91	3.00	43	71	0.91

Pearson correlations between the two combined sources of information were close to or higher than 0.9 for four dimensions—Explanation/Justification, Grouping, Hands-On, and Overall. Thus, the two methods revealed a strong linear relationship among the seven classrooms on these dimensions

We also computed two summary correlations to indicate the overall correspondence between notebooks plus transcript, and gold standard ratings. First, we considered each of the eleven dimensions for each of the seven teachers as a separate piece of information (77 in all); this correlation was 0.76. Second, we computed the correlation between the average notebook-plus-transcript rating and average gold standard rating for each classroom; this correlation was 0.86. The high values for both correlations suggest that judgments of reform-oriented practice based on the combination of Scoop Notebook and transcripts are similar to judgments based on the combination of Scoop Notebook and classroom observations (gold

standard judgments). Because of the small sample size, we are hesitant to draw inferences about correspondence between the two sets of ratings for individual dimensions of classroom practice.

Conclusions

Reliability of Notebook Ratings

The results indicate that Scoop Notebooks can be rated with reasonable levels of reliability on most dimensions if three readers are used. Readers had the lowest level of agreement rating Structure of Lessons, and the factor analyses results showed that Structure of Lessons did not relate closely to the other dimensions. It might be the case that improvements to the scoring guide for this dimension could resolve the problem. However, the fact that the dimension itself does not seem to be closely aligned with the other features of reform-oriented instructional practice in science may indicate that it is not being implemented consistently or that it is not clearly evident in the notebook, and therefore that notebooks may not be an adequate mechanism for measuring this dimension. Alternatively, it could reflect a lack of variance in the dimension—i.e. most teachers' lessons are well structured, so that the dimension does not discriminate well between classrooms that are more or less reform-oriented within this narrow range. Very high average ratings on this dimension lend some support to the latter explanation.

Readers also failed to agree on their ratings of a small number of the Scoop Notebooks. On inspection, these notebooks contained fewer artifacts than the typical notebook completed during this field study. Completeness of the notebook does not appear to be a major factor in rating correspondence. However, although the vast majority of teachers followed the directions in the notebook and provided rich collections of artifacts, some did not. Better, more detailed directions might help overcome this problem.

Teachers' reflections, in particular, may warrant further attention. Many teachers were clear, complete, and insightful in reflecting on their daily lessons and on student work; others were not. We anticipated that we might have some difficulty obtaining thoughtful reflections from all teachers, in part, because completing the reflections was probably the most time-intensive aspect of compiling the Scoop Notebook. To address this potential difficulty we included examples of "rich" reflections in the instructions, providing teachers with models to follow. Further, to accommodate different styles of composition, we permitted teachers to submit

reflections in writing, as computer files, or on audiotape—even supplying tape recorders to teachers who wanted to use that method but did not have access to equipment. Finally, we provided teachers with a generous honorarium ($250) for participating in the project to motivate them to complete the assignments fully.

Nevertheless, even under the favorable conditions of our research studies, some teachers were very limited in their reflective comments. It would appear that this problem stems from factors that are difficult to control, such as differences in teachers' willingness to write full explanations, differences in available time from day to day, the level of insight they have into their own practice, and their willingness to share these insights.

Our small study of transcripts yielded mixed results. Most importantly, transcripts were not rated as reliably as notebooks on a number of dimensions. These results may be due to the quality of the audio recordings, which was relatively poor. To keep data collection practical, we used a single wireless microphone worn by the teacher, which could be used simply and without complicated apparatus and support personnel. Although this equipment captured everything the teacher said, it was often impossible to hear students' responses. Having "half" the conversation was helpful to raters, but not as helpful as it might have been to have the complete verbal exchanges. Despite these limitations, the combination of notebooks and transcripts resulted in increased reliability for some dimensions (including those for which evidence is likely to be found in verbal exchanges). The number of cases in which we had both sources of information was small, however (only seven classrooms), so the study did not produce a strong test of the added value of transcripts.

Overall, the results suggest that artifacts can be collected in a systematic manner and scored consistently-enough to be used to compare science teaching across classrooms.

Reliability of Classroom Observation Ratings

The most appropriate analysis to consider when drawing conclusions about the reliability of classroom observations is the analysis of the summary observation ratings assigned by observers after three visits to the classroom. These global ratings for a series of lessons are most similar to the Scoop Notebooks in that they average over multiple observations.

The reliability of summary observation ratings was generally comparable to that of notebook ratings. As was the case with notebooks, the dimension with the lowest reliability was Structure of Lessons. With the exception of that dimension, all dimensions were rated reasonably consistently by pairs of observers who saw the same lessons. For four dimensions (Inquiry, Hands-on, Scientific Resources, and Structure of Lessons) the generalizability of summary observation ratings was higher than that of ratings based on the notebook, while for three other dimensions (Assessment, Discourse Community, and Explanation/Justification) the reverse was true. For the remaining three dimensions and the *Overall* dimension similar generalizability coefficients were obtained with both data collection methods. It should be noted that there were small modifications to the dimensions between the 2003-04 study and the 2004-05 study. Thus, differences in the reliability estimates of notebooks and observations could-reflect differences between the two methods of data collection, differences between the definitions of the dimensions across studies, or a combination of both.

Validity of Notebooks as Measures of Reform-Oriented Instructional Practice

We found a moderate to high degree of correspondence among ratings of reform-oriented classroom practice based the various data sources we examined. In general, Scoop Notebooks (and notebooks plus transcripts) yielded portrayals of practice that were similar to portrayals based on observations (and observations plus notebooks). This provides evidence that the notebook ratings are valid indicators of reform-oriented practice, as judged by direct observation.

At the broadest level (averaging across dimensions), the judgments about reform-oriented practice from these sources were highly correlated. That is, the different data sources portray the practices of individual teachers in similar ways. The correlations were as follows:

- Average Notebook and Summary Observation: r = 0.71
- Average Notebook +Transcript and Summary Observation: r = 0.74
- Average Notebook and Gold Standard: r = 0.72
- Average Notebook + Transcript and Gold Standard: r = 0.86
- Summary Observation and Gold Standard: r = 0.95

These summary correlations provide useful information about similarities among the various sources of information. First, there is very little difference between notebooks and notebooks plus transcripts when making comparisons to observations. In general, adding transcripts did not enhance the validity of ratings, although there were selected dimensions for which ratings based on notebooks plus transcripts were more like those based on observations than were ratings based on notebooks alone. Second, summary observation ratings and gold standard ratings were the most similar in terms of ranking teachers. This similarity probably reflects the fact that both ratings include evidence gained from observations, which seem to be the most persuasive source of information to raters.

Overall, based on evidence from the two studies presented in this report it is reasonable to conclude that there is a moderate degree of correspondence between judgments of classroom practice based on the Scoop Notebook and judgments based on direct classroom observation. Correspondence is particularly high for dimensions that do not exhibit great variation from one day to the next. Furthermore, judgments based on the Scoop Notebook correspond moderately well to our "gold standard" ratings, which include all the information we have about practice. Thus the results of these field studies suggest that the Scoop Notebook is a reasonable tool for describing instructional practice in broad terms. For example, it could be useful for providing an indication of changes in instruction over time that occur as a result of program reform efforts.

However, the evidence is not strong enough to support use of the Scoop Notebook for making judgments about individual teachers. Neither the reliability of the notebook ratings nor the correspondence between notebook rating and other evidence about classroom practice are high enough to justify using the notebooks for individual rating. It may be possible that something like the Scoop Notebook, combined with either direct classroom observation or transcripts of lessons, would provide sufficiently robust evidence for more high stakes decisions but further research would be needed to validate using combined data sources for that purpose.

Results of the two field studies suggest several modifications to the Scoop Notebook that warrant further exploration. Completion of the Scoop Notebook imposes a moderate burden on teachers. Further, in the context of a large research project, it might not be possible to offer honoraria sufficient to motivate complete responses. Additional research should be conducted to determine whether the process can be streamlined and still permit consistent judgments. Another change to

explore is the provision of more training for the teachers. Prior to data collection, we met with all science teachers who participated in these studies individually or in small groups for approximately one-half hour to review the notebook instructions. It might be possible to improve the quality of the Scoop Notebook by asking teachers to complete notebook materials for one day and then reviewing their materials and providing feedback. However, each additional training may not be practical for most large-scale research studies.

Several additional lines of research would be helpful in further exploring the feasibility of the Scoop Notebook as a research tool. First, further information is needed about the costs—in terms of time, resources, and money–associated with the collection and analysis of the Scoop Notebook, classroom observations, and transcripts of classroom discourse. The efficacy of the artifact collection as a research tool will depend on cost as well as reliability and validity. Second, we need additional research to address the question of time sampling, i.e., how many observations of a science classroom over what period of time during the school year are needed to provide a stable portrayal of instruction in that classroom. Finally, additional analyses are needed to determine whether instructional practices in science can be characterized as reliably and validly using fewer dimensions than the eleven dimensions that comprise our current rating system. The process might be simplified considerably (and perhaps made more reliable and valid) if fewer broad dimensions could be used. The results from factor analyses of ratings based on notebooks and observations were similar to those observed in our previous study of the mathematics Scoop Notebook (Stecher et al., 2005). In both cases there was considerable overlapping variance among the 10 dimensions; this suggests that it might be possible to distill a general reform instruction construct, with one or two additional dimensions reflecting unique features (and variance).

References

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment, 10*(2), 73-104.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3).

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.), *Handbook of research on teaching*. New York, NY: McMillan.

Stecher, B. M., Borko, H., Kuffner, K. L., Wood, A. C., Arnold, S. C., Gilbert, M. L., & Dorman E. H. (2005). *Using classroom artifacts to measure instructional practices in middle school mathematics: A two-state field test* (CSE Technical Report 662). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Appendix A
Teacher Reflection Questions

(Note: Explanations and examples have been removed for brevity.)

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?*

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

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

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

Daily Reflection Questions

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

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

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

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

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

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

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

7. *Have you completed the Daily Calendar and Photograph Log entries for today?*

Post-Scoop Reflection Questions

To be answered at the end of the Scoop timeframe.

1. *How does this series of lessons fit in with your long-term goals for this group of students?*

2. *How representative of your typical instruction was this series of lessons (with respect to content, instructional strategies and student activities)? What aspects were typical? What aspects were not typical?*

3. *How well does this collection of artifacts, photographs, and reflections capture what it is like to learn science in your classroom? How "true-to-life" is the picture of your teaching portrayed by the Scoop?*

4. *If you were preparing this notebook to help someone understand your teaching, what else would you want the notebook to include? Why?*

Appendix B
Scoring Guide

Science SCOOP Rating Guide
CRESST Artifact Project

February 15, 2005

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

In all dimensions (unless otherwise specified)...

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

General Notes for Observation and Summary Ratings:

1. Take notes during the observation of each lesson.
2. Use a separate observation rating each day and then a summary at the end.
3. At the end of the observations (after all days of observation) the "overall" rating is a holistic rating (rather than mathematical average). [For the Spring 2005 ratings, each observer should work independently and do his/her own summary rating.]
4. It is sometimes difficult to rate a dimension based on the observation of one lesson, especially when the dimension description includes a "series of lessons."

General Notes for Gold Standard Ratings:

1. Use your rating forms and notes from the class observations, as well as the notebook, in order to determine the rating for each dimension. In other words, use "everything that you know" to determine the gold standard ratings. [For the Spring 2005 ratings, each observer should work independently and do his/her own gold standard rating.]

2. When explaining your rating of each dimension, describe the evidence you used from both your observations (summary rating) and the notebook (student work, reflections, pictures, lesson plan, etc.) to determine your rating.
3. Two additional areas for rating should be included in the gold standard: notebook completeness and confidence.

Quick Reference Guide for CRESST Observation Ratings

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

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related scientifically and build on one another in a logical manner.
NOTE: The focus of this dimension is on design, rather than enactment. Ratings should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.

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

4. "Hands-On". The extent to which students participate in activities that allow them to physically engage with the scientific phenomenon by handling materials and scientific equipment.
NOTE: The emphasis is on direct observation and interaction with scientific equipment and physical objects, to address the substance of the science lesson. Acting out a scientific phenomenon does count. Computers don't unless use involves equipment such as probes.]

5. Inquiry. The extent to which the series of lessons involves the students actively engaged in posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.
NOTE: There is a "high bar" on this one. The focus is on the enactment of the lesson and student engagement. A key question is whether the unit/activity is designed so that all phases of inquiry are part of the unit, not whether we observe all phases during the Scoop days. To be true to the intent of this dimension, we should make inferences about the features of inquiry that are incorporated into the entire investigation.

6. Cognitive Depth. Cognitive depth refers to a focus on the central ideas of the unit, generalization from specific instances to larger concepts and connections and relationships among science concepts. There are two aspects of cognitive depth: the lesson design and teacher enactment. Thus, this dimension considers extent to which lesson design focuses on cognitive depth and the extent to which teacher consistently promotes cognitive depth.

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

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

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

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

11. Overall. How well the series of lessons reflect a model of instruction consistent with dimensions previously described. This dimension takes into account both the curriculum and the instructional practices.
NOTE: The rating on this dimension is implicitly a weighted average of the ratings on the first ten dimensions, with greater weight being given to Inquiry, Cognitive

Depth, Scientific Discourse Community, and Explanation/Justification to the extent that the rater felt he/she could rate these dimensions accurately.

For Gold Standard and Notebook Ratings:

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

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

Description of CRESST Notebook Rating Levels

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

High: Teacher designs activities to be done in groups that are directly related to the scientific goals of the lesson. The majority of students work on these activities in groups.

Example: The class is divided into groups, with each group focusing on a different planet. Students conduct research to design a travel brochure, describing the environment of their planet. Students are then reorganized into groups, with one student from each planet group in each of the new groups, to explore how the distance from the Sun affects characteristics of planetary environments such as the length of a day, the length of a year, temperature, weather, and surface composition.

Example: Students are divided into small groups to brainstorm how animals in different habitats are adapted to the unique features of their environments. Each group is considering a different environment (desert, mountain, woodland, etc). The class reconvenes to consider what characteristics of animals are important to examine when thinking about how an animal is adapted to its environment. Armed with the class list, students work in pairs to examine a spider and hypothesize about where this animal might live.

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

Example: In a unit on the solar system, each day the teacher delivers a lecture on the solar system, students read about it in their textbooks, and then work in groups of 3-4 to complete a worksheet.

Example: Students read about spiders in their textbook, and then they break into groups of 3-4 to study real spiders in terrariums. They return to their desks to complete a worksheet about their observations.

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

Example: The teacher delivers a lecture on the solar system, students read about it in their textbooks, and complete an individual worksheet.

Example: Students watch a video about the anatomy of spiders. They form into groups to practice for the upcoming state test by bubbling in answer sheets.

2. Structure of Lessons. The extent to which the series of lessons is organized to be conceptually coherent such that activities are related scientifically and build on one another in a logical manner.
NOTE: The focus of this dimension is on design, rather than enactment. Ratings should take into account interruptions for procedural activities that are not part of the instructional unit, when these interruptions consume a non-trivial amount of time.

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

Example: A unit of instruction on air pressure begins by engaging students through a provocative event in which they experience the effects of air pressure (trying to drink orange juice out of a cup through two straws in which one straw is placed outside of the cup). This activity includes opportunities for students to explore and raise questions about their experiences with the orange juice. The teacher then involves students in a logical sequence of experiments and class discussions about air pressure. Lessons culminate in conclusions or generalizations made through evidence gained during students' exploration of the effects of air pressure, current scientific explanations provided, and opportunities to apply their developing understanding of air pressure to new phenomena, events or activities.

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

Example: A unit of instruction on air pressure begins with the teacher explaining air pressure and its effect on our lives. The next day the teacher hands back a test on force and motion and the class discusses the results. Following that, the teacher involves students in a series of disjointed activities in which they experience or witness the effects of air pressure. Lessons culminate in opportunities for students to demonstrate what they have learned about air pressure.

Low: The series of lessons does not appear to be logically organized and connected.

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

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

High: The use of a variety of scientific resources forms a regular and integral part of instruction throughout the lesson/series of lessons. The lesson is meant to engage all students (e.g. teacher demonstration in which all students are watching). [NOTE: there are at least two categories of resources – scientific lab resources and print-based resources. Variety could be variety within each category or across the two categories.]

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

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

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

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

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

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

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

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

4. "Hands-On". The extent to which students participate in activities that allow them to physically engage with the scientific phenomenon by handling materials and scientific equipment.
NOTE: The emphasis is on direct observation and interaction with scientific equipment and physical objects, to address the substance of the science lesson. Acting out a scientific phenomenon does count. Computers don't count unless use involves equipment such as probes.

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

Example: As part of an investigation of water quality in their community, students bring water samples into class. They set up the appropriate equipment and measure the pH levels of the samples. In class the next day, students discuss how pH is related to water quality. The following day, they perform the same tests at a local stream and observe aquatic life in the stream.

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

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

Example: As part of an investigation of water quality in their community, the teacher brings water samples into class and sets up equipment to measure its pH. The teacher selects several students who then measure the pH levels of these water samples while the others observe. The following day, the teacher takes them outside to watch a few students test the pH of water in a local stream.

Example: As part of a discussion of plate tectonics, students model plate boundaries by acting them out with their bodies. Later in the unit, students supplement their reading about faults by using wooden blocks to represent different fault types.

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

Example: As part of a unit on water quality, the teacher brings water samples into class, sets up equipment to measure its pH, and performs the measurements while students observe.

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

5. Inquiry. The extent to which the series of lessons involves the students actively engaged in posing scientifically oriented questions, designing investigations, collecting evidence, analyzing data, and answering questions based on evidence.
NOTE: There is a "high bar" on this one. The focus is on the enactment of the lesson and student engagement. A key question is whether the unit/activity is designed so that all phases of inquiry are part of the unit, not whether we observe all phases during the Scoop days. To be true to the intent of this dimension, we should make inferences about the features of inquiry that are incorporated into the entire investigation.

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

Example: As part of a unit on motion, students are designing an amusement park. One group has chosen to work on a swinging Viking ship ride, and they are worried that the number of people on the ride (and their weight) will affect how fast the ride swings. They construct a simple pendulum and design an experiment to answer the question, "How does the weight at the end of a pendulum affect the amount of time it takes to complete ten swings?" They conduct the investigation and use the results to inform their design.

Example: The class has been discussing global warming. As a class, they decide to investigate how the temperature in their city has changed over the past 100 years. Students debate about what data they should gather, and different groups of students end up approaching the problem in different ways. The groups collect the data, analyze them, and present their results

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

Example: Students are asked, "What is the relationship between the length of a pendulum and the period of its swing? Between the weight at the end of the pendulum and the period?" To answer the questions, students follow a carefully scripted lab manual, taking measurements and graphing the data. They use their results to formulate an answer to the question.

Example: As part of a series of lessons on global warming, the teacher asks the students to show how the temperature of different cities has changed over the past 100 years. They select cities, gather data from the library, graph the information and report what they found.

EXAMPLE OF A "2" RATING:

Example: Students follow a carefully scripted lab manual to verify the formula for the period of a pendulum's swing given in a lecture the day before. They follow a carefully scripted lab manual, taking specific measurements and making specific graphs of their data. They conclude with answering factual questions in the lab manual.

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

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

Example: Students read in their textbook that the temperature of the Earth is rising x degrees per decade. At the back of the book, there is a table of data on which this statement was based. Following specific instructions, students graph this data to verify the statement in their book.

6. Cognitive Depth. Cognitive depth refers to a focus on the central ideas of the unit, generalization from specific instances to larger concepts and connections and relationships among science concepts. There are two aspects of cognitive depth: the lesson design and teacher enactment. Thus, this dimension considers extent to which lesson design focuses on cognitive depth and the extent to which teacher consistently promotes cognitive depth.

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

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

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

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

Low: The series of lessons focuses on discrete pieces of scientific information, e.g., disconnected vocabulary, definitions, formulas, and procedural steps. These are elements of science that can be memorized without requiring an understanding of the larger concepts. Teacher rarely attempts to engage students in instructional activities that demonstrate the connectedness of scientific concepts and principles. Teacher's interactions with students focus on correctness of their answers rather than on conceptual understanding.

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

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

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

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

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

High: Students consistently are encouraged to express their scientific reasoning to other students and the teacher, and they are supported by the teacher and other students in their efforts to do so. Students' ideas are solicited, explored, and

attended to throughout the lesson, and students consistently use appropriate scientific language. Emphasis is placed on making scientific reasoning public, raising questions and challenging ideas presented by classmates.

> Example: Students work in groups, investigating plant growth. The teacher moves around the room listening to their discussions and, at times, joining them. In answer to student questions, the teacher responds with suggestions or her own questions, keeping the focus on thinking and reasoning. Following the group work, students present their findings to the class. Classmates actively engage in a critique of each presentation by raising questions, challenging assumptions, and verbally reflecting on their reactions to the findings presented. The teacher asks probing questions, and pushes the scientific thinking of both presenters and peers. These discourse patterns appear to be the norm.
>
> Example: In a class discussion on the behavior of gases, the teacher asks students to share their thinking about why the diameter of a balloon increases when placed in hot water and decreases when placed in cold water. The teacher uses wait time to allow students to formulate their thinking. When students share their ideas, the teacher listens carefully and asks other students to reflect on, build on, or challenge the ideas presented by their classmates. Teacher may offer suggestions or alternative ways of thinking about the question when gaps in student thinking are evident, but does not engage in correcting students' ideas, or in giving the "real/right" answer.
>
> Example: During a lesson on cell structure and function, the teacher asks students to work in pairs on a lab activity. Their task is to determine the effect of a salt solution on green plant cells. Prior to the activity, each pair creates a hypothesis statement. They prepare their microscope slide and write down observations; describing the effects of salt and identifying various cell structures, and discuss the lab directed questions challenging each other's scientific reasoning and formulating their conclusions together.

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

> Example: Students work in groups, investigating plant growth. The teacher moves around the room listening to their discussions. When students stop her and ask questions, the teacher responds by providing suggestions or answers. Following the group work, students present their findings to the class. Their classmates listen to presentations, but do not ask questions, challenge results or react to the findings. The teacher tends asks questions to elicit both procedural and conceptual understanding from the presenters. The teacher supplements students' answers with content if it is missing from the presentations, or asks leading questions trying to prompt presenters into filling in the missing content.
>
> Example: In a class discussion on the behavior of gases, the teacher asks students to reflect on how air particles might be affecting the diameter of a balloon when it is moved from a bowl of hot water to a bowl of cold water. One student suggests that it has something to do with the air particles slowing down in the cold. The teacher responds to the student by saying "yes, and when the air particles slow down, they don't push against the balloon as much." Teacher follows this statement with a question like, "and how would that affect the diameter of the

balloon… if the air isn't pushing as hard, would the diameter of the balloon increase or decrease?" When most of the class responds with "decreases," the teacher goes on to ask, "So why then do you think the diameter of the balloon increases when we place it in a bowl of hot water?"

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

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

Example: Following an investigation on plant growth, the teacher holds a whole class discussion in which she asks students to recall important facts about plant growth that they learned in the process of their investigations. All of the teacher's questions have known answers, and teacher evaluates the "correctness" of each student response as it is given. If "correct" answers are not given, the teacher asks the question again or provides the answer.

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

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

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

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

Example: Following a whole class discussion on plate boundaries, the teacher poses a question for students to begin in class and complete for homework. The teacher asks the students to explain how the geologic features found near Nepal were created. Using maps in

the classroom one student indicates that there is a mountain range present in this region. The student compares a map of plate boundaries with a world map and points out that Nepal is located along a plate boundary. For homework, she uses data found from the Internet about the recent tectonic activity and is able to further her argument of converging plates with the data. The next day, she explains to the class, using her evidence from the maps and Internet search that two continental plate boundaries are converging to create mountains.

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

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

Example: Following a whole class discussion on plate boundaries, the teacher poses a question for students to begin in class and complete for homework. The teacher asks the students to explain how the geologic features found near Nepal were created. The student looks in her textbook and on the Internet to help answer the question. She finds a diagram of converging plate boundaries. The next day she shows this diagram to the class, as well as reads aloud the caption below the diagram. The teacher poses similar questions at the end of each lesson and students respond with similar concrete explanations.

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

One possibility for a "2" rating:
Teacher sometimes asks students to explain/justify their scientific reasoning, but students rarely provide such explanations.

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

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

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

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

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

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

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

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

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

Example: A week-long unit on chemical change involves three activities that are graded with teacher comments: a homework assignment, an in-class writing assignment, and an exam consisting of multiple choice items and one essay. Results count toward grades but are not

otherwise used. There is no evidence that the students were asked to revise any of the work based on the teacher's comments.

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

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

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

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

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

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

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

Medium: Teacher or students sometimes make connections between the science they are learning in class and their own experiences, OR the world around them. Students have some opportunities to learn to apply classroom science in contexts that are relevant to their own lives or to consider the role of science in society (for example, how science can be used to inform social policy). However, these opportunities occur only occasionally, or the examples are potentially relevant to

the students' own lives or to the role of science in society, but these connections are not made explicit.

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

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

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

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

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

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

FOR GOLD STANDARD AND NOTEBOOK RATINGS:

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

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

- Summary of content for the Scoop period
- Information about each day's lesson (content, instructional activities, materials used, student work in class, grouping of students for class work, homework assigned, and projects worked on)

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

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

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

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

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

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

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

Science SCOOP Rating

Rater: _____ Date: _____

Teacher: _____

1. Grouping	**(Circle one) 1 2 3 4 5**
Justification	

2. Structure of Lessons	**(Circle one) 1 2 3 4 5**
Justification	

3. Use of Scientific Resources	(Circle one) 1 2 3 4 5
Justification	

4. Hands-On	(Circle one) 1 2 3 4 5
Justification	

5. Inquiry	(Circle one) 1 2 3 4 5
Justification	

6. Cognitive Depth	(Circle one) 1 2 3 4 5
Justification	

7. Scientific Discourse Community	(Circle one) 1 2 3 4 5
Justification	

8. Explanation/Justification	(Circle one) 1 2 3 4 5
Justification	

9. Assessment	(Circle one) 1 2 3 4 5
Justification	

10. Connections/Applications	(Circle one) 1 2 3 4 5
Justification	

11. Overall	(Circle one) 1 2 3 4 5
Justification	

12. Notebook Completeness	(Circle one) 1 2 3 4 5
Justification	

13. Confidence	(Circle one) 1 2 3 4 5
Justification	