WORKING
PAPER

Reliability of Ratings of the Scoop Notebooks and Transcripts

BRIAN STECHER

WR-261-EDU

April, 2005

EDUCATION

Preface

This paper was presented at the Symposium "Using classroom artifacts to measure instructional practice in middle school science: a two-state field test" at the annual meeting of the American Educational Research Association, Montreal, Canada, April 15, 2005.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and improvement, U.S. Department of Education.

Introduction

The previous paper in this symposium provided background information on the Scoop research project and the approach we are using to try to develop rich descriptions of classroom practice in middle school mathematics and science (Creighton, 2005). The researchers collected data from 36 middle school science classrooms, including "Scoop" notebooks of artifacts collected by teachers, transcripts of audiotapes of classroom discourse, and direct observations of instruction. A framework for judging eleven dimensions of "reform oriented" classroom practice was developed, along with a detailed scoring guide to use to rate each type of artifact. These dimensions were used to rate observations, notebooks and transcripts. In addition, readers of notebooks and transcripts rated each source of information on "completeness" on the same one-to-five scale, and assigned a "confidence" score to their overall set of ratings. This presentation will focus on the procedures that were used to rate the notebooks, transcripts, and classroom observations and the reliability of the judgments that were made. This assessment of reliability is the first step in determining whether the notebooks and/or transcripts can serve as a reasonable surrogate for direct classroom observation. If these sources of information cannot be rated reliably, it is unlikely that they can prove to be a valid indicator of practice. The validity question will be examined in a subsequent paper (Borko, 2005).

The primary question that guided this analysis is how consistent are the ratings of notebooks and transcripts. For a variety of practical reasons, only one observer visited each classroom that participated in the study. Although the observer visited the classroom for the equivalent of three lessons, because there was only one observer, we cannot compute a quantitative indicator of interrater reliability for observations. Efforts were made to train observers (see below) to rate classrooms as accurately as possible.

Methods

Rating the Observations, Notebooks and Transcripts

As described in the previous paper, researchers identified 11 dimensions of reform-oriented science instruction that they believed could be measured using artifacts (Creighton, 2005). The dimensions were as follows:

- 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).
- 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).
- 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.
- Discourse Community. Extent to which the classroom social norms foster a sense of community in which students feel free to express their scientific ideas openly. 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.

- Explanation/Justification. Extent to which teacher expects and students provide explanations/justifications either orally or on written assignments.
- 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.)
- "Hands-On". Extent to which students participate in activities that allow them to physically engage with the scientific phenomenon by handling materials and scientific equipment.
- Inquiry. 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.
- 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.
- Structure of Lessons. 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.

- 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.

Each dimension was rated on a five-point scale from low (1) to high (5), and the same dimensions were used for rating the classroom observations and the notebooks. To facilitate the rating process, a guide was developed 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 were provided, as well (see Figure 1). Readers also rated the notebooks on a five-point "completeness" scale indicating how many of the requested elements were present.

Training Observers. To improve the quality of the observation ratings, training sessions were held prior to the classroom visits. The same procedures were followed at training meetings in California and Colorado. In both locations, the observer training process began with a review of the scoring guide; each observer read the guide and the embedded examples. Then a videotape of a middle school science lesson was shown, and each observer took free form notes while watching the tape. When the tape was completed, each observer rated the lesson using the guide, including writing descriptions of observed classroom occurrences that justified each rating. At the conclusion of this process, all ratings were posted on a chalkboard and the group discussed the set of ratings. The discussion continued until agreement was reached on the right rating for the lesson. In addition, 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 guide. The process was repeated with a second videotape. At that point we were satisfied that observers

Explanation/Justification. Extent to which teachers expect and students provide explanations/justifications that incorporate conceptual knowledge or the use of scientific evidence (e.g. data collected in a laboratory activity), either orally or on written assignments.

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

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.

Figure 1. Scoring Guide for Explanation/Justification

were able to apply the rating guide in a consistent manner, and they were permitted to begin actual observations.

Training Readers. After all scoop notebooks were collected and the audio tapes were transcribed, the researchers assembled at a single location to begin the process of rating the notebooks and transcripts. Seven readers participated in the study, and all were trained prior to rating. As in the case of the observations, training began with reading the scoring guide. Then three notebooks were chosen to use for initial calibration. Working individually, all readers rated each of the three notebooks. Then the readers met as a group to discuss the ratings. All the ratings 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 for notebooks was descriptions of the evidence that would be relevant to rating certain dimensions, such as discourse. There was not direct evidence of discourse in the notebook, but it was possible to make inferences based on teachers' daily reflections, assignments, and student work contained in the notebook. Guidance regarding this dimension was added to the scoring guide.

Once the initial ratings and discussion was completed, the process was repeated with another set of notebooks. After two rounds of calibration, it appeared that the readers understood the scoring guide and were applying it consistently. At that point, the official scoring study began. Transcripts were rated in the same manner as the notebooks, and there was not a separate training session for rating the transcripts. During the study, readers worked independently and on their own schedule. They were given a 7-10 day period to complete all ratings, and they checked out the notebooks and transcripts and rated them on their own.

Design for Scoring Study. Twenty-eight notebooks were used in the scoring study. Twenty came from classrooms without transcripts, and eight were selected from classrooms that had been audio-taped. The notebooks without transcripts were each read by three raters, as were the transcripts. The notebooks and transcripts were assigned to raters according to incomplete, but balanced, designs shown in Table 1. Each of the seven readers began at the top of the identified column and rated fifteen notebooks and/or transcripts in the order listed.

Table 1.
Design for Reading Notebooks and Transcripts

Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Reader 6	Reader 7
Atkinson	Baker (N+T)	Milton (N+T)	Douglas	Schmid	Lesner (TO)	Reginald
Elder	Lapp	Coolidge	Taylor (TO)	Lesner (N+T)	Sleeve (N+T)	Good
Saunders	Bennett	Shaker	Cook (N+T)	Newman	Taylor	Sleeve (TO)
Taylor (TO)	Shephard	Baker (TO)	Shafer	Garman	Beck	Jones (N+T)
Martin (N+T)	Cook (TO)	Kretke	Walters	Baker (TO)	Vonnne	Solaris
Lapp	Jones (TO)	Shafer	Garman	Milton (TO)	Good	Martin (N+T)
Baker (N+T)	Kretke	Douglas	Schmid	Taylor	Milton (TO)	Atkinson
Shephard	Milton (N+T)	Cook (N+T)	Baker (TO)	Beck	Solaris	Lesner (TO)
Sleeve (TO)	Shaker	Walters	Newman	Sleeve (N+T)	Jones (N+T)	Saunders
Bennett	Coolidge	Lesner (TO)	Lesner (N+T)	Vonnne	Reginald	Elder
Douglas	Newman	Martin (TO)	Solaris	Martin (N+T)	Baker (N+T)	Shaker
Shafer	Martin (TO)	Vonnne	Jones (N+T)	Elder	Cook (TO)	Coolidge
Walters	Schmid	Taylor	NA	Jones (TO)	Lapp	Milton (N+T)
Cook (N+T)	Garman	Beck	Good	Atkinson	Bennett	NA
Milton (TO)	Lesner (N+T)	Sleeve (N+T)	Reginald	Saunders	Shephard	Kretke

Note: TO = Transcript only; N+T = notebook only, followed by notebook and transcript together; NA = not applicable

The notebooks with transcripts were rated in two different ways. First, each transcript was read and rated without reference to the notebook. Then, the same reader reviewed the notebook and made a second rating on the basis of the information contained in both sources. This provided transcript only ratings for eight classrooms, and transcript plus notebook ratings

for the same classrooms (due to a procedural mistake, only seven of eight were re-rated). The data sources and number of raters are summarized in Table 2.

Table 2.
Sources of Information about Reform-Oriented Classroom Practice

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

Analysis

Two approaches were used to estimate reliability. First, we compared each pair of ratings directly and determined the level of agreement for each dimension. 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 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, considering raters and classrooms as random facets. These analyses were conducted separately for each dimension (because the dimensions were selected purposefully from a universe that we considered to be limited). SAS PROC VARCOMP was used to estimate variance components and then generalizability coefficients were estimated for designs using two, three or four raters. Both absolute and relative g-coefficients 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.

Results

Reliability of Notebook Ratings

We found moderate agreement among notebook readers on most of the dimensions (see Table 3). 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%).[1] All the results in Table 3 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.

Table 3 also reveals differences in reliability among the dimensions. Six dimensions had exact agreement at or above 40% and eight had within one agreement at or above 80%. In fact, with the exception of Explanation/Justification all of the agreement results were fairly similar. It is interesting to note that the differences in agreement among the dimensions were not as large as in a previous study of middle school mathematics teachers (Borko, et al., 2004). 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 subtle 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. It may also be the case that by

[1] These estimate 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.

taking averages across 28 classrooms we have obscured differences in rating reliability that are related to individual teaching styles. That will be investigated below.

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).

Table 3.
Percent Agreement in Notebook Ratings (3 Raters) by Dimension (28 Classrooms)

Dimension	Average Score	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

As suggested above, we found larger variation in agreement across classrooms than we did across notebooks (see Table 4). For example, readers of the notebooks from 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.)

Table 4.
Percent Agreement in Notebook Ratings (3 Raters) by Classroom (11 Dimensions)

Classroom	Average Score	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

In addition, rater agreement for notebooks was not a function of the average score of the notebook, i.e., the presence or absence of reform-oriented practices. Some notebooks with low average scores had high levels of agreement (e.g., Milton), while others had lower levels of agreement (e.g., Sleeve).

Percent agreement is not the only way to analyze rating consistency. Generalizability theory allows us to analyze simultaneously the variation in ratings among teachers and raters and to determine how much is associated with these two factors and how much is unsystematic or random error or another source of systematic but unmeasured variance, such as occasions. The first step in a generalizability analysis is to partition the variance among notebooks, raters, and residual/error. Table 5 shows the percent of variance due to each facet for each dimension. High generalizability occurs when most of the variance is associated with classrooms, and little is associated with unmeasured facets (residual). Variance associated with raters can be reduced by adding raters.

Table 5.
Percent of Estimated Notebook Rating Variance Attributed to Each Facet by Dimension (28 Classrooms, 3 Raters Per Notebook)

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%	0.8%	38.2%
Hands-On	59.6%	2.0%	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%

Fortunately we do not have to interpret the results in Table 5 directly, we can use them to compute specific generalizability coefficients for designs using different numbers of raters. Table 6 shows the generalizability coefficient that would be applicable if we were to use two, three or four raters for each notebook. Relative decisions are those that are based on ranking classrooms along the dimension (i.e., putting them in order) but not considering their absolute score. Absolute decisions are ones in which the absolute score on the five-point scale is of interest, not just the relative standing.

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 any other dimension. Using three raters, all dimensions except Structure of Lessons (and perhaps Explanation/Justification and Assessment) can be rated to a reasonable level of consistency for relative decisions (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 there might not be much variance among teacher to detect. Also, the reader's knowledge of subject matter may influence their interpretation of the logical connections between the topics covered during the scoop period.

For absolute decisions, two other dimensions-- Assessment and Explanation/Justification--also fall below a 0.80 threshold. Assessment was more 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 referred to it with a one or two comments in their reflections, but readers did not always note or credit these comments. Explanation/Justification may have been

more difficult to rate on the basis of the notebooks because it is manifest, in large part, in teachers' and students' verbal behavior, which is not apparent in most notebooks. The aspects that are evident from the notebook, e.g., "why" questions in written assignments, are easy to overlook when rating the notebooks.

Table 6.
Notebook Rating Generalizability Coefficients for Absolute and Relative Decisions Using Two, Three or Four Raters by Dimension
(28 Classrooms, 3 Raters Per Notebook)

	Relative Decisions			Absolute Decisions		
Dimension	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

Reliability of Transcript Ratings

In a smaller sample of classrooms, we also collected audiotapes, which were transcribed to provide a supplemental source of information on classroom practices. In these seven classrooms separate ratings were done on the basis of the transcripts alone, and on the basis of the transcripts combined with the notebooks. To find out the quality of these transcript ratings we conducted similar analyses of levels of agreement. The smaller sample size makes these results themselves less conclusive, i.e., we have less confidence in the estimates reported in the tables.

Ratings based on transcripts alone were more reliable than ratings based on notebooks on some dimensions and less reliable on others (see Table 7). Relatively higher agreement was achieved for Explanation/Justification and Discourse. 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. However, transcripts reveled 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 that students and teachers engage in. This suggests that a combination of notebooks and transcripts may achieve higher reliability than either source alone. However, data presented below show that is not necessarily the case.

Table 7.
Percent Agreement in Transcript Ratings (2 or 3 Raters) by Dimension (8 Classrooms)

Dimension	Average Score	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

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.

There was considerable variation in the level of agreement in transcript ratings summarized by classroom (see Table 8). In some cases (e.g., Sleeve) the notebooks and transcripts had comparable levels of agreement. However, for most classrooms the levels of agreement for notebooks and transcripts were different.

Table 8.
Percent Agreement in Transcript Ratings (2 or 3 Raters) by Classroom (11 Dimensions)

Classroom	Average Score	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

We conducted a generalizability analysis for transcript ratings, although there were too few raters 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 9). This is consistent with the results obtained in the previous analysis.

Table 9.
Transcript Rating Generalizability Coefficients for Absolute and Relative Decisions Using Two, Three or Four Raters by Dimension
(8 Classrooms, 2 or 3 Raters Per Transcript)

	Relative Decisions			Absolute Decisions		
Dimension	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

Reliability of Notebook Plus Transcript Ratings

It appears that combining information from notebooks and transcripts does not yield results that are more reliable than notebooks alone. Table 10 shows the level of rater agreement by dimensions for the combination of these two data sources, and Table 11 shows the agreement by classroom. This negative result may be due, in part, to the smaller number of classrooms that were included in the combined analysis. It may also be due to the fact that the two sources of information may provide conflicting points of view on a dimension, which readers resolve differently. 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.

Table 10.
Percent Agreement in Notebook Plus Transcript Ratings (3 Raters) by Dimension (7 Classrooms)

Dimension	Average Score	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

Comparing only those seven classrooms included in Table 11 with the same classrooms in Table 5 shows similar levels of agreement.

Table 11.
Percent Agreement in Notebook Plus Transcript Ratings (3 Raters) by Classroom (11 Dimensions)

Classroom	Average Score	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

Conclusions

The results indicate that notebooks can be rated reliably on almost all dimensions using three readers. Readers had the greatest problem rating Structure of Lessons, which suggests that the scoring guide needs to be improved for this dimension, the dimension itself is not well conceptualized, or notebooks are not an adequate mechanism for measuring this dimension.

Readers also found it difficult to rate the notebooks of a few of the teachers, such as Elder and Jones. On inspection, these notebooks contain fewer artifacts than the typical notebook completed during this field study. While completeness does not appear to be a major factor in rating consistency, it appears that some teachers were more likely than others to assemble artifacts that were easy to rate. Perhaps better directions would overcome this problem, although the vast majority of teachers followed the directions correctly.

For example, we might be able to improve the instructions so that teachers were more forthcoming and provided detailed reflections. We included examples of "rich" reflections in the instructions, so teachers had models to follow. We also permitted teachers to submit reflections in writing, on audiotape or as computer files. In these ways we attempted to make the process flexible enough to accommodate different styles of composition. We even supplied tape recorders to teachers who wanted to use that method but did not have access to equipment. The problem does not seem to be method as much as differences in the tendency to discuss, the length of commentary, and the kinds of insights teachers have and are willing to share. Perhaps we could have heightened teachers' motivation to put more energy into reflections with larger incentives. However, we provided teachers with a generous honorarium ($250) for participating, for this exact reason. We tried to test the notebooks under the most favorable conditions. Yet, even under these circumstances some teachers were very limited in their reflective comments.

We do not think additional remuneration would have made a big difference for some teachers, and greater incentives would be impractical in most large-scale research studies.

Finally, we could have trained teachers better. In the case of science, we met with teachers individually or in small groups for approximately one-half hour to review the notebook instructions. It is not clear to us how the training could have been improved, but training is another area in which improvement is possible.

Transcripts were not rated as reliably as notebooks on a number of dimensions, although our test of transcripts was more limited in scope. However, transcripts were effective sources of information for Discourse, which was difficult to rate on the basis of notebooks alone. One of the problems with rating the transcripts was that the quality of the recordings was poor. While we were able to hear everything the teacher said, it was often impossible to hear students' responses. This was the best we could do with wireless equipment that could be used simply and without complicated apparatus and support personnel. Having "half" the conversation was helpful, but not as helpful as it might have been to have the complete verbal exchanges.

Oddly enough, the combination of notebooks and transcripts was better for some dimensions, including those where evidence was found in verbal exchanges, but not others. The number of cases in which we had both sources of information was small (only seven classrooms) and we do not believe this provided a strong test of their effectiveness.

Overall, this study suggests that artifacts can be collected in a systematic manner and scored consistently enough to be used to compare science teaching across classrooms. However, the Scoop Notebook imposes a moderate burden on teachers, and additional research should be conducted to determine whether the process can be streamlined and still permit consistent judgments. The next paper in this seminar looks at the correspondence between ratings based on

different data sources (observations, notebooks, transcripts) to see whether the impressions drawn from notebooks provide valid inferences about instructional practices. Subsequent papers examine the use of the notebooks as tools for professional development.

References

Borko, H. (2005). Validity of artifact packages for measuring instructional practice. Paper presented at the Symposium "Using classroom artifacts to measure instructional practice in middle school science: a two-state field test" at the annual meeting of the American Educational Research Association, Montreal, Canada, April 15, 2005.

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., and McClam, S. (Forthcoming.) Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment*.

Creighton, L. (2005). Artifact packages for measuring instructional practice: Introduction. Paper presented at the Symposium "Using classroom artifacts to measure instructional practice in middle school science: a two-state field test" at the annual meeting of the American Educational Research Association, Montreal, Canada, April 15, 2005.

Wood, A. Data collection methods and procedures: The scoop notebook. Paper presented at the Symposium "Using classroom artifacts to measure instructional practice in middle school mathematics: a two-state field test" at the annual meeting of the American Educational Research Association, San Diego, April 15, 2004.