Looking Into Students' Science Notebooks: What Do Teachers Do With Them?

CSE Technical Report 562

Maria Araceli Ruiz-Primo, Min Li, and Richard J. Shavelson
CRESST / Stanford University

April 2002

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

Project 1.1. Models-Based Assessment Design: Individual and Group Problem Solving—Science
Richard Shavelson, CRESST/Stanford University, Project Director

The work reported herein was supported in part under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education and in part by the National Science Foundation (ESI-9910020).

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

LOOKING INTO STUDENTS' SCIENCE NOTEBOOKS: WHAT DO TEACHERS DO WITH THEM?

Maria Araceli Ruiz-Primo, Min Li, and Richard J. Shavelson
CRESST/Stanford University

Abstract

We propose the use of students' science notebooks as one possible unobtrusive method for examining some aspects of teaching quality. We used students' science notebooks to examine the nature of instructional activities they encountered in their science classes, the nature of their teachers' feedback, and how these two aspects of teaching were correlated with students' achievement. We examined the characteristics of students' science notebooks from 10 fifth-grade classrooms. Six students' notebooks in each classroom were randomly selected. Each entry of each student's science notebook was analyzed according to the characteristics of the activity, quality of student's performance as reflected by the notebook entry, and the teacher feedback in the notebook. Results indicated that (a) raters can consistently classify notebook entries despite the diversity of the forms of communication (written, schematic or pictorial). They can also consistently score the quality of a student's communication, conceptual and procedural understanding, and the quality of a teacher's feedback to the student. (b) The intellectual demands of the tasks required by the teachers were, in general, low. Teachers tended to ask students to record the results of an experiment or to copy definitions. (c) Low student performance scores across two curriculum units revealed that students' communication skills and understanding were far from the maximum score and did not improve over the course of instruction during the school year. And (d) teachers provided little, if any, feedback. Only 4 of the 10 teachers provided any feedback to students' notebook entries, and when feedback was provided, comments took the form of a grade, checkmark, or a code phrase. We concluded that the benefits of science notebooks as a learning tool for students and as a source of information for teachers were not exploited in the science classrooms studied.

The success of science education reform relies on the quality of instruction that takes place in the class. It is expected that the opportunities to learn science that students have be appropriate, meaningful, and rich (see National Research Council, 1996). Classroom observations are the most direct way to assess the quality of teaching. However, this is an expensive and time-consuming method.

In this paper we propose the use of students' science notebooks as an unobtrusive method to explore some aspects of the quality of teaching. We think that teachers should consider science notebooks a natural strategy, among others, to monitor their students' progress. The notebooks should reflect, at least partially, the instructional activities carried out in class. If teachers communicate to students their progress and encourage them to improve their learning, at least some evidence of this communication should be found in the notebooks. If teachers adjust their instructional practices based on the information gained as they monitor student progress, these adjustments should also be partially reflected in the students' notebook entries.

In this study we examined the characteristics of students' science notebooks from 10 fifth-grade classrooms. Six students' notebooks in each classroom were randomly selected. Each entry of each student's science notebook was analyzed according to the characteristics of the activity, quality of the student's performance as reflected by the entry, and teacher feedback. Based on this analysis, we describe the types of entries most commonly found across classrooms and provide information about how the characteristics of the notebook entries and teacher feedback are related to students' learning.

On Students' Science Notebooks

We defined a science notebook (Ruiz-Primo, 1998) as a compilation of entries (or items in a log) that provide a partial record of the instructional experiences a student had in her or his classroom for a certain period of time (e.g., unit of study). Baxter, Bass, and Glaser (2000) provided evidence that science notebooks reflect with great fidelity what students do and what teachers focus on in science classes. Since notebooks are generated during the process of instruction, the characteristics of their entries vary from entry to entry as they reflect the diversity of activities in a science class. In their notebooks, then, students may describe problems they tried to solve, procedures they used, observations they made, conclusions they arrived at, and their reflections. Notebooks are viewed mainly as a written account, in more or less detail and with diverse quality, of what students do and, hopefully, learn in their science class.

This study is part of a larger effort to evaluate the feasibility of using science notebooks as an assessment tool (see Ruiz-Primo, Li, Ayala, & Shavelson, 1999, 2000). Notebooks as an assessment tool can be considered at two levels: (a) At the *individual level,* they may provide evidence bearing on student performance over the course of instruction; and (b) at the *classroom level,* they may provide evidence of opportunities students have to learn science, including both exposure to the science content students have to learn as specified in the curriculum / program adopted, and the quality of teacher feedback on the students' performance as observed in their notebooks.

From these two levels we can provide three measures: (a) *Unit implementation* —What intended instructional activities were implemented as reflected in the student's notebooks? Were any other additional activities implemented appropriate to achieving the unit goal? (b) *Student performance*—Were students' communications in the notebooks complete, focused, and organized? Did students' communications indicate conceptual and procedural understanding of the content presented? (c) *Teacher feedback*—Did the teacher provide helpful feedback on students' performance? Did the teacher encourage students to improve their scientific communication? (See Ruiz-Primo et al., 1999, and Ruiz-Primo, Li, et al., 2000, for detailed information.)

Documentation about the *implementation* of science activities can be found in different forms: reports of hands-on activities, reports and / or interpretations of results, predictions, reflections about the activity, and the like. To be able to draw conclusions about the activities implemented, information across individual notebooks within a class is aggregated. If none of the students' notebooks from a class had any evidence that an activity was carried out, most likely the activity was not implemented. A *student's performance* can be assessed from an analysis of the student's notebook entries (e.g., the student's notes, written reports, diagrams, data sets, explanation of procedures or results reported). Each notebook entry is evaluated according to the *quality of the communication* (e.g., Did a student's communication correspond to the appropriate communication genre?) and the *conceptual and/or procedural understanding* reflected in the communication (e.g., Did a student's explanation apply the concepts learned in the unit correctly? Did the student's description provide accurate examples of a concept? Was the student's inference justified based on relevant evidence?). Finally, evidence on *teacher feedback* can be found in teachers' comments in the students' notebooks.

Results of previous studies (Ruiz-Primo et al., 1999; Ruiz-Primo, Li, et al., 2000) indicated that (a) students' science notebooks could be reliably scored. *Unit implementation, student performance,* and *teacher feedback* scores were highly consistent across scorers. (b) Inferences about unit implementation using notebooks were justified. A high percent of agreement with independent sources of information (e.g., teachers' unit logs and teachers' verification lists) on the instructional activities implemented indicated that the unit implementation score was valid for this inference. (c) Inferences about students' performance were also very encouraging. High and positive correlations with performance assessment scores indicated that the notebook performance score could be considered as an achievement indicator. And (d) teacher feedback scores helped to identify teacher feedback practices across classrooms. Based on these results, we tentatively concluded that notebooks provided reliable and valid information on student performance and opportunity to learn.

In this study we focused on the types and characteristics of the entries observed in the students' science notebooks drawn from 10 classrooms. Specifically, we focused on the appropriateness of notebook entries and types of teacher feedback in the context of teaching and learning science and level of students' performance.

Method

Student Notebooks

Eight schools in a medium sized urban school district in the Bay Area of California participated in the study with 10 teachers/classrooms and 60 fifth-graders. All 10 teachers/classrooms implemented two Full Option Science System ([FOSS], 1993) units: "Variables" in the fall, and "Mixtures and Solutions" (henceforth Mixtures) in the spring. All teachers reported that they regularly used science notebooks in their science classes. No directions were provided to them on how to use science notebooks or the characteristics notebooks should have. Teachers were asked to sort their students into five ability groups—from the top 20% to the bottom 20%—according to science proficiency. Students' notebooks were collected at the end of the school year. For this study, two students each from the top-, middle-, and low-proficient students were randomly selected from each class.

Each student in the sample had two notebooks, one for Variables, generated during the fall, and another for Mixtures, generated during the spring. A total of 120 science notebooks (1,804 pages), comprising 60 Variables notebooks (961 pages) and 60 Mixtures notebooks (843 pages), were analyzed in this study. For each student, information about performance assessment scores on a pretest and posttest for each unit was obtained (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2000). Effect sizes based on the pretest and posttest performance assessment scores were calculated by classroom for each unit.

Analysis of Science Notebook Entries

Notebooks are a compilation of communications with diverse characteristics. Each of these communications is considered as a *notebook entry,* which can be a set of definitions, a set of data, an interpretation of results, a description of an experiment, or a quick note about what was learned in the class on a particular day. The characteristics of notebook entries vary since each entry may ask students to complete different tasks depending on the instructional activity implemented on a particular day (e.g., write a procedure or define a concept). After reviewing dozens of students' science notebooks from different classrooms (Ruiz-Primo et al., 1999), and looking into the types of activities that students were supposed to do in a science class (see National Research Council, 1996), we identified 13 general entry categories: Defining, Exemplifying, Applying Concepts, Predicting/Hypothesizing, Reporting Results, Interpreting Results and/or Concluding, Reporting & Interpreting Results and/or Concluding, Reporting Procedures, Reporting Experiments, Designing Experiments, Content Questions/Short Answers, Quick Writes (e.g., reflections), and Assessments. Each type of entry was considered to have its own set of characteristics that make it identifiable. For example, reporting results focuses on the description of observations or presentation of data, whereas interpretation focuses more on summarizing and generalizing from the data, or highlighting specific cases (e.g., Penrose & Katz, 1998).

Some of the categories proposed also include subtypes of entries according to the form or the characteristics of the communication. Notebook entries can be found in different forms of communication: verbal—written/text—(e.g., explanatory, descriptive, inferential statements); schematic (e.g., tables, lists, graphs); or pictorial (e.g., drawing of apparatus). For example, a definition can be verbal (e.g., defining, verbally, the pendulum system) or pictorial (e.g., drawing the pendulum system); therefore, the type of entry *definition* includes both subtypes of definitions. Reporting procedures has three subtypes: A procedure can be found as a "recount" description (e.g., I put the screen over the cup . . .), as an instruction (e.g., you put the screen on the top of the cup . . .), or as a direction (e.g., put the screen on the top of the cup . . .). Including the subtypes of entries defined, we had a set of 23 categories.

Each notebook entry was coded at two levels. First, a code was used to identify the type of entry (e.g., an entry in which a procedure-recount was reported was coded as "11"). Once the type of entry was identified, a set of second-level codes was used to define the characteristics of the entry. Second-level codes were of three types: (a) the characteristics of the investigations/experiments reported in the entry, if appropriate (e.g., replications of the experiments were implied, or more than one level of the independent variable was studied, or both); (b) the format of the entry (e.g., only students' responses are found, without a formal prompt; or format of the entry is provided by teachers or curriculum developers); and (c) general characteristics of the entry (e.g., the entry was repeated in another part of the notebook, or had a supplemental picture/graph, or content of entry was clearly copied from textbook). For example, an entry could be coded as 15—reporting an experiment—and the code ".3" could be added (i.e., 15.3) if the entry had evidence that replications of the experiment/investigation were done. Also, the entry could have an additional code, for instance, ".6" (i.e., 15.3.6), if the format of the entry was provided to the students (e.g., a printed sheet on which to report the experiment).

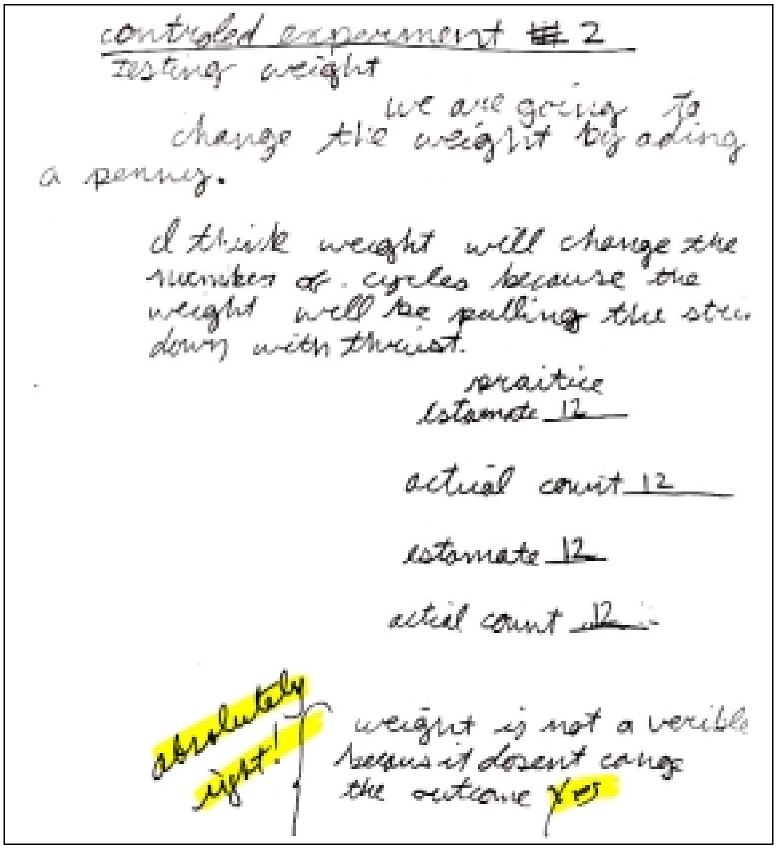
Once the type and the characteristics of a notebook entry were identified, we scored quality of communication, conceptual understanding and/or procedural understanding if appropriate, and teacher feedback for each journal entry. Quality of communication was evaluated on a 4-point scale: 0—*incoherent and not understandable communication* (e.g., incomplete sentences); 1—*understandable but not using the characteristics of the genre* (e.g., examples are provided, but the category to which the examples belong is not provided); 2—*understandable and uses some of the basic characteristics of the genre* (e.g., category to which the examples belong is provided, but only in the form of a title, not making the logical relationship explicit); and 3—*understandable and uses all the basic characteristics of the genre* (e.g., category to which the examples belong is provided and makes the logical relationship explicit). If a student's communication was scored zero, we did not attempt to score the student's understanding.

Conceptual and procedural understanding were evaluated on a 4-point scale: 0—*no understanding* (e.g., examples or procedures described are completely incorrect); 1—*partial understanding* (e.g., relationships between concepts or descriptions of observations are only partially accurate or are incomplete); 2—*adequate understanding* (e.g., comparisons between concepts or descriptions of a plan of investigation are appropriate, accurate, and complete); and 3—*advanced understanding* (e.g., communication focuses on justifying responses/choices/decisions based on the concepts learned, or the communication provides relevant data/evidence to formulate the interpretation); plus (NA)—*not applicable* (i.e., instructional task does not require any conceptual or procedural understanding).

We assessed the quality of teacher feedback by using a 6-level score: –2—*feedback provided, but incorrect* (e.g., teacher provides an A+ for an incorrect notebook entry); –1—*no feedback, but it was needed* (e.g., teacher should point out errors / misconceptions / inaccuracies in student's communication); 0—*no feedback;* 1—*grade or code phrase comment only;* 2—*comment that provides student with direct, usable information about current performance against expected performance* (e.g., comment is based on tangible differences between current and hoped performance, "Don't forget to label your diagrams!"); and 3—*comment that provides a student with information that helps her to reflect on/construct scientific knowledge* (e.g., "Why do you think it is important to know whether the material is soluble for selecting the method of separation?").

To explain how we approached the analysis of students' notebooks we present in Figure 1 an example of a student's notebook entry. The entry was clearly linked to Experiment 2 (i.e., testing weight) of the Swingers activity in the FOSS Variables unit. The entry was classified as "Reporting an Experiment" because it had important elements of a report: (a) a title, "Testing weight"; (b) an hypothesis, "I think weight will change the number of cycles because the weight will be pulling the string down with thwist [sic]"; (c) a procedure, "we are going to change the weight by ading [sic] a penny"; (d) results, "practice estamate [sic] 12, actual count 12…"; and (e) conclusions, "weight is not a verible [sic] becaus [sic] it doesn't cange [sic] the outcome." Furthermore, the entry provided evidence that two replications were carried out during the experiment. The entry shows the teacher's written comments. Both can be found in the conclusion part (i.e., "absolutely right!" and "yes").

The code for this entry was 15.3. The first part of the code indicates that the entry is reporting an experiment, and the second part, ".3," indicates that there is evidence of replications. The student's performance was scored on the two aspects described above, quality of communication, and conceptual / procedural understanding. The communication quality of this student's report was poor. For example, the procedure was not replicable since the description was incomplete; it is not clear how the outcome was measured, and subtitles were missing. Furthermore, for this student a variable is only a variable if it has an effect on the outcome; if it does not, then the variable studied is not a variable! The teacher feedback score focused on both the quality of the teacher's comment to the student's communication and the student's understanding. It was clear that the student needed feedback from the teacher not only for improving the quality of the



controled experiment #2
Testing weight

we are going to change the weight by ading a penny.

I think weight will change the number of cycles because the weight will be pulling the stri down with thrust.

praitice
estamate 12

actual count 12

estamate 12

actiel count 12

absolutely right!

weight is not a verible becaus it dosent chang the outcome Yes

Figure 1. An example of a student's science notebook entry for the Swingers activity of the FOSS Variables unit.

communication of the experiment, but also for helping the student understand the concept of variable. Teacher's feedback to this student was scored as "provided but incorrect" because the teacher rewarded the student's response despite the evidence of the student's misunderstanding. Student's quality of communication was scored as 1, understanding as 1, and teacher feedback as –2.

Two independent scorers analyzed the notebooks. Scorers were experts in the unit activities. Students' notebooks within units were mixed and randomly ordered for scoring. Scorers were unaware of the classroom to which a student belonged or her performance level (top, middle, or low, according to teacher ranking).

Results

The report of results focuses on (a) describing and analyzing the types of entries most frequently found in students' science notebooks, (b) describing teacher feedback notebook practices, and (c) looking for associations between types and quality of entries and teacher feedback with students' learning. We first present evidence about the agreement on the classification of entries and the reliability of the scores between the two scorers.

Agreement and Reliability Across Raters

Eighteen (236 notebook pages) of the 120 notebooks were used to train and calibrate scorers, and 50 notebooks (774 notebook pages) were used to evaluate interrater agreement and reliability. We evaluated interrater agreement in classifying notebook entries according to the type and characteristics, and interrater reliability for each type of score (student performance and teacher feedback) across units. After training, 23 notebooks from Variables and 29 from Mixtures were scored independently by two raters. Table 1 provides the percentage of agreement across units and the reliability coefficients by type of score and unit.

Results indicated that interrater agreement in type of entry was above 85% and consistent across the two units. Scorers could consistently identify and classify notebook entries despite the variety of the characteristics of the entries across students and classrooms.

The magnitude of the interrater reliability coefficients varied according to the type of score. However, coefficients were never below .80. We concluded that raters were consistent in identifying the type of entry across students' notebooks and that they consistently scored students' performance and quality of teacher feedback.

Table 1

Percent of Agreement for Type of Entry and Interrater Reliability Across Types of Scores

		Interrater reliability				
		Student performance			Teacher's feedback	
Unit	% of agreement / type of entry	Quality of communication	Conceptual understanding	Procedural understanding	Communication	Understanding
Variables	85.47	.86[a]	.88[b]	.80	.80[c]	.91
Mixtures	85.18	.81[a]	.89	.82[a]	.94	.88[c]

[a] Three outliers were dropped from the sample, two of them with very similar scores within raters.

[b] Two outliers were dropped from the sample.

[c] One outlier was dropped from the sample.

Types of Notebook Entries

For characterizing types of entries, within each classroom the total number of entries across the students' notebooks was counted, entries were classified, and percentage by type of entry was calculated. Percentages for each type of entry were averaged across classes (Table 2).

We found that most of the science notebook entries, across all the classrooms and units, were pertinent/appropriate to the learning of science. Less than .5% of the entries were classified as entries unrelated in any way to the units' goals—the *Don't care about activity* category.

Although the profiles of types of entries varied from classroom to classroom, in all classrooms and across both units the types of entries most frequently found were *reporting data* (34.99%), *definitions* (18.98%), and *content questions/short answer* (15.05%). The types of entry least frequently found were *designing experiments* (0.12%), and *reporting and interpreting data* (0.78%). On the rest of the categories the percentage varied across the two units. Only in three classrooms did we find evidence that formal assessments were provided at the end of the units. All of the assessments found were classified as simple forms of assessment (e.g., short-response, matching exercises).

Table 2

Percentage of Type of Entry by Unit

Type of Entry	Variables (*n* = 60)	Mixtures (*n* = 60)
Defining	20.58	17.38
Exemplifying	6.90	0.99
Applying concepts	1.80	4.30
Predicting/hypothesizing	1.15	0.90
Reporting results	32.34	37.63
Interpreting results/concluding	5.02	2.29
Reporting and interpreting results/concluding	0.95	0.61
Reporting procedures	2.78	4.33
Reporting experiments	6.34	3.07
Designing experiments	0.20	0.05
Content questions/short answers	8.87	21.29
Quick writes-reflections, affective questions	8.91	5.81
Assessments	3.69	0.98
Don't care about activity	0.48	0.33

Some of the inconsistencies in the percentage across the two units might be due to the content of the units. For example, the activities in which students were involved in the Variables unit were more likely to ask students for predictions (e.g., predict the number of cycles given a certain length of the string) than the activities in the Mixtures unit, which focused more on physical and chemical changes of substances. However, there is no clear justification about the difference on the percentages across the two units in certain types of entries. For example in both units it was possible to apply the concepts learned, to report, interpret, and make conclusions, or to report an experiment.

Based on these results it was clear that students did not have many opportunities in their classrooms to engage in important forms of communication (e.g., interpreting data) that might help them improve understanding (e.g., evidence to support explanations, design experiments). Comparing percentages, students recorded data about seven times more than understanding what the data meant. If students are not asked to make the appropriate connections between the data, the evidence they provide, and the conclusions, how can they learn to interpret a data table? How can they learn that conclusions can be validated only when evidence is provided to support them? Moreover, students have many concept definitions in their notebooks, but few entries show that they are required to apply those concepts (e.g., relating, contrasting, comparing, justifying, or explaining the concepts). An instructional activity such as asking students to interpret data and draw conclusions provides an opportunity to assess whether students are understanding the concept at hand (see Figure 1). Unfortunately, these opportunities were hardly found.

Table 3 provides percentages for those entry categories that have subcategories (see Method section). As expected, most of the definitions were verbal and very few pictorial. Pictorial definitions were mainly found in the Variables unit. However, one wonders whether students’ understanding could be improved if they had to represent more concepts, when appropriate, in both modalities, verbal and pictorial (Shavelson, Carey, & Webb 1990).

Forms of reporting data, verbally or graphically, varied according to the unit. Graphical data were mainly found in the Variables unit, but the opposite was true in the Mixtures unit. We believe this is congruent with the content of the units. The data students collected in the Variables unit were more suitable to being organized and represented in a table or a graph (string length on the horizontal axis and number of cycles on the vertical axis) than those in the Mixtures unit, in which most

Table 3

Percentage of Type of Entries by Unit and Sub-Category

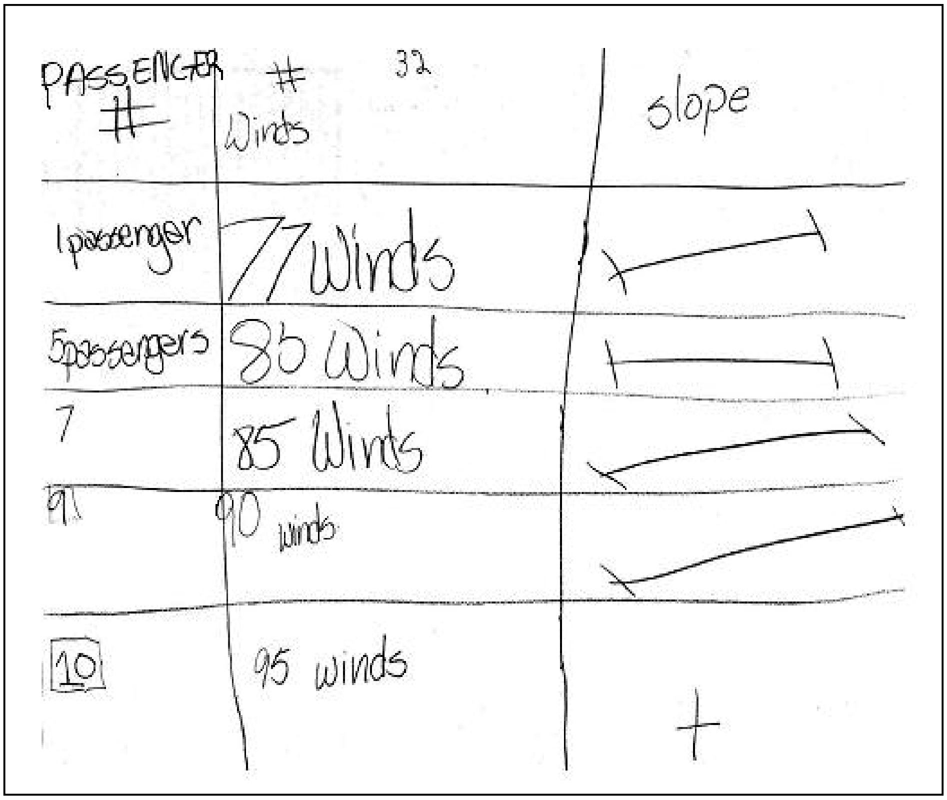
Type of entry	Variables (n = 60)	Mixtures (n = 60)
Defining-verbal	19.80	17.30
Defining-pictorial	0.77	0.08
Reporting results-verbal	5.66	30.05
Reporting results-graphic	26.68	7.58
Reporting verbal results & interpreting / concluding	0.48	0.44
Reporting graphic results & interpreting / concluding	0.47	0.17
Reporting procedures-recount	2.41	2.18
Reporting procedures-instructions	0.33	1.42
Reporting procedures-directions	0.03	0.73
Reporting experiments-complete	2.01	0.14
Reporting experiments-incomplete	4.33	2.93
Quick writes-reflections	3.93	1.32
Quick writes-affective questions	0.15	0.04
Assessments-simple forms	3.69	0.98
Assessments-complex forms	0.00	0.00

of the data collected were observations (e.g., description of the salt crystal left after evaporation).

Unfortunately, most of the procedures reported across the two units were classified as *narrative recounts procedures,* instead of instructions or directions. This finding is important. Research (e.g., Martin, 1989, 1993) has found that narrative descriptions of procedures in science (e.g., Today we put two cups of water . . .) decreased generalizability of the procedure since they are presented as a recollection of events and in past tense; therefore, they typically express temporality. The accurate and appropriate reporting of procedures is essential to the work of scientists, as is reporting experiments. However, we found not only that on few occasions were students asked to report an experiment, but also that most of the experiments reported were incomplete. They usually lacked the purpose / objective of the experiment or data interpretation and conclusions.

These findings indicated that the development of students' scientific literacy and understanding was restricted as a consequence of the type of instructional activities implemented (Marks & Mousley, 1990). Instructional activities were not cognitively demanding, and they were limited in the use of scientific written genres.

Furthermore, we found that instructional activities also reflected the level of the teachers' content knowledge. Figure 2 provides a student's notebook entry that shows what the students did in one classroom for the Plane Sense activity in the FOSS Variables unit. It is important to mention that all the students' notebooks within the same class had pretty much the same entry. In this activity, students built a plane with jumbo straws, rubbers bands, a propeller, popsicle sticks, and a hook and made the plane fly on a string taped between two chairs. Students were supposed to conduct controlled experiments to determine which variables affected how far the plane flew on the line. Students worked with variables such as passengers (i.e., number of clips), inclination of the flight line, the number of winds in the propeller, and the like. Look carefully at the student's notebook entry. Notice that the table in the notebook entry shows that the student varied all the variables—number of passengers, winds, and the slope—at the same time on each trial. Remember that all students' notebook entries in this class were exactly the same.



PASSENGER #	# 32 Winds	slope
1passenger	77 Winds	
5passengers	85 Winds	
7	85 Winds	
9	90 winds	
10	95 winds	+

Figure 2. A student's notebook entry for the Plane Sense activity of the Variables unit.

What can one infer about this teacher's knowledge of variables, controlled variables, and controlled experiments based on this activity? Are these the opportunities students are given to learn science? Further, what did students do after this "learning" experience? There was no interpretation of the data or any conclusion after this entry. Of course, interpretation and conclusions could be discussed in class. Still, why was there no record of the conclusions? Students' understanding in an entry like this one was hard to assess since the instructional activity was incorrectly designed. In the next section we explore in more detail the characteristics of the students' notebook entries.

Characteristics of the Notebook Entries

Table 4 provides the percentage of entries with a particular characteristic (e.g., nature of investigation, format) coded. It is important to mention that some characteristics were necessarily linked to certain types of entries. For example, *replications implied* was a characteristic that was applied only to those entries in which the nature of the information provided allowed the scorer to infer that more than one trial was carried out to test the consistency of the results. This information could not be obtained in entries that focused, for example, on definitions, but could be found in entries that described a procedure, provided data, or reported an experiment.

The characteristics *replications implied*, *experiment design 1*, and *experiment design 2* focused on the quality of the investigations students carried out in their science class. We found that only 1.98% of the entries (average across the three characteristics and the two units) showed evidence that replications and the manipulation of more than one value of the variables were considered in the experiments. We acknowledge that it is not always possible to conduct replications or to study different values of the manipulated variables. However, there were certain experiments in which the systematic replications were critical for understanding the concept at hand (e.g., variables) or the relevance of certain process skills (e.g., carefulness in measuring and manipulating the variables). How can students understand better the concept of a variable if they only manipulate one value of a variable? How can students understand the importance of care in manipulation when their results differed from those expected? How can students understand that whenever scientists carefully measure any quantity many times, still they expect that most measurements will be close but not exactly the same? We

Table 4

Characteristics of Entries in Percentage

Characteristic	Variables (*n* = 60)	Mixtures (*n* = 60)
Characteristics of the investigations carried out		
Replications implied. More than one trial was carried out within the same experiment.	4.76	0.67
Experiment design 1. More than one level of the independent variable was manipulated, but no replications were implied.	2.02	0.27
Experiment design 2. More than one level of the independent variable was manipulated and replications were implied.	4.09	0.04
Format of the entries		
Informal entry. A scientific communication was embedded in an informal narrative (e.g., What we did today in class?).	5.81	1.98
Without a prompt. No questions or formal prompts were found in the notebook, only responses. Applies only to content questions and quick writes.	5.79	3.82
Entry format provided. Format provided to the student by the curriculum developers (activity sheets) or by the teacher.	29.49	39.30
General characteristics of the entries		
Supplemental picture/graph provided. Applied to any entry.	5.15	7.25
Repeated notebook entry. Without any Improvement. Applied to any entry.	3.54	3.60
Copied definitions. Content of definition was clearly copied from textbook or dictionary.	72.93	66.13

believe that there were critical experiments across the two units in which replications and the manipulation of more than one value of the variables were necessary for providing students with more meaningful and rich instructional experiences.

Informal entry referred to scientific communications that were embedded in informal narratives. For example, the students reported a procedure in a narrative describing “What I Liked About Today’s Science Class.” Unfortunately, in some classes this type of “quick writes” was the only evidence found about the implementation of an instructional activity. However, the percentage of this characteristic was not very high. Notice that relevant information learned in the science class could be lost, or at least not valued as it should be, if the emphasis is put on what students liked about the class, instead of what was learned and how that learning could be formalized for better understanding. We believe it is more meaningful to ask students to describe the procedure learned and to explain the conditions in which the procedure is applicable, than only to ask students to report what they liked about the class. Of course, this does not imply that it is inappropriate having quick-write entries. What is important is also to have entries focusing on formal communications (e.g., describing a procedure).

The characteristic *without a prompt* referred to those entries without a title or a prompt that could be used as a hint on what was the entry about. This characteristic was applied only to entries that focused on content questions and quick writes, which meant that only students' responses were found as notebook entries. If students' notebook entries do not have a title or a prompt that can help students to track what they have done, how can they use this information later as a reference? How can these entries be useful later in the course? Although the percentage of this characteristic was not high (4.81% on average across the two units), it would be better if it were non-existent.

Notebook entries varied not only according to the type of the entry, but also according to who provided the format: curriculum developers, the teacher, or the student. For example, a science unit/module adopted by a teacher/class/district had, as part of its design, activity sheets that students filled out as they carried out the activity. At other times the teacher provided students with an entry format; for example, a table for reporting data or a printed sheet on which to report an experiment. When the format was provided by the curriculum developers or the teachers we coded the entry as *entry format provided.* Unfortunately, a high percentage of the entries (34.40% averaged across the two units) had to be coded as having format provided. Most of the entry formats were provided by the curriculum developers, and only a few were provided by the teachers. Furthermore, all the activity sheets that accompanied the units focused mainly on recording data without requiring students to interpret the data or make conclusions. We already have discussed the impact of this type of activity on students' performance and understanding. Still, there is another issue to consider when students are provided with the entry formats: the lack of experience in organizing and presenting information according to the scientific genres and the audience at hand (e.g., Is a notebook entry for a student's own reading, for the teacher to read, or for sharing with peers? What are the characteristics of the scientific written genre most appropriate for a particular communication?).

On average across the two units, 6.20% of the entries had a *supplemental picture or graph.* In most of the entries the supplemental pictures or graphs were related to the content to the entry. However, they were not necessarily useful. That is, they did not add any information to the content of the entry, nor were they helpful for better understanding the topic at hand.

We found that 3.57% of the entries, averaged across the two units, were exactly repeated notebook entries and not revised entries in which students provided an improved communication based on the teachers' comments. Some entries were repeated up to four times. Most of them were definitions. We wondered how students could benefit from repeating/copying the same entry on more than one occasion without further elaboration of the entry?

Finally, we found that 69.53% (averaged across the two units) of the definitions in the students' notebooks were copied from the science textbook or from a dictionary. Further development of the concepts defined was rarely found (e.g., students could apply the concept by providing examples, or by relating or comparing it to, or contrasting it with, other concepts). Whether a discussion about the concepts took place in the class is hard to know. We believe that if concepts were discussed, some evidence should be found in the students' notebooks so students could use that information later in the course to justify answers in an argument, or in providing a conclusion.

Assuming notebook entries reflect what students do in their science class (and they do—see Baxter et al., 2000; Ruiz-Primo et al., 1999; Ruiz-Primo, Shavelson, et al., 2000), a natural question to ask is whether the type of instructional activities in which students were involved is associated with the level of their performance. In the next section we provide evidence about this issue.

Types of Entries and Students' Performance

To evaluate whether there was a relation between types of entries and students' performance we used the students' scores obtained in the performance assessments administered before and after instruction for each unit (Ruiz-Primo, Shavelson, et al., 2000). The effect size was calculated using all students in the classrooms, not only the students sampled for this study. Students took, in a pretest-posttest design, the Pendulum performance assessment for the Variables unit and the Saturated Solutions performance assessment for the Mixtures unit (see Appendix). We calculated the effect size for each classroom. Table 5 provides the effect size for each class across the two units. Based on the magnitude of the effect sizes obtained in each unit, we blocked the ten classrooms into three levels, high, medium, and

Table 5

Effect Sizes by Class Across the Two Units

Class	Variables	Mixtures
1	.95	1.60
2	1.11	1.62
3	.60	2.21
4	.48	2.17
5	.53	.59
6	.60	.66
7	.57	1.68
8	.32	1.16
9	.62	1.37
10	.73	1.16
Average	.65	1.30

low.[1] Effect sizes varied across classrooms within and across units. Notice also that the ranking of the classes varied between the two units; some classes did better in the Variables unit, others in the Mixtures unit.

Table 6 provides the percentage of type of entries according to the ranking of the classes by unit. On both units, the high-ranked classes had fewer definitions than the middle- and low-ranked classes. Notice that low classes had the highest percentage of definitions. High classes had the highest percentage on exemplifying and applying concepts, which indicates that students elaborated more on the concepts instead of merely defining them. The lower the rank of the class, the lower the percentage of exemplifying and applying concepts.

The same pattern was found for the categories *reporting, interpreting data,* and *concluding*. High-ranked classes had more entries in which students interpreted the data collected and drew conclusions than middle- and low-ranked classes. However, we were expecting high classes to have more entries focusing on reporting experiments than the other two groups, and this was the case only for the Mixtures unit. Furthermore, the low classes had the highest percentage on this category in the Variables unit.

[1] Examination of effect sizes provides clues to actual pretest-posttest change free of sample size limitations. An effect size of .20 is considered small, .50 is considered medium, and .80 large. Formula:

$$X' = {X_{post} - \overline{X}_{pre}} \Big/ {SD_{pre}}$$

Table 6

Percentage of Type of Entry by Class Level Across Units

	Variables[a]			Mixtures[b]		
Type of entry	High (*n* = 12)	Middle (*n* = 30)	Low (*n* = 18)	High (*n* = 18)	Middle (n = 30)	Low (*n* = 12)
Defining	16.64	21.14	22.27	16.42	17.05	19.66
Exemplifying	14.52	5.68	3.85	1.76	0.79	0.34
Applying concepts	4.14	1.18	1.28	5.50	4.48	2.04
Predicting / hypothesizing	1.20	0.49	2.22	1.17	1.11	0.00
Reporting data	26.11	38.60	26.07	31.27	38.67	44.57
Interpreting data / concluding	4.35	6.41	3.13	2.77	2.37	1.36
Reporting, interpreting, and concluding	1.96	0.70	0.69	0.80	0.73	0.00
Reporting procedure	4.02	2.10	3.08	3.99	5.18	2.71
Reporting experiments	5.41	5.28	8.74	5.11	2.25	2.04
Designing experiments	0.00	0.40	0.00	0.18	0.00	0.00
Content questions / short answers	11.62	4.59	14.18	20.58	20.14	25.24
Quick writes	9.32	9.44	7.73	6.90	6.74	2.04
Assessment	0.00	3.33	6.76	3.27	0.00	0.00
Don’t care about activity	0.72	0.67	0.00	0.27	0.50	0.00

Note. Ranking of classes is based on the magnitude of the effect sizes.

[a] Classes 1 and 2 were classified as High; classes 3, 6, 7, 9, and 10 as Middle; classes 4, 5, and 8 as Low.

[b] Classes 3, 4, and 7 were classified as High; classes 1, 2, 8, 9, and 10 as Middle; classes 5 and 6 as Low.

It is clear that the pattern of types of entries found in students’ notebooks was not always in the direction we expected or consistent across units. However, we believe we have provided some evidence that students whose notebooks had more meaningful entries performed better on the performance assessments than students whose notebooks had less meaningful and rich entries. In the next section we present information about the students’ notebook performance and teacher feedback scores.

Students’ Performance Scores and Teacher Feedback Scores

The students’ performance was scored on the two aspects described above, *quality of communication* (i.e., Does the student’s communication correspond to the appropriate scientific genre, in this case, to reporting an experiment?), and *conceptual/procedural understanding* (e.g., Did the student interpret the data correctly? Was the student’s conclusion justified based on the evidence?). The *teacher feedback* score focused on the quality of the teacher’s comment on both the student’s communication and the student's understanding. Mean scores are provided in Table 7.

Based on the mean scores it is clear that students' notebook entries across the 10 classrooms did not provide evidence of high quality of scientific communication, no matter the type of entry at hand. Notebook entries were understandable. They had some of the distinctive elements of the written genre, so the scorers were able to classify the entries (e.g., definitions had the concept and the meaning of it). However, students rarely used technical terms or the suitable grammatical characteristics according to the genre (e.g., use present tense in defining a concept). Nor did they use the appropriate genre structure and characteristics (e.g., in reporting an experiment, they did not provide the experiment purpose or conclusion). For example, we found that most of the data reported in the students' notebooks were strings of numbers without any organization at all (e.g., 12, 12, 12; meaning number of cycles on three trials), or a very rudimentary form of organization (e.g., 28 – 12, 25 – 13; the first number is the length of the string, and the second is the number of pendulum cycles). As an example of this, see Figure 3.

Figure 3 provides an example of a student's notebook page with data recorded. The rudimentary table presented in the page does not have any label. Therefore, it is hard to interpret "1, 12" unless the reader is familiar with the activity. The student is reporting the number of cycles (12) obtained over three trials. If the student were actually reporting data, a title would be needed, and the table columns would need

Table 7

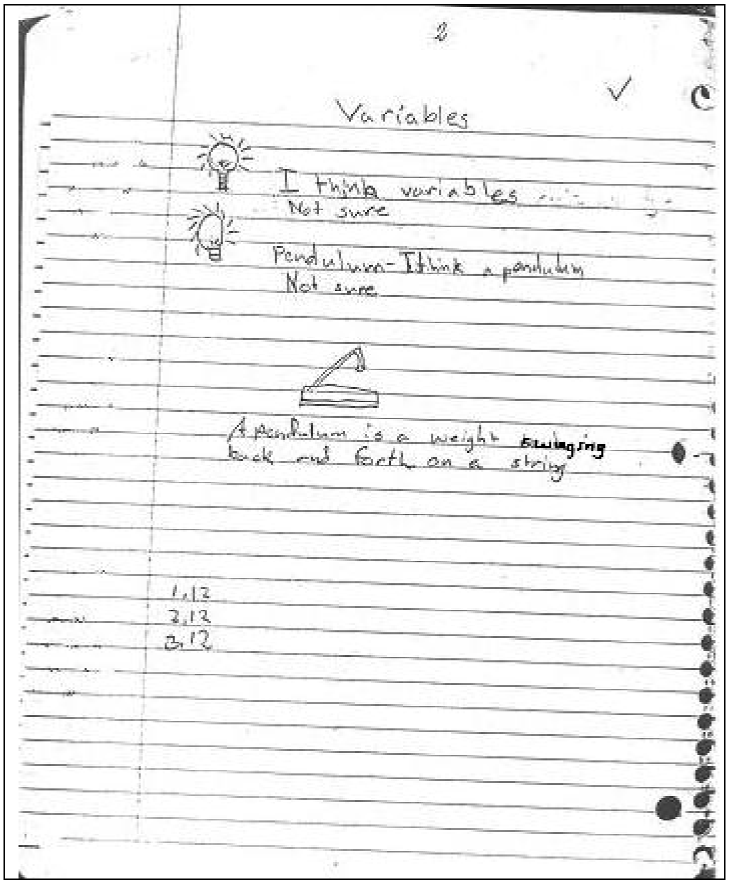
Mean Scores and Standard Deviations for Communication and Understanding

		Student performance						Teacher feedback			
		Quality of communication		Conceptual understanding		Procedural understanding		Communication		Understanding	
Unit	*n*	Mean[a]	*SD*	Mean[a]	*SD*	Mean[a]	*SD*	Mean[a]	*SD*	Mean[a]	*SD*
Variables	60	1.30	.30	1.49[b]	.55	1.32	.34	.07	.16	.10	.50
Mixtures	60	1.21	.31	1.10[c]	.48	1.20	.30	.01	.05	.06	.43

[a] Max = 3.

[b] There were no notebook entries that focused on conceptual understanding in two classrooms for the Variables unit, nor in two students' notebooks in a third classroom. Therefore *n* = 44 for this calculation.

[c] There were no notebook entries that focused on conceptual understanding in three students' notebooks for the Mixtures unit. Therefore *n* = 57 for this calculation.



2

✓ C

Variables

I think variables ...
Not sure

Pendulum- I think a pendulum
Not sure

A pendulum is a weight swinging
back and forth on a string

1,12
2,12
3,12

Figure 3. A student’s notebook entry for the Swingers activity of the Variables unit.

to be labeled so the numbers in them could be interpreted appropriately. In terms of the student’s procedural understanding, the teacher should ask the student to provide an interpretation of and conclusion for these results. In this case, if interpretation of the data and discussion of conclusions were done with the whole class, there is no notebook record of the conclusion at which they arrived. If this further step (i.e., interpreting and concluding) was not done, what was the purpose of writing numbers? This student’s notebook did not have any subsequent entry in which such information was provided. It is difficult for science notebooks with entries like this one to be used by students as sources of useful information and as a means for them to track their thinking and reasoning behind a scientific conclusion.

Mean scores for conceptual and procedural understanding indicate the partiality of students’ knowledge. It is important to remember that most of the students’ notebook entries that focused on conceptual understanding were mainly definitions. As mentioned, notebook entries rarely focused on providing explanations and examples, applying the concepts learned, or defining a concept in a different way than in the textbook. For those definitions coded as copied, understanding was not scored since we decided it was difficult to find any evidence of whether or not the student understood the concept at hand. Still, the mean score on conceptual understanding indicates that students' communications that focused on conceptual understanding (e.g., providing examples of solutions) were only partially correct (e.g., some of the examples provided by the student were incorrect).

Procedural understanding mean scores also indicated that students' entries did not provide accurate, appropriate, and complete information in their communications. This was the case with the entries that focused on scientific process skills (e.g., interpreting data) and with the entries that focused on describing the implementation of a known procedure (e.g., screening a mixture).

Now consider what the teachers did to improve students' understanding and performance. Did they provide appropriate feedback to students? Was there any written evidence of this feedback? If there was no written evidence, is it possible to infer that they provided verbal feedback to students? We found that in 6 of the 10 classrooms studied, there was no evidence of teacher feedback in any notebook entry across the two units despite the fact that students' entries revealed poor communication and partial understanding (see Ruiz-Primo, Li, & Shavelson, 2001).

We explored in more detail the quality of teacher feedback in those four classes in which feedback was provided. We found that feedback focused mainly on students' understanding. In spite of its importance, quality of communication was basically ignored across the two units by all the teachers. Some of the teachers paid attention to spelling errors, but not to the quality and the characteristics of the students' written communications (for details see Ruiz-Primo et al., 2001).

Figure 4 shows the percentage of types of feedback teachers provided across the students' notebook entries for students' understanding. We have explained the 6-level scoring scale used (–2 to 3). However, the figure has two additional categories, *inconsistent* and *not necessary*. The former refers to those cases in which, for the same notebook entry, the teacher provided correct and incorrect feedback (e.g., –1 and +2). The latter refers to those entries for which both (a) the teacher did not provide feedback and (b) feedback was not necessary based on the characteristic of the entry (e.g., a copied definition).

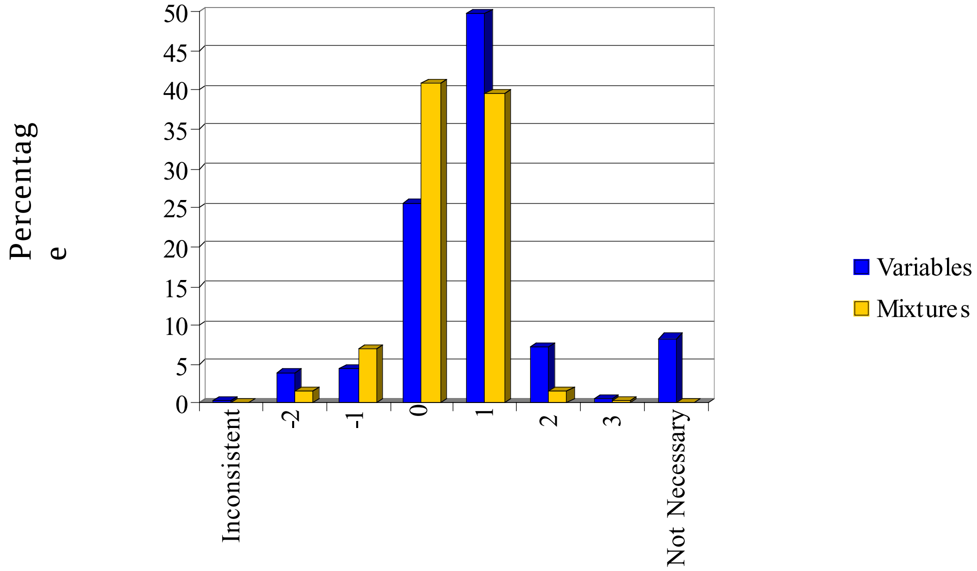


Figure 4. Percentage of types of feedback focusing on students' understanding used by the four teachers in the Variables and Mixtures units.

The type of feedback with the highest percentage across the four classrooms was type 1—grade, checkmark, or code phrase (44.59% averaged across the two units). Literature on feedback (e.g., Black & Wiliam, 1998; Sadler, 1989, 1998) has emphasized that providing only a grade, a checkmark, or a general comment cannot help students reduce the gap between where they currently are and where they should be (the learning goal).

We found very low percentages of helpful comments across the four classes (averaged across the two units, 4.41% for type 2—usable comments to improve performance, and .49% for type 3—comments that provide students with information that helps to reflect on/construct scientific knowledge). Unfortunately, we also found that 3.73% (over the two units) of feedback provided by teachers was incorrect (e.g., teacher provided an A+ to an incorrect response, a –2 score). Furthermore, 5.74% (over the two units) was type –1 (i.e., teachers did not provide feedback but they should have). As expected, a high percentage (33.15% over the two units) was type 0 (no feedback provided, indicating that the four teachers did not provide feedback for approximately one third of the entries in their students' notebooks).

Despite the general agreement that science notebooks allow teachers to assess students' understanding and provide the feedback students need for improving their performance (e.g., Audet, Hickman, & Dobrynina, 1996; Dana, Lorsbach, Hook, & Briscoe, 1991; Fellows, 1994; Heinmiller, 2000; Hewitt, 1974; McColskey & O'Sullivan, 1993; Shepardson & Britsch, 1997), we concluded that in our sample, there was not enough evidence to show that teachers had taken advantage of using notebooks as an assessment tool.

Some readers may argue that due to time constraints, teachers could not provide written feedback but did read the students' notebooks and provided verbal feedback to the whole classroom, or even individually. In another analysis (see Ruiz-Primo et al., 2001), we tested this hypothesis by comparing students' performance over the course of the school year. We acknowledged that the content of the science units was different. Therefore, we focused on quality of communication and procedural understanding, assuming that these two aspects should improve from the beginning to the end of the school year, independent of the content. For example, if students do not know how to report a procedure at the beginning of the school year, they should know how at the end of the year. If they only provide conclusions without using any evidence at the beginning of the year, we want them to learn that conclusions can only be validated when evidence is provided to support them. In sum, if appropriate verbal feedback was provided to students in the classes, they should have performed better, at least in the quality of their communications.

To test this hypothesis we carried out two occasion-by-class MANCOVAs controlling for reading scores, one for quality of communication score and the other for procedural understanding score (for details see Ruiz-Primo et al., 2001). Results indicated a significant interaction between occasion and class for the quality of communication score (Hotelling's $T = .63$, $p = .003$). This means that in some classes, students did better in the Variables unit, and in other classes, students did better in the Mixtures unit. Unfortunately, we found that in most classes, the quality of communication mean score was lower for the Mixtures unit than for the Variables unit: Students did better at the beginning of the year than at the end. The picture for the procedural understanding score was exactly the same. If any verbal feedback was provided in the classes, this was not an effective feedback that could help students improve their quality of communication or their understanding.

Conclusions

In this study, we used students' science notebooks to examine the nature of the instructional activities they did in their science classes, the nature of their teachers' feedback, and how these two aspects of teaching related to the students' performance.

Each entry of each student's science notebook was analyzed according to the characteristics of the activity, quality of student performance, and teacher feedback. Results indicated that (a) raters could consistently classify notebook entries despite the diversity of the forms of communication (written, schematic or pictorial). They also could consistently score students' quality of communication, conceptual and procedural understanding, and the quality of teachers' feedback. (b) The intellectual demands of the tasks required by the teachers were, in general, low. Teachers tended to ask students to record the results of an experiment or to copy definitions. These types of tasks by themselves can hardly help students to improve their understanding. (c) Low student performance scores across the two curriculum units revealed that students' communication skills and understanding were far from the maximum score and did not improve over the course of instruction during the school year. And (d) this latter result may be due, in part, to the fact that teachers provided little, if any, feedback. Indeed, teachers did not provide feedback despite the errors or misconceptions that were evident in the students' performance. Therefore, there was no effort to close the gap between student performance at the time a notebook entry was produced and the desired performance. Results indicated that only four of the ten teachers provided any feedback to students' notebook entries. When feedback was provided, comments were reduced to a grade, checkmark, or a code phrase. We concluded that the benefits of science notebooks as a learning tool for the students and as a source of information for teachers were not exploited in the science classrooms studied.

Our findings may be useful in suggesting areas for professional development. What we learned about teachers' use of science notebooks does not speak very highly about the quality of science instruction students are getting. It is clear that keeping a science notebook is a widespread science teaching practice; however, its use is neither effective nor efficient. Results indicated that writing in science notebooks was mainly mechanical. For almost every instructional activity, students were asked to write down the results found, for example, by recording data or observations. The quality of the students' communications was poor. Procedures were hardly replicable. Results were almost never organized in a way that could help students find patterns and were rarely used as evidence in conclusions, when a conclusion could be found at all.

Science notebooks can assist students' thinking, reasoning, and problem solving if used appropriately. The ongoing accurate and systematic documentation of the development of ideas, concepts, and procedures is a powerful scientific tool for replicating studies, for discussing and validating findings, and developing models and theories; in sum, for developing scientific inquiry. To this end, we believe that teachers need first to carefully select the type of entry to work on with students. Which types of entries are more useful and effective for promoting understanding, scientific inquiry, and improving performance? Second, teachers need to think of options for assisting themselves and helping students move toward self-monitoring (e.g., self- and peer-assessment; Sadler, 1989). Third, educators and researchers need to think carefully about how science notebooks can be conceptualized, implemented, and assessed in a form that most effectively reflects their main purpose.

References

Audet, R. H., Hickman, P., & Dobrynina, G. (1996). Learning logs: A classroom practice for enhancing scientific sense making. *Journal of Research in Science Teaching, 33,* 205-222.

Baxter, G. P., Bass, K. M., & Glaser, R. (2000). *An analysis of notebook writing in elementary science classrooms* (CSE Tech. Rep. No. 533). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7-74.

Dana, T. M., Lorsbach, A. W., Hook, K., & Briscoe, C. (1991). Students showing what they know: A look at alternative assessments. In G. Kulm & S. M. Malcom (Eds.), *Science assessment in the service of reform* (pp. 331-337). Washington, DC: American Association for the Advancement of Science.

Fellows, N. (1994). A window into thinking: Using student writing to understand conceptual change in science learning. *Journal of Research in Science Teaching, 31*, 985-1001.

Full Option Science System. (1993). *Britannica Science System*. Chicago, IL: Encyclopaedia Britannica Educational Corporation.

Heinmiller, B. (2000). Assessing student learning—and my teaching—through student journals. Retrieved from the Eisenhower National Clearinghouse Web site: www.enc.org/focus/topics/assessment/articles/a06/index.htm

Hewitt, H. C. (1974). The journal. *Science and Children, 11*(8), 30-31.

Marks, G., & Mousley, J. (1990). Mathematics education and genre: Dare we make the process writing mistake again? *Language and Education, 4*, 117-135.

Martin, J. R. (1989). *Factual writing: Exploring and challenging social reality*. Oxford: Oxford University Press.

Martin, J. R. (1993). Literacy in science: Learning to handle text as technology. In M. A. K. Halliday & J. R. Martin (Eds.), *Writing science: Literacy and discursive power* (pp. 166-202). Pittsburgh, PA: University of Pittsburgh Press.

McColskey, W., & O'Sullivan, R. (1993). *How to assess student performance in science: Going beyond multiple-choice tests*. Tallahassee, FL: Southeastern Regional Vision for Education (SERVE).

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

Penrose, A., & Katz, S. (1998). *Writing in the sciences*. New York: St. Martin's Press.

Ruiz-Primo, M. A. (1998). *On the use of students' science journals as an assessment tool: A scoring approach*. Unpublished manuscript, Stanford University, School of Education.

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (1999, March). *Students' science journals and the evidence they provide: Classroom learning and opportunity to learn*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston.

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (2000, April). *Students' science journals as an assessment tool*. Paper presented at annual meeting of the American Educational Research Association, New Orleans.

Ruiz-Primo, M. A., Li, M., & Shavelson, R. J. (2001, March). *Exploring teachers feedback to students' science notebooks*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Saint Louis, MO.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2000). *On the evaluation of systemic science education reform: Searching for instructional sensitivity*. Paper submitted for publication.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sadler, R. D. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5,* 77-84.

Shavelson, R. J., Carey, N. B., Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71,* 692-697.

Shepardson, D. P., & Britsch, S. J. (1997). Children's science journals: Tool for teaching, learning, and assessing. *Science and Children, 34*(5), 13-17, 46-47.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessment in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19,* 1-14.

Appendix

Description of the Performance Assessments

In the *Pendulum* assessment students are asked to identify the variable that affects the time it takes a pendulum to complete 10 cycles (Stecher & Klein, 1997). Students explore the relationship between the length of a string, the weight of the suspended object, and the periodicity of a pendulum. The scoring system focused on the correctness of the variable identified, the accuracy of the measurements, and the interpretation of the data and the accuracy of the prediction required.

The *Saturated Solution* assessment asked students to find out which of three powders was the most and the least soluble in 20 ml of water. Students are asked to provide information about how they conducted the investigation, the results they obtained, and two other questions about solubility (e.g., how they can dissolve the maximum possible powder in a saturated solution). The scoring system focused on the accuracy of the results and the quality of the procedure used to solve the problem.