INT. J. SCI. EDUC., 8 OCTOBER 2004,
VOL. 26, NO. 12, 1477–1506

Taylor & Francis
Taylor & Francis Group

RESEARCH REPORT

Evaluating students' science notebooks as an assessment tool*

Maria A. Ruiz-Primo, School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305–3096, USA; e-mail: aruiz@stanford.edu; Min Li, 312D Miller Hall, College of Education, University of Washington, Seattle, WA 98195, USA; e-mail: minli@u.washington.edu; Carlos Ayala, Sonoma State University, 1801 East Cotati Avenue, Rohnert Park, CA 94928, USA; e-mail: carlos.ayala@sonoma.edu; Richard J. Shavelson, School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305–3096, USA; e-mail: richs@stanford.edu

The idea of using science notebooks as a classroom assessment tool is not new. There is general agreement that science notebooks allow teachers to assess students' conceptual and procedural understanding and to provide the feedback students need for improving their performance. In this study we examined the use of science notebooks as an unobtrusive assessment tool that can also be used by individuals outside the classroom (for example, school district personnel), and as a means for obtaining information about students' learning and their opportunities to learn. More specifically, in this study students' science notebooks were used as a source of data about the (a) implementation of a curriculum's intended activities, (b) students' performance, and (c) quality of teachers' feedback. Our results indicated that: (1) Students' science notebooks can be reliably scored. Unit implementation, student performance, and teacher feedback scores were highly consistent across raters and units. (2) High and positive correlations with other performance assessment scores indicated that the student performance score can be considered as an achievement indicator. And (3) low performance scores across the two units revealed that students' communication skills and understanding were far away from the maximum score and did not improve over the course of instruction during the school year. This result may be due, in part, to the fact that no teacher feedback was found in any of the students' notebooks across the six classrooms studied. This may reflect some characteristics of the teachers' assessment practices that may require further professional development.

Introduction

Science notebooks or journals are seen as a log of what students do in their science class. (We prefer the former term, notebook, to stay away from the meaning of science journal as diary.) Keeping science notebooks encourages students to write as a natural part of their daily science class experience. Students may describe the problems they are trying to solve, the procedures they use, observations they make, conclusions they arrive at, and their reflections. As expected, there are several variations on this basic idea. (For example, thinking journals (Lozaukas and Barell 1992), affirmational dialogue journals (Hanrahan 1999), log sheets (Lucido 1992),

* The report herein was supported by the National Science Foundation (No. SPA-8751511 and TEP-9055443). The opinions expressed, however, are solely those of the authors.

International Journal of Science Education ISSN 0950–0693 print/ISSN 1464–5289 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0950069042000177299

dialectical journals, think-aloud journals, and team journals (see Rivard 1994).) Still, science 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. In this study we defined notebooks as a compilation of entries (or items in a log) that provide a record, at least partially, of the instructional experiences students have in their classroom for a certain period of time (for example, unit of study). Since notebooks are generated during the process of instruction, the characteristics of their entries vary from entry to entry as they reflect the diverse set of activities in a science class.

There is general agreement that science notebooks allow teachers to assess students' understanding (for example, Audet et al. 1996, Dana et al. 1991, Fellows 1994, Heinmiller 2000, Hewitt 1974, McColskey and O'Sullivan 1993, Shepardson and Britsch 1997) and to provide the feedback students need for improving their performance (for example, Audet et al. 1996).

We moved a step further and proposed another perspective and function of science notebooks (Ruiz-Primo 1998, Ruiz-Primo et al. 1999). We proposed notebooks as an unobtrusive assessment tool to be used not only by teachers but also by individuals outside the classroom (for example, school district personnel). What information can be collected from an outsider's perspective when using students' science notebooks as an assessment tool? The most evident answer is information on students' performance. However, we proposed (Ruiz-Primo et al. 1999) that students' notebooks could also be used to collect information about opportunity to learn. Consistent with the National Science Education Standards (National Research Council 1996), we believe that students should not be held accountable for achievement unless they are given adequate opportunity to learn science. Therefore, both students' performance and opportunity to learn science should be assessed.

In what follows we: (a) propose a framework for conceptualizing students' notebooks as an assessment tool, (b) propose an approach for scoring them, and (c) provide evidence on the technical quality of science notebook scores.

On student performance

According to the National Science Education Standards (National Research Council 1996), inferences about students' achievement can be based on an analysis of their classroom performances and work products. Communication and understanding are considered in the Standards as fundamental for both performance- and product-based assessments. If science notebooks are considered as one of the possible products of students' work, evidence about students' communication and understanding might be collected from the written/schematic/pictorial accounts of what they do in a science class.

Focusing on the characteristics of students' scientific communications is not irrelevant. Constructing sound and scientifically appropriate communications helps students not only to better understand scientific concepts and procedures, but also to participate in a scientific community. Not knowing the "rules of the game" alienates students from the scientific culture and keeps them scientifically illiterate (for example, Bybee 1996, Lemke 1990, Martin 1989, 1993). Results across different studies (see Rivard 1994) have suggested that writing science in an expository fashion (for example, explaining, taking notes, summarizing results)

enhances students' learning. Although expressive writing appears to be useful, its effectiveness in improving science learning is not conclusive (see Rivard 1994). This means that if science notebooks invite students to write expressively rather than focus on what is being learned and how to communicate it, the notebook's usefulness may be diminished. The more emphasis that is placed on "doing science in your own words" the less students are helped to understand the fundamental role of scientific language in doing science (for example, Lemke 1990, Marks and Mousley 1990, Martin 1989, 1993). In sum, a necessary part of becoming a proficient science student is learning to read and write the various genres in science, such as writing a report (for example, Bybee 1996, Lemke 1990, Martin, 1989, 1993). Therefore, one aspect of the students' performance we proposed to focus on is the quality of communication in their notebooks.

The second aspect is understanding. An analysis of students' writing can provide insight into the state of students' understanding, misconceptions, and other factors associated with learning (for example, Audet et al. 1996; Dana et al. 1991, Fellows 1994, Heinmiller 2000, Hewitt 1974, McColskey and O'Sullivan 1993, Shepardson and Britsch 1997). Following some of Bybee's (1996) dimensions of scientific literacy we focused on conceptual and procedural understanding. Conceptual understanding involves the functional use of scientific words/vocabulary appropriately and adequately as well as relating the concepts, represented by those words (i.e., understanding facts, concepts, and principles as parts of conceptual schemes). Procedural understanding emphasizes the abilities of inquiry–the processes of science. These abilities include not only observing, hypothesizing, and experimenting, but also using evidence, logic, and knowledge to construct explanations (Duschl 2003).

On opportunity to learn

Opportunity to learn focuses on evidence that the curricular objectives are translated into topics actually taught in the classroom (i.e., Do teachers provide instruction on the knowledge stated in the adopted science curriculum?). Inferences about opportunities to learn using students' notebooks are based on the assumption that science notebooks are an account of what students do in their science classroom. If this assumption holds, it should be possible to map instructional activities implemented in a science classroom when information from individual science notebooks is aggregated at the classroom level. If none of the students' notebooks from a class has any evidence that an instructional activity was carried out, it is unlikely that the activity was implemented.

According to the National Science Education Standards (National Research Council 1996), another aspect of opportunity to learn is the quality of teaching. If science notebooks allow teachers to assess students' understanding, we think some evidence of this assessment should be found in the students' notebooks in the form of teacher's written comments. Indeed, Black and Wiliam (1998) provide strong evidence on the relation of the nature of feedback and student achievement. Black (1993) has shown that formative evaluation of student work (for example, feedback) can produce improvements in science learning. However, teachers' effective use of formative evaluation is hard to find (for example, Black 1993, Black and Wiliam 1998). Furthermore, classroom teachers are rarely good at providing useful feedback (for example, Sadler 1989, 1998, Wiggins 1993). Most of the time

feedback is considered as a comment in the margin that involves praise and/or blame or code phrases for mistakes (for example, "seg. sentence!"). Research has found that quality of feedback (for example, helpful comments, comments with grade, or grade only) affects its effectiveness for improving students' performance (for example, Butler 1987). Teachers' feedback, such as a grade (for example, B-) or a code phrase (for example, "incomplete!" or a happy face sticker), can hardly help students redirect their efforts to meet the needs revealed by their notebook entries (for example, Sadler 1989, 1998).

Based on these arguments, we proposed (Ruiz-Primo 1998, Ruiz-Primo et al. 1999) two indicators to evaluate opportunity to learn using science notebooks: (1) exposure to the science content as specified in the curriculum/program adopted, and (2) quality of teachers' feedback on students' performance. (We acknowledge that there are many indicators of opportunity to learn at the classroom level (for example, teachers' content and pedagogical knowledge, and their understanding of students). Science notebooks are seen as one source of evidence, among others, that can be used as an indicator of opportunity to learn.) We (Ruiz-Primo et al. 1999) named the first indicator unit implementation, and the second, teacher feedback on student performance.

Evidence of the implementation of an instructional activity can be found in different forms in a student's notebook: description of a procedure, hands-on activity report, interpretation of results, and the like. Variation in these forms is expected across activities and students' notebooks. Furthermore, notebook entries may vary from one student to the next within the same classroom for a number of reasons (for example, the student was absent when a particular instructional activity was implemented). The variety of notebook entries can be even wider when students' science notebooks are compared across different classrooms. To tap the variation in notebook entries within- and between-classes, entries in the notebooks are linked to the intended instructional activities specified in the curriculum adopted.

Students' science notebooks can also be used as a source of evidence on teachers' feedback. If science notebooks allow teachers to assess students' understanding, we would expect to see some evidence of feedback in the students' notebooks. If teachers do not respond, probe, challenge, or ask for elaborations of notebook entries, the benefit of the notebook as a learning tool and as an instrument to inform students about their performance may be lost.

Assessment approach

In previous papers (Ruiz-Primo 1998, Ruiz-Primo et al. 1999) we pointed out that science notebooks could be viewed as an assessment tool at two levels: (1) individual level–a source of evidence bearing on a student's performance over the course of instruction; and (2) classroom level–a source of evidence of opportunities students had to learn science.

Our assessment approach focuses on three aspects of students' science notebooks: (1) Unit Implementation–What intended instructional activities were implemented as reflected in the students' notebooks? Were any other additional activities implemented that were appropriate to achieve the unit's goals? (2) Student Performance–Were students' notebook communications appropriate according to the scientific genres? Did students' communications indicate conceptual and

procedural understanding of the content? And (3) Teacher Feedback–Did the teacher provide helpful feedback on students' performance? Did the teacher encourage students to reflect on their work?

Notebooks as an assessment tool

We conceived of an assessment as a combination of a task, a response format, and a scoring system (for example, Ruiz-Primo and Shavelson 1996a). Based on this triple, a notebook used as an assessment tool can be thought as a: (a) task that allows students to provide evidence bearing on their knowledge and communication skills; (b) format for students' communication (that is, entry); and (c) scoring system by which the students' notebook entries can be evaluated accurately and consistently.

In contrast to other forms of assessment (for example, multiple-choice, performance assessments or concept maps), the identification or correspondence of the triple is not straightforward for notebooks (see Ruiz-Primo 1998, Ruiz-Primo and Shavelson 1996a, b). How can a notebook assessment task be defined? What would the response format be? How can a scoring system be defined? As mentioned before, notebooks are a compilation of communications with diverse characteristics. Each of these communications is considered as a notebook entry and we view each entry as an instance of the triple. A notebook entry 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.

A notebook assessment task varies according to the notebook entry, which is viewed as the "what" students were asked to do. For example, reporting the data from the experiment carried out on a particular day. The entry format also varies according to the type of entry. Reporting data may take the form of a table or graph (a schematic communication), or simply a description of observations (a verbal communication). Finally, the scoring system is the criteria used to judge the notebook entry. The scoring system should allow scoring the same aspects across different types of entries. For example, each student's notebook entry should be considered as evidence of the student's understanding; some entries will focus more on conceptual understanding (for example, explaining a concept), some others on procedural (for example, conducting an experiment properly). In what follows, we explain how our approach incorporates the assessment triple to form an assessment tool.

Science notebook assessment tasks

As mentioned before, 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 (for example, write a procedure or explain a concept). The key issue, from the assessment perspective, is to identify the notebook entries according to the "what" students were asked to do. After reviewing dozens of students' science notebooks from different classrooms, and looking into the types of activities that students are supposed to do in a science class (see National Science Education Standards/National Research Council 1996), we identified fourteen general entry categories. We acknowledge that many different schemes can be used to analyze students' notebooks communications (see Audet et al. 1996, Keys 1999).

Table 1. Types and sub-types of notebook entries.

Genre	*Type of Entry*	*Code*	*Sub-Types*
Minor Genre	Defining	1	• Defining, verbal
		2	• Defining, pictorial
	Exemplifying	3	–
	Applying Concepts	4	–
	Predicting/Hypothesizing	5	–
	Reporting Results	6	• Reporting results, verbal
		7	• Reporting results, graphic
	Interpreting Data and/or Concluding	8	–
	Reporting and Interpreting Data and/or Concluding	9	• Reporting and interpreting, verbal
		10	• Reporting and interpreting, graphic
	Content Questions/Short Answer	17	–
	Quick Writes	18	• Contextualizing science
		19	• Narrative affective
		20	• Narrative reflections
Major Genre	Reporting Procedure	11	• Procedure recount
		12	• Procedure instructions
		13	• Procedure directions
	Reporting a Quasi Experiment	14	–
	Reporting an Experiment	15	–
	Designing an Experiment	16	–
NA	Assessment	21	• Simple forms (for example, short answer)
		22	• Complex forms (for example, performance assessments)
	Don't Care About Activity	23	–

NA: Not applicable
– No sub-type of entry

Notebook entries can be found in different forms of communication: verbal–written/text–(for example, explanatory, descriptive, inferential statements); schematic (for example, tables, lists, graphs showing data); or pictorial (for example, drawing of apparatus). Moreover, some of the categories proposed also include sub-types of entries according to the form of communication. For example, a definition can be verbal or pictorial (for example, drawing of a pendulum system). Therefore, the type of entry, definition, includes two sub-types of definitions. Table 1 presents the types and sub-types that we have identified. We assumed that all the entries provide information, at least partially, about the students' conceptual and procedural understanding and communication skills.

Each type of entry is considered to have its own characteristics that make it identifiable. For example, reporting results focuses on describing observations or presenting data, whereas interpretation focuses on summarizing and generalizing the data, or highlighting specific cases (for example, Penrose and Katz 1998). Once a notebook entry is identified as an instance of a particular type, the "what" that the student was asked to do is inferred. For scoring purposes and assuming that entry

types can be more or less identifiable, each type of entry was assigned a code (see table 1).

Science notebook entry formats

Entry formats vary according not only to the type of entry, but also to who provides the format–curriculum developers, the teacher, or the student. For example, a science unit/module adopted by a teacher/class/district may have, as part of its design, activity sheets that students fill out as they carry out the activity. Other times the teacher provides students an entry format; for example, a table for reporting data or a printed sheet on which to report an experiment. In these cases the entry formats are given to the student. Still, in other cases students are not provided with any response format. They are asked by the teacher to write about, say, the procedure used that day, and the students write about it with no response format imposed. We have, then, identified those possible sources of response formats with corresponding codes (for example, when the format is provided by the teacher the scorer provides the code ".6" after the type of entry, say 5.6).

Science notebook scoring system

The scoring system focuses on collecting information on students' performance and opportunity to learn. Therefore, for each entry identified, students' performance can be scored as to the quality of the communication–Did the student's communication correspond to the appropriate communication genre?–and the understanding–Did students' communications indicate conceptual and procedural understanding of the content presented? Both aspects are scored according to the requirements of the task. Opportunity to learn focuses on identifying entries that reflect the implementation of the intended instructional activities, as well as some aspects of the quality of instruction students received.

Student performance. In a pilot study we scored each communication as to: completeness, clarity, and organization (Ruiz-Primo et al. 1999). Completeness and clarity in communication were dichotomously scored 0 (No) or 1 (Yes). Organization of communication was evaluated using a three-level score: 0–No Organization (i.e., no sign of organization); 1–Minimal Organization (for example, student only uses dates to separate information or only lists information); and 2–Strong Organization (for example, student uses titles, subtitles, and labels appropriately). Unfortunately, results indicated that students' scores varied little since most communications were scored as complete and clear (score of 1) and with minimal organization (score of 1). The criteria did not accurately discriminate the quality of communication across students.

In this study, we approached the scoring of the quality of communication from the perspective of genres in scientific communication (for example, Lemke 1990, Martin 1989, 1993). (Lemke (1990) classifies the scientific genres into: (1) minor genres, short or simpler forms of communication, such as descriptions, comparisons, and definitions, and (2) major genres, usually longer, more complex, and more specialized communications, such as lab reports.) We focused on linking types of entries with scientific communication genres (see table 1). We defined the general characteristics of each genre so as to develop scoring criteria that went beyond

completeness, clarity, and organization (that is, Does the student's communication have the distinctive elements of the written genre in hand?). (The approach did not intend to focus on the functional analysis of the students' written communication (for example, lexical density or characteristics of the clauses; see Halliday and Martin 1993, Keys 1999). Instead it used only the general characteristics of the genres as criteria for scoring the quality of the communications.)

Take as an example, "definitions". Definitions are considered a minor scientific genre with clauses (the term to be defined and the definition) that explain what is what or who is who (for example, Lemke 1990, Martin 1993). The general characteristics of definitions are that they: (1) are always reversible (for example, "solutions are mixtures that . . .", or "mixtures that have the same properties . . . are called solutions"); (2) use technical terms, when appropriate, to condense information (for example, it is better to refer to mixtures than to substances that can be easily separated without making any new chemicals); and (3) are timeless (verbs in present tense) since technical/scientific definitions do not apply only to the "here and now". (We acknowledge that people may complain that focusing on the use of technical terms, when appropriate, may focus attention on "jargon" instead of understanding. First, translating jargon into common sense is responsibility not only of scientists but also of teachers. Second, scientists could not do their job without technical discourse. Not only it is compact, and therefore efficient, but, most importantly, it codes an alternative perspective on reality to common sense, a perspective accumulated over centuries of scientific inquiry (Martin 1993).) Using these general characteristics of the "definition genre" we developed scoring criteria (table 2). Criteria such as those presented in table 2 were developed for each type of entry. table 3 provides some examples of the criteria developed for minor genre ("defining pictorial" and "reporting data graphic") and major genre ("reporting procedures" and "designing experiments").

Some entry categories were considered to be "narrative" communications (for example, affective) rather than scientific communications and we also developed criteria to score them. Once a notebook entry was identified, the corresponding genre criteria were applied. If a student's quality of communication was 0 no further attempt was made to score understanding as evidenced by that entry.

Each entry/communication was also scored as to the conceptual or procedural understanding it reflected (for example, Does a student's explanation apply the concepts learned in the unit correctly? Does the student's description provide correct examples of a concept? Is the student's inference justified based on relevant evidence?). Scoring focused on conceptual understanding when the communication in the entry referred to defining, exemplifying, relating, comparing, or contrasting unit-based concepts, and on procedural understanding when the communication referred to reporting procedures carried out during an activity/experiment, reporting observations/results/outcomes, interpreting results, or concluding.

Conceptual and procedural understanding was evaluated on a four-point scale: (NA)–Not applicable (i.e., instructional task does not require any conceptual or procedural understanding); 0–No Understanding (for example, examples or procedures described were completely incorrect); 1–Partial Understanding (for example, relationships between concepts or descriptions of observations were only partially accurate or incomplete); 2–Adequate Understanding (for example, comparisons between concepts or descriptions of a plan of investigation were

Table 2. Example of the criteria used to score "quality of communication for definitions".

Score	*Criteria*	*Example*
0 Incoherent, incomplete, and not understandable communication	Definition is incomplete, not understandable.	*Mixture. When you put . . . (Incomplete sentence.)*
1 Understandable but does not use the characteristics of the genre	Definition is complete–the two parts of a definition are identifiable, BUT the definition does not have technical terms when appropriate. Definition may or may not have verbs in present tense OR may or may not make reference to the here and now.	*Mixture. When you put two or more things together.*
2 Understandable and uses ***some*** of the basic characteristics of the genre	Definition is complete–the two parts of a definition are identifiable, AND the definition has technical terms if appropriate. Definition may have verbs in present tense OR may not make reference to the here and now, BUT not both.	*Mixture. When you put two or more materials together.*
3 Understandable and uses ***all*** of the basic characteristics of the genre	Definition is complete–the two parts of a definition are identifiable, AND has technical terms if appropriate, AND has verbs in present tense AND does not make reference to the here and now.	*Mixture. Combining two or more materials together forms a mixture.*

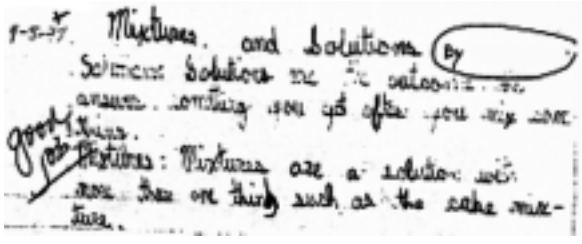
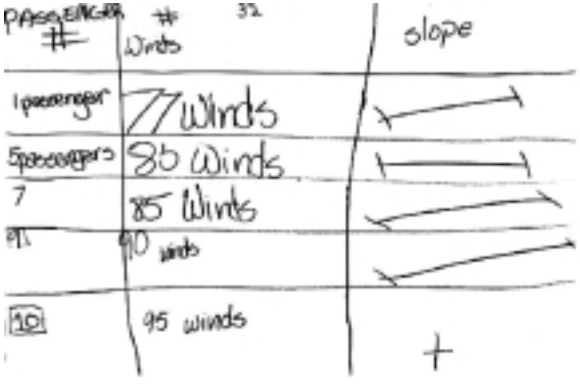
appropriate, accurate, and complete); and 3–Advanced Understanding (for example, communication focused on justifying student's responses/choices/decisions based on the concepts learned or the communication provided relevant data/evidence to formulate the interpretation). Table 4 provides examples of students' notebook entries focusing on conceptual and procedural understanding and the scores assigned.

Opportunity to learn. To answer the question about the implementation of the intended curriculum (i.e., What intended instructional activities, as specified by a particular curriculum/unit, were implemented as reflected in the student's notebook?), we first defined the instructional activities to be considered as evidence that the unit was implemented. The specification of these activities was based on an analysis of the intended curriculum. An inventory of the major activities served as a verification list for capturing the implementation of the basic instructional activities, as well as "other" activities implemented but not required by the curriculum (i.e., Were any other additional activities implemented that were appropriate to achieve the curriculum/unit goals?). Implementation was evaluated dichotomously based on the question, "Is there any evidence in the student's

Table 3. Examples of the criteria used to score quality of communication.

Codes	*Type of Genre*	Score			
		0	**1** *Understandable but not using the characteristics of the genre*	**2** *Understandable and uses* some *of the basic characteristics of the genre*	**3** *Understandable and uses* all *the basic characteristics of the genre*
2	Defining Pictorial	Incoherent, incomplete, not understandable communication	Representation can be easily identified (for example, as a drawing of a pendulum), BUT most of the important parts are not labeled (i.e., it can have one or more labels, but not the most important ones). Representation may or may not have a title, AND may or may not have technical terms if appropriate (for example, student uses "little pieces of glasses" instead of "crystals").	Representation can be easily identified AND most of the important parts are labeled. Representation may or may not have a title, OR may or may not have technical terms if appropriate, but not both.	Representation can be easily identified AND most of the important parts of the representation are labeled, AND has a title, AND has technical terms if appropriate.
7	Reporting Results, Graphic	Incoherent, incomplete, not understandable communication	Representation is clearly a table, a graph, or a schematic representation, BUT is not labeled properly (for example, columns and rows are not labeled), AND data is not presented in a way that shows a data trend (for example, data are not in ascendant/descendent order). Representation may or may not have a title AND/OR may or may not have technical terms if appropriate.	Representation is clearly a table, a graph, or a schematic representation, AND is labeled properly, AND data is presented in a way that shows a data trend (for example, data is in ascendant/descendent order). Representation may or may not have a title OR may or may not have technical terms if appropriate, BUT not both.	Representation is clearly a table, a graph, or a schematic representation, AND is labeled properly, AND data is presented in a way that shows a data trend (for example, data is in ascendant/descendent order), AND representation has a title AND has technical terms if appropriate.
11	Reporting a Procedure	Incoherent, incomplete, not understandable communication	Procedure covers some of the important steps, BUT steps are not presented in a clear sequence (i.e., numbered), so it is difficult to replicate procedure. Procedure may or may not have a title, AND/OR may or may not have technical terms when appropriate.	Procedure covers most of the important steps, AND steps are presented in a clear sequence (i.e., are numbered or clearly sequenced), so procedure can be replicated. Procedure may or may not have a title, OR may or may not have technical terms if appropriate, BUT not both.	Procedure covers all of the important steps, AND steps are presented in a clear sequence (i.e., are numbered), so procedure can be replicated, AND has a title, AND has technical terms if appropriate.
16	Designing an Experiment	Incoherent, incomplete, not understandable communication	Experiment staging is incomplete because only has procedure, but there is not purpose. Description of procedure may or may not be in future tense or imperatives AND/OR may or may not have technical terms if appropriate.	Experiment staging is complete, has a purpose and a procedure BUT verbs tense or forms in both descriptions are not appropriate. Descriptions may or may not have technical terms if appropriate.	Experiment staging is complete, has a purpose and a procedure AND verbs tense or forms in both descriptions are appropriate. Descriptions have technical terms when appropriate.

Table 4. Examples of students' notebook entries focusing on conceptual and procedural understanding and the scores assigned.

Type of Understanding	*Example of Students' Notebook Entries*	*Score*
Conceptual: Defining	Example 1 * 	***Definition of Solution*** *1 – Partial Understating:* Definition does not provide indication that student considers solutions as a special type of mixture where a material dissolves in a liquid and cannot be separated by screening or filtering (two types of separation methods studied in this activity). ***Definition of Mixture*** *1 – Partial Understating:* Definition is partially accurate since a mixture is not a type of solution. However, the student provides two extra pieces of information: Mixtures include more than one material (i.e., "thing") and a correct example.
Procedural: Reporting results, graphic	Example 2 	*1 – Partial Understating:* The student records most of the results collected from the experiment. However, the student seems to conduct the controlled-experiment inappropriately by changing more than one variable at the same time. In addition, the student does not report the slope information precisely.

* This example was taken from the pilot study. We erased the student's name for confidentiality concern.

notebook that activity "X" was implemented?" A score of 1 denoted the affirmative and 0 denoted absence.

We also assessed the quality of teacher feedback for each notebook entry identified. We used a six-level score: –2–feedback provided, but incorrect (for example, teacher provided a positive feedback, say A+, to a student's communication that was incorrect); –1–no feedback, but it was needed (for example, teacher should point out errors/misconceptions/inaccuracies in student's communication); 0–no feedback; 1–grade or code phrase comment only; 2–comments that provided students with direct, usable information about current performance against expected performance (for example, comment is based on tangible differences between current and hoped performance, "Don't forget to label your diagrams!"); and 3–comments that provided students with information that helped them reflect on/construct scientific knowledge (for example, "Why do you think is important to know whether the material is soluble for selecting the method of separation?). Rules were created for those entries in which one instructional task had more than one type of feedback.

Pilot study

The notebook assessment approach proposed was applied in an exploratory study with a sample of 32 students' science notebooks from six classrooms (see Ruiz-Primo et al. 1999). Three of the classrooms taught the Full Option Science System (FOSS) Variables Unit and the other three the FOSS Mixtures and Solutions Unit. The results of this exploratory study indicated that: (1) Students' science notebooks could be reliably scored. Unit implementation, student performance, and teacher feedback scores were highly consistent across scorers and units. Interrater reliability was, on average across the two units, 0.91 for the unit implementation score, .86 for the student performance score, and 0.88 for the teacher feedback score. (2) Inferences about unit implementation using notebook scores were accurate and hence justified. We found a high agreement between the unit implementation score and independent sources of information (for example, teachers' unit logs). (There is no information available about the validity of the teacher logs. The Teacher Logs were developed by another institution who at that moment was partner in the research project in which this pilot study was immersed. However, it is important to mention that the main focus of Teacher Logs was on the curriculum activities implemented by the teachers and the percentage of time they spent on them. Teachers only had to check those activities implemented on a particular day. We considered that the information provided by the Teacher Logs could be useful as a source of validation of the implementation score.) On average, the percentages of agreement between the activities reported in the teachers' logs and the activities captured in the notebook scoring form were: 93 for Variables and 97 for Mixtures and Solutions. (3) Evidence for the validity of inferences about student performance was also encouraging. High and positive correlations with hands-on performance assessment scores indicated that the student performance score could be considered an achievement indicator. (4) The unit implementation score helped explain differences in the performance across classrooms. Those classrooms in which notebooks showed that more instructional activities were implemented were also associated with higher performance assessment mean scores. And (5) teacher feedback scores corresponded with the variation observed on students' performance

assessment scores. Those classrooms with higher scores on teacher feedback were those with higher performance assessment mean scores.

The purposes of this study were to: (1) replicate, if possible, the pilot study results using a larger sample of students' notebooks, (2) provide evidence about the technical quality of notebook scores, and (3) track improvement of student learning over a course of a school year using the notebooks as an assessment tool.

The study

Students' notebook selection

Six fifth-grade classrooms in the Bay Area participated in this study. These classrooms were selected among the 20 classrooms (484 fifth-graders) that participated in a larger study to evaluate the impact of inquiry science curricula reform (Ruiz-Primo et al. 2002b). (In Ruiz-Primo et al. (2002b), we proposed a multilevel-multifaceted approach to evaluating the impact of education reform on student achievement that would be sensitive to context and small treatment effects. The approach uses different assessments based on their proximity to the enacted curriculum. The rationale behind this approach is the spread of effect of reform. If there is a reform effect to be found, first it should be found centrally and then the effect should trail off in regions increasingly distant to the enacted curriculum. To provide evidence about the sensitivity of the approach in ascertaining outcomes of hands-on science programs we administered close, proximal, and distal performance assessments to evaluate the impact of instruction based on two FOSS units.) The school district in which the study was conducted had received support from NSF to implement hands-on science since 1990. All teachers had been provided with professional development to support the implementation of the FOSS science curriculum adopted by the school district for the elementary school.

In Ruiz-Primo and colleagues' (2002b) study all students were administered three assessments that differed in their proximity to the curriculum enacted: close assessment–close to the content and activities of the unit; proximal assessment–tap knowledge and skills relevant to the curriculum, but specific topics can be different from the ones studied in the unit; and distal assessment–based on state/national standards in a particular knowledge domain but topics not related to those studied. In all 20 classrooms two FOSS units (1993) were implemented, Variables in the fall, and Mixtures and Solutions (henceforth Mixtures) in the spring. The close and proximal assessments were administered before and after the instruction of each unit, Variables and Mixtures. (See Appendix A for a detailed description of the assessments.)

We selected the six classrooms based on the magnitude of the effect sizes (Δ) observed (for details see Ruiz-Primo et al. 2002b). Assuming that effect sizes indicated the students' performance improvement from pretest to posttest, we wanted to take a closer look into those classrooms whose effect size magnitudes indicated not much growth, some growth, and large growth across the two units. Based on the effect sizes observed for the close assessments across the two units, Variables (V) and Mixtures (M), we selected two top-classrooms (Top 1 – Δ_V = 1.11, Δ_M = 1.62 and Top 2 – Δ_V = .95, Δ_M = 1.60), two middle-classrooms (Middle 3 – Δ_V = .73, Δ_M = 1.16 and Middle 4 – Δ_V = .62, Δ_M = 1.37), and two low-classrooms (Low 5 – Δ_V = .60, Δ_M = .66 and Low 6 – Δ_V = . 53, Δ_M = . 59)

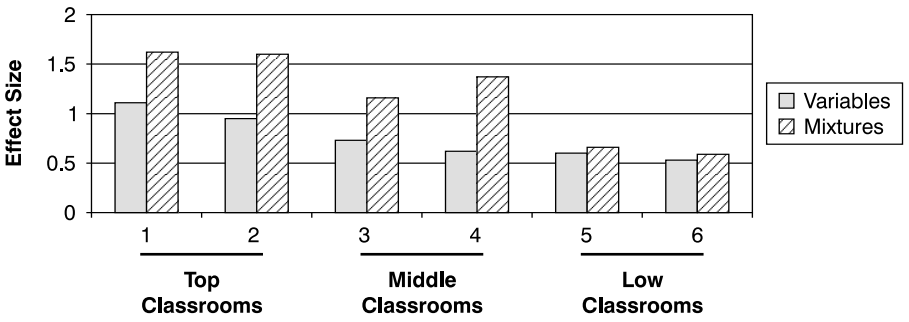


Figure 1. Effect size for close performance assessments across units for the classrooms selected for the study.

(see figure 1). (A scatterplot of the effect sizes across the two units helped in selecting the classrooms and classifying them on the three levels.)

Teachers reported that they regularly used notebooks in their science classes. No instructions were provided to teachers on how to use science notebooks or on the characteristics notebooks should have. Students' notebooks were collected at the end of the school year. Teachers were asked to sort students into five ability groups–from the top 20 percent in the class to the bottom 20 percent–according to their science classroom performance on each unit. Criteria used for this classification were based on the teachers' own evaluation system.

For each of the six classes we randomly selected notebooks from two top-, two middle-, and two low-student groups. Each student in the sample had two notebooks, one for Variables, generated during the fall, and another one for Mixtures, generated during the spring. A total of 72 science notebooks (877 pages), 36 for the Variables unit and 36 for the Mixtures, were scored for this study.

Performance assessments scores were available for each student. As mentioned, we administered the close and proximal assessments for each unit before and after instruction. Students within each classroom were randomly assigned to one of two sequences of pretest and posttest: (1) close–close or (2) proximal–proximal (for example, those students who took the close assessment as a pretest, also took the close assessment as a posttest). The distal assessment was administered to all students after instruction in both units during the spring. (Distal performance assessment scores were provided by the CSIAC project director, Kathy Comfort.)

Students' notebook scoring

The scoring materials consisted of two parts: (1) Notebook Scoring Form–a verification list that included, as rows, the instructional activities to be considered as evidence that the unit was implemented and, as columns, the aspects to be scored; and (2) a Criteria Table–a table that specifies codes, criteria, and examples to be used in scoring. To score students' notebooks two scoring forms, one per unit, were developed. The instructional activities specified in the Notebook Scoring Form were based on the description of the implementation presented in the teacher guide for each FOSS unit. The Notebook Scoring Form followed the units' organization: one

Mixtures Unit—Activity 1—Separating Mixtures	1 Basic/Extra Implemented? 0-1	2 Complete Report? 0-3	3 Type of Entry 1-23	4 Quality of Communication 0-3	. . . .
Part 1 Make and Separate Mixtures					
Making Mixtures					
Activity Sheet: Separating -- Mixtures Part 1					
Activity Sheet: Separating -- Mixtures Part 3					
1.1.1					
1.1.2					
Screening Mixtures					
1.2.1					
1.2.2					
…					
Defining: Solutions					
Extra Activity					
Part 2 Weight and Separate a Salt Solution					
Making and weighting a Salt Solution					
Activity Sheet: Separating a Solution -- Part 3					
2.1					
2.2					
Review Question: What is a mixture?					
…					
Extra Activity					
Part 3 Salt Crystals					
Evaporating a Saltwater Solution					
Activity Sheet: Separating a Solution -- Part 6					
Activity Sheet: Separating Mixtures Review – Part 1					
3.1					
3.2					
…					
Extra Activity					
Reflections on the Activity—Questions at the end of the Act.					
Integrating: How are screen and filter similar/different?					
…					. . .

Figure 2. Example of a portion of the notebook scoring form for Activity 1, Separating Mixtures, of the Mixtures Unit.

verification list for each activity and one for assessments suggested in the guide (i.e., hands-on assessments, pictorial assessments, reflective questions). Each activity-verification list contained different Parts (P) that corresponded to the organization of the activity (figure 2. As an example, only the first four columns of the scoring form are presented).

For each instructional activity specified on the Notebook Scoring Form, seven questions are asked according to the three aspects of the notebook evaluated: (1) Unit Implementation (Is there any evidence that the unit-based instructional activity or that an appropriate extra-instructional activity was implemented? Is the activity sheet/report complete? What type of entry is identified in the evidence provided?). (2) Student Performance (Quality of the communication–Is the communication appropriate according to the genre at hand? Understanding Reflected in the Communication–Is there any evidence of conceptual understanding in the communication? Is there any evidence of procedural understanding

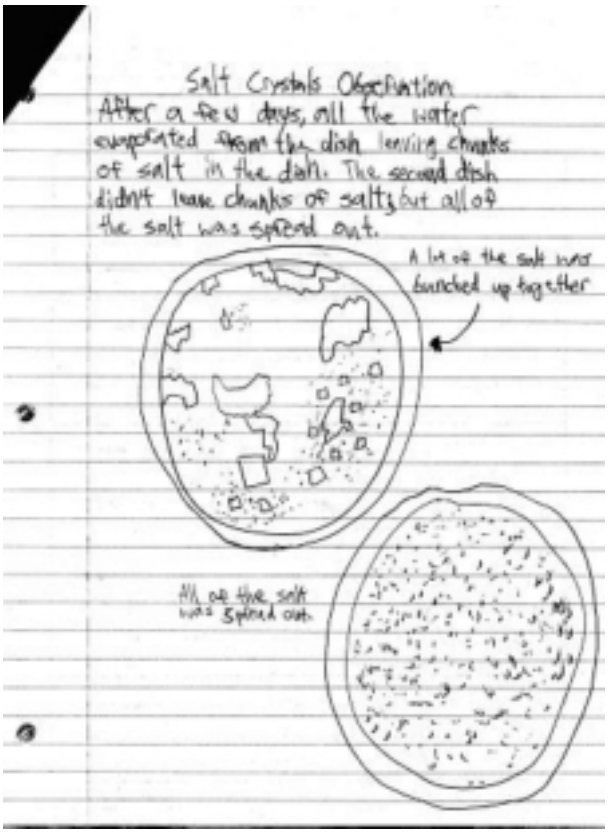


Figure 3. An example of a student's notebook entry for the Mixtures and Solutions unit.

in the communication?). And (3) Teacher Feedback (Is there any evidence that the teacher provided feedback on the student's communication?).

Each notebook entry was linked first to an instructional activity listed on the Notebook Scoring Form. For example, in figure 3, the student's Mixtures notebook entry can be linked to Activity 1, Separating Mixtures, Part 3, Salt Crystals (see figure 2). The student's entry focused on reporting observations about evaporation, therefore a "1" is placed in column 1, row 22 and the evidence (i.e., observations) is located in row 3.1, under "Evaporating a Saltwater Solution". (If the notebook entry cannot be linked to an instructional activity specified in the scoring form, then it is considered an extra activity and the scorer needs to determine whether or not the extra activity is relevant to the unit.) Notice that the Notebook Scoring Form is designed to capture all the different notebook entries that can provide information about the implementation of a particular instructional activity, but the activity is scored only once as implemented or not (1 or 0). By doing this the implementation score accurately reflects which instructional activities, as prescribed by the FOSS teacher guide, were implemented, not boosting the score by considering all notebook entries done as different instructional activities. Given the context of the entry it is more or less easy to assign to which part (P) of the activity the entry belongs.

Once a notebook entry was linked to an instructional activity, the next step was to identify the type of entry it reflected. In our example, the type of entry can be identified as "Reporting results, verbal". Therefore, Code "6" (see table 1) was written in Column 3, Row 25–3.1. Once the type of entry was identified, we scored quality of communication (Column 4), conceptual understanding if appropriate (Column 5), procedural understanding if appropriate (Column 6), and teacher feedback (Column 7 for feedback). If a student's communication was scored "0" we did not attempt to score the student's understanding.

The shaded boxes (figure 2) in the Notebook Scoring Form mean that the criteria do not apply to the notebook entries in hand. For example, the criterion, "Completeness of Report," only applies to the "Activity Sheet". Activity sheets are provided by FOSS for students to fill out for each activity and they are considered an essential piece of the implementation of any unit activity.

Types of scores

Five types of scores were obtained with the approach: unit implementation, quality of communication, conceptual understanding, procedural understanding, and teacher feedback. Unit implementation and teacher feedback scores were the sum of scores obtained for each notebook entry identified whereas the three student performance scores were mean scores (that is, the sum score divided by the number of entries identified in each student's notebook; Li et al. 2000). The advantage of these "mean scores" was that they shared the same scale (from 0 to 3) making it easy to compare the level of a student's performance on the different aspects.

Results

Analyses focused on two main issues: (1) The technical quality of the notebook assessment–Can different raters reliably score student's science notebooks? Can scores on quality of communication, conceptual understanding, and procedural understanding be interpreted as reflecting students' academic performance? And, do notebook scores bearing on student performance correlate positively with other measures of their performance? And (2) students' improvement over a course of the school year using notebooks as an assessment tool.

Reliability

Eighteen of the 72 notebooks (236 pages) were used to train and calibrate scorers. After training, 24 notebooks (394 notebook pages), 12 from Variables and 12 from Mixtures, were scored by three independent raters. Notebooks were sampled across classrooms and students' performance level. We evaluated interrater agreement in classifying notebook entries according to entry type. Also, interrater reliability was calculated for each score across units (table 5). Results are consistent with our previous studies (Ruiz-Primo et al. 1999, Li et al. 2000). In general the magnitude of the coefficients were high across types of scores. Magnitudes were lower for student performance scores than for implementation and teacher feedback scores.

We interpreted these results as evidence that despite the variability in the number and type of entries in a notebook and the diversity of the forms of communications (written, schematic or pictorial), raters consistently identified

Table 5. Percent of agreement and interrater reliability across types of scores.

Unit	*Percent of Agreement**	*Interrater Reliability***				
	Type of Entry	*Unit Implementation*	*Quality of Communication*	*Conceptual Understanding*	*Procedural Understanding*	*Teacher Feedback*
		Total Score	*Mean Score*	*Mean Score*	*Mean Score*	*Total Score*
Variables	81.29	.99	.82	.88	.85	.91
Mixtures	84.79	.99	.82	.84	.83	.83

* Percent of exact agreement based on two raters. Percentages of agreement across three raters were 65.39 for Variables and 74.12 for Mixtures.
** Interrater reliability across three raters.

whether or not an instructional task was implemented, and consistently classified the type of entry identified. Raters also consistently scored student performance.

Based on these results, the remaining 30 students' notebooks (247 pages) were randomly distributed among the three raters and only one rater scored each notebook. Distribution of notebooks was done in such a way that all three raters scored students' notebooks from all classrooms across the two units.

Notebook Scores

Table 6 provides descriptive information for each type of notebook score. The maximum score for Unit Implementation was the total number of FOSS instructional activities. Extra activities were not included in this analysis.

Only about 20 percent of the instructional activities described in the FOSS teacher guides for both units were implemented in the classrooms (20.34 percent for Variables and 21.20 percent for Mixtures). Low student performance scores

Table 6. Means and standard deviations for each type of scores across units and classrooms.

Type of Score	*n*	*Variables Notebooks*			*Mixtures Notebooks*		
		Max	*Mean*	*SD*	*Max*	*Mean*	*SD*
Unit Implementation	36	75	15.27	9.52	93	19.72	12.58
Student Performance							
Quality of communication	36	3	1.31	.37	3	1.10	.29
Conceptual understanding	20*	3	1.16	.58	3	.99	.55
Procedural understanding	36	3	1.28	.39	3	1.11	.31
Teacher Feedback	36	3	–.19	.19	3	–.30	.18

* There was no evidence of entries focusing on conceptual understanding in two classes (12 students) and four students of other classes.

across units revealed that students' communication skills were not well developed and that students only partially understood the different topics addressed in the units. Teacher feedback scores were negative across the two units, meaning that teachers tended not to provide feedback to students despite errors or misconceptions that were evident in the students' communications. In fact, no teacher feedback was provided in any of the students' notebooks in this sample. (Only in one notebook entry in one student's notebook did we find evidence of feedback. The teacher only corrected student's spelling errors). If one aspect of good teaching is informative feedback, this finding says something about the quality of instruction these students received. Since no teacher feedback was provided, we only focus on unit implementation and student performance scores.

Validity

The validity analysis focused on the following issues: If scores on quality of communication, conceptual understanding, and procedural understanding can be interpreted as reflecting students' academic performance, they should correlate positively with one another. And, if notebook scores bear on student performance, they should correlate positively with other measures of their performance.

Correlations among student performance scores

If quality of communication, conceptual understanding, and procedural understanding are related, we should expect positive correlations among the three types of scores. However, if they are not tapping overlapping aspects of achievement, the direction may not be positive and the magnitude of the correlations should not be high.

We found positive correlations (table 7). (We acknowledge that since the sample size is not very large, deviant scores could change the magnitude of the correlations. Therefore, we checked every scatterplot to decide whether or not to delete a case.) The magnitude of the correlations (0.49–0.73) indicated that the three aspects were related but still tapping somewhat different aspects of student performance. The correlation between quality of communication and procedural understanding was the highest in magnitude, especially in the Mixtures unit. This was probably due to the type of entries that Mixtures notebooks had. Entries that focused more on process skills require better communication skills than those focusing on definitions or examples.

Based on the magnitudes of the correlations, we concluded that the scoring system tapped, as claimed, somewhat different aspects of the student achievement. Still, since we wanted to create a composite score that reflected a student's overall performance, we averaged the three scores and created a total score for student performance, named simply, "student performance score." We used this score, with a maximum of 3, in the rest of the analyses.

Notebook student performance scores as achievement indicators

As mentioned, students within each classroom were randomly assigned to one of the two types of performance assessments (i.e., close or proximal assessments) administered before and after the instruction of the units (i.e., Variables and

Table 7. Correlations among types of mean scores.

	Type of Score		
	QC	*CE*	*PE*
Variables			
Quality of communication (QC)	–		
Conceptual understanding (CE)	.53* (n = 20)	–	
Procedural understanding (PE)	.55**[a] (n = 33)	.49*[a] (n = 17)	–
Mixtures			
Quality of communication	–		
Conceptual understanding	.52**[b] (n = 31)	–	
Procedural understanding	.73** (n = 36)	.51**[b] (n = 32)	–

** Correlation is significant at. 01 level.
* Correlation is significant at. 05 level.
a Three outliers dropped.
b Two outliers dropped.

Mixtures; see Appendix A for characteristics of the performance assessments). All students in each classroom took the distal assessment. To examine whether the notebook scores bearing on student performance behaved as achievement indicators, these scores were correlated with scores students obtained on the posttest performance assessments. Table 8 shows the correlations obtained across units by the proximity of the assessments to the enacted curriculum.

In general, the correlations observed in this study, when outliers were deleted, were consistent with the ones observed in our previous study (Ruiz-Primo et al. 1999): all were positive, as expected, and the pattern varied according to the proximity of the assessments. This means that, on average, the higher the student performance score obtained in the notebook, the higher the score obtained by the student in the performance assessments, independent of the proximity of the assessment to the curriculum studied. However, the magnitude of the correlations varied from one study to the next. In our previous study (Ruiz-Primo et al. 1999) the magnitude of the correlations across the three types of performance assessments was lower for the Variables unit (0.34, on average), and higher for the Mixtures unit (0.76, on average). Furthermore, the lowest correlations were observed for the proximal assessments across both units. In this study the pattern of the correlations observed is closer to the expected pattern–higher correlations are observed with the more proximal assessments (see Appendix A for details about the characteristics of the assessments used.). The pattern of correlations for the Variables unit was just as expected, but for the Mixtures unit, the correlation between the notebook performance score and the proximal performance assessment score was higher than the correlation with the close performance assessment score. Even when the correlations were adjusted for general ability (for example, reading score), the pattern and the magnitude did not change dramatically for the Variables unit, but

Table 8. Correlations and partial correlations between notebook student performance score and performance assessment scores of different proximities.

Proximity of Performance Assessments	*Notebook Student Performance Score*					
	Correlations				*Partial Correlations*[a]	
	Complete Sample		*Without Outliers*		*Without Outliers*	
	Variables	*Mixtures*	*Variables*	Mixtures	*Variables*	*Mixtures*
Close	.09 (n = 14)	.35 (n = 20)	.89***[b] (n = 10)	.58**[d] (n = 19)	.83***[b] (n = 7)	.25[d] (n = 14)
Proximal	.54** (n = 22)	.49 (n = 16)	.71***[c] (n = 20)	.61**[d] (n = 15)	.55**[c] (n = 16)	.63**[d] (n = 16)
Distal	.34 (n = 29)	.26 (n = 29)	.49**[c] (n = 27)	.43*[d] (n = 28)	.50**[b] (n = 22)	.30[d] (n = 24)

*** Correlation is significant at .005 level.
** Correlation is significant at. 01 level.
* Correlation is significant at. 05 level.
a Reading scores were controlled.
b Four outliers dropped.
c Two outliers dropped.
d One outlier dropped.

they dropped in the Mixtures unit almost by half in the close and distal assessments.

A possible explanation for the difference in the magnitude of the correlations between the two studies may be the study designs used. In the first study three classrooms studied the Variables unit and another three the Mixtures unit; therefore, students were different across the two units. In this study, however, students are the same across the two units since all of the classrooms studied both units.

We interpreted these results as indicating that notebook performance scores may serve as an achievement indicator, even at a distal level, when the content of the assessment is not based on the content of the curriculum students studied in their science classes.

Opportunity to learn and performance scores

To address the opportunities students had to learn the units' content, we examined only the unit implementation score since no teacher feedback was found in this sample of notebooks. First we evaluated whether implementation scores varied according to the rank of the classrooms based on the effect sizes. Figure 4 and table 9 provide information about unit implementation and effect sizes on the close performance assessment according to the rank of the classroom. The pattern of the histograms across implementation and effect size seems to be closer to the expected for the Variables unit, but not for the Mixtures unit (figure 4). Despite the high

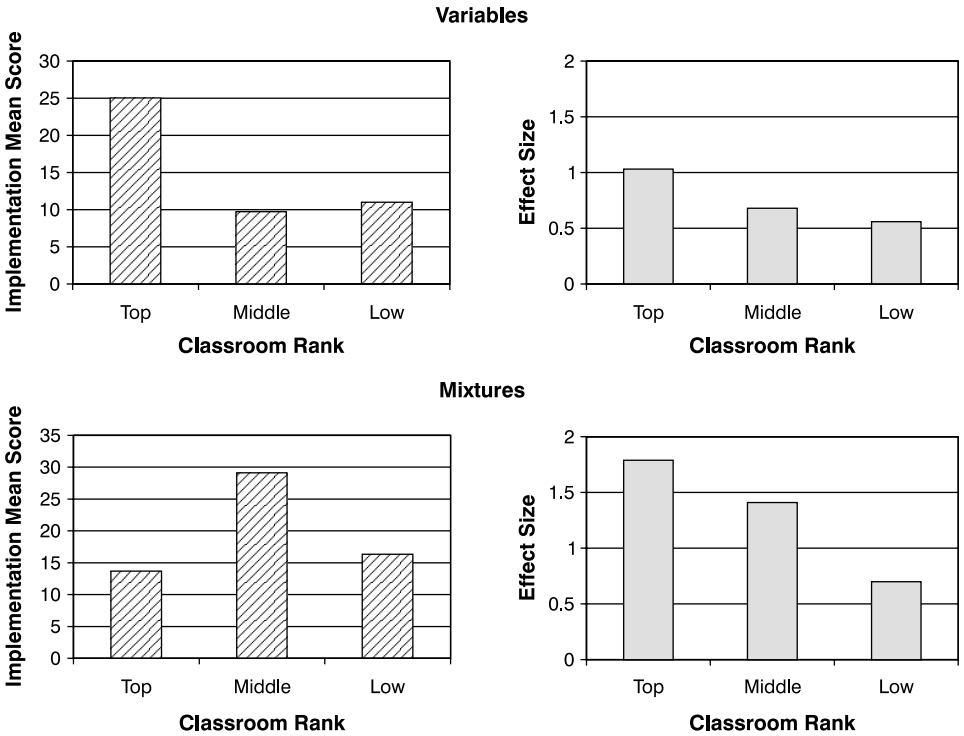


Figure 4. Histograms comparing unit implementation and close performance assessments effect sizes across units.

implementation mean score for the middle-classes in this unit, students did not perform in the performance assessment as well as it would be expected (table 9). In sum, high implementation mean scores were not always associated with high classroom performance. These results were not consistent with what we observed before. In our previous study (Ruiz-Primo et al. 1999) we found that those

Table 9. Unit implementation mean score across units and classrooms.

		Unit Implementation Score			
		Variables (Max = 75)		*Mixtures (Max = 93)*	
*Class Rank**	*n*	*Mean*	*SD*	*Mean*	*SD*
Top	12	25.06	5.70	13.69	5.38
Middle	12	9.75	2.78	29.13	12.56
Low	12	11.00	9.53	16.33	12.96
All	36	15.27	9.53	19.72	12.58

* Rank based on effect sizes observed across the two units.

classrooms in which more instructional activities were implemented, higher performance assessment scores were observed.

The issue, then, might be not on how much was implemented, but the quality of the implementation. To explore some aspects of quality of implementation we analyzed the type of activities done in the classrooms as reflected in the students' notebooks. Was there a difference in the activities carried out in the classrooms according to their rank? Across the three groups the types of activities most frequently found were "reporting data" (on average, 38.11 percent of the notebook entries in the Variables notebooks and 40.83 percent in the Mixtures notebooks), "definitions" (on average, 22.92 percent in the Variables notebooks and 22.90 percent in the Mixtures notebooks), and "short-answer questions" (on average, 8.17 percent in the Variables notebooks and 18.90 percent in the Mixtures notebooks). (See Appendix B for number of entries by classroom and unit.) The differences across the three groups of classroom were observed more clearly in other types of activities. For example, the top classrooms had a higher percentage of entries under the category of "applying concepts" (for example, 7.92 percent in the Mixtures unit, versus 1.31 percent on average for the other two classroom groups), "predicting/hypothesizing" (for example, 2.04 percent in the Variables unit, versus 0 percent for the other two groups), and "interpreting data/concluding" (for example, 4.08 percent in the Variables unit, versus 0.98 percent on average for the other two classroom groups). It seems that students from top classrooms had slightly better opportunities to explore other types of activities that might helped them improve understanding, than students from the other two groups.

Differences in performance across units

A final issue to consider is whether students improved their performance over the course of instruction during the school year. It is clear that the content of the units differed (Variables was taught in the fall and Mixtures in the spring), so it might be difficult to compare students' understanding from one unit to the next. However, students' communication skills should improve over the course of the school year. That is, independent of science unit content, students should, for example, improve the quality of their reports of an experiment or data collected.

To assess the difference between communication skills across the two units, we carried out an overall- (that is, across all classrooms) and within classroom-dependent *t*-tests (table 10). Results of the overall *t*-test indicated a significant difference in the quality of communication score ($t = 3.12$; $p = 0.004$). Unfortunately, the means for quality of communication was lower for the Mixtures notebooks, taught during the spring, than for Variables notebooks, taught in the fall.

Overall, students' quality of communications not only did not improve, but performance decreased over the course of instruction. At the class level, the pattern of means was reproduced–quality of communication mean score was higher for Variables than for the Mixtures. However, only in Classroom 1, one of the two top classrooms, the difference was significant ($t = 8.57$, $p = 0.000$; see table 10). If definitions, reporting data, and short-answer questions were the type of entries most frequently found in students' notebooks across the Variables and Mixtures units, it should be expected that those entries that focused on reporting data be of much higher quality, assuming appropriate attention and feedback was provided to

Table 10. Quality of communication mean scores across units by classrooms rank.

		Quality of Communication				
		Variables (Fall)		*Mixtures (Spring)*		*Sig.*
Classroom Rank[a]	*n*	*Mean*	*SD*	*Mean*	*SD*	
Top						
Class 1	6	1.54	.12	.97	.16	⋆
Class 2	6	1.00	.18	.93	.26	
Middle						
Class 3	6	1.49	.42	1.26	.36	
Class 4	6	1.34	.53	1.22	.20	
Low						
Class 5	6	1.11	.32	1.02	.19	
Class 6	6	1.35	.27	1.19	.46	
All	36	1.31	.37	1.10	.29	⋆

[a] Rank based on effect sizes observed across the 2 units.
⋆ Significant difference at .05 level.

students. However, most of the definitions were copied and most of the reporting data was on tables of low quality (that is, not labeled, no titles, no units of measurement). This means that based on the type of entries it is clear that students did not have many opportunities in their classrooms to better explore other forms of communication that may help them to improve understanding (for example, evidence to support explanations or design of experiments). More specific results about the types of entries found are presented in Ruiz-Primo et al. (2002a).

Conclusions

In this study we explored the use of students' science notebooks as an assessment tool for providing evidence bearing on their performance over the course of instruction and on the opportunities they have to learn science. We examined whether students' notebooks could be considered a reliable and valid form of assessment and whether they could be used to explain, at least partially, between-class variation in performance.

Our results, in short, indicated that: (1) Students' science notebooks can be reliably scored. Unit implementation, student performance, and teacher feedback scores were highly consistent across raters and units; (2) High and positive correlations with other performance assessment scores indicated that the student performance score can be considered as an achievement indicator. Although the pattern of correlations was not the same across the two units, in general, correlations were in the right direction; (3) Student's communication skills and understanding were far away from the maximum score and did not improve over the course of instruction during the school year; and (4) this latter result may be due, in part, to the fact that no teacher feedback was found in any of the students'

notebooks. Therefore, there was no effort to close the gap between the student performance at the time that a notebook entry was produced and the desired performance.

Overall, our approach for assessing students' science notebooks is promising in its capacity to describe important aspects of student performance and opportunity to learn in science. The technical quality of the notebook scores is adequate. Although interrater reliability and agreement was appropriate, the scoring approach can be improved to reduce scoring time and improve consistency between raters. The three aspects of student performance–quality of communication, conceptual understanding, and procedural understanding–correlated positively and moderately. Therefore, this distinction seems to be pertinent and should be maintained in scoring the next set of data. The results indicated that the notebook student performance and opportunity to learn scores have the potential of being educational indicators of classroom practice and student achievement. Notebook scores correlated high and positively with other achievement measures (that is, science performance assessments).

Our findings also appeared useful in suggesting areas for professional development. It is clear that keeping a science notebook is a wide spread teaching science practice. All the classrooms in which we have collected information used science notebooks. (It is important to remember that teachers who have participated in our studies were not trained on how to use science notebooks or were not asked to create them for these studies.) However, the main issue is how science notebooks are used.

Results from our studies indicated that writing in science notebooks was mainly mechanical. Almost for every instructional activity, students were asked to write down the procedure used and the results found. As mentioned, the quality of the descriptions was poor. Procedures were hardly replicable, results were almost never organized in a way that could help students to find patterns, and almost never were used as evidence in explanations or conclusions. Furthermore, explanations and conclusions were difficult to find across students' notebooks and classrooms.

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. Furthermore, it has been demonstrated that writing in science in an appropriate, purposeful, and relevant way can improve students' learning and understanding (Lemke 1990, Martin 1989, 1993, Rivard 1994). Unfortunately, teachers in our sample were not using science notebooks in an efficient and effective manner to help students improve their performance and understanding in science. Students' notebooks hardly had entries focused on the understanding of the concepts learned that day. The only entry related to the concepts learned were definitions, mainly copied from the textbook or a dictionary. Students were never asked, for example, to contrast and compare concepts (for example, mixtures and solutions), or to apply them in different contexts (to improve transferability of knowledge). In sum, notebook entries were not intellectually challenging or coherent. Notebook entries were mainly a set of unconnected activities within each unit that could hardly reflect an alignment between the tasks and the unit goals.

The fact that teacher feedback was not found in any of the classrooms scored for this study, tells something about the teachers' classroom assessment practices. Some readers may argue that feedback might be provided verbally during the discussion of the notebook entries in the class. If this was true, and we are sure in some cases it was, why was there no improvement on students' communications over the course of instructions? How much do teachers know about effective feedback and its impact in improving students learning (for example, Black and Wiliam 1998, Sadler 1989, 1998)? Time constraints may be an argument that almost any teacher can make for not providing feedback to students. To overcome this problem, we believe that first teachers need to carefully select the type of entry to work with students. (Which types of entries are useful and effective for promoting understanding, scientific inquiry, and improving performance?) Second, teachers need to think of options for assisting their assessment practices and helping students moving toward self-monitoring (for example, self- and peer-assessment, Sadler 1989). Third, the educators and research community need to think carefully about how science notebooks can be conceptualized, implemented, and assessed in forms that most effectively reflect their main purpose. If science notebooks are to be used as an unobtrusive assessment tool we need to make an effort to help teachers coordinate the power of purposeful recording and thoughtful reflection about students' work with helping students improve their understanding and performance, as well as the meaning of science inquiry.

References

Audet, R. H., Hickman, P. and Dobrynina, G. (1996). Learning logs: A classroom practice for enhancing scientific sense making. *Journal of Research in Science Teaching*, 33(2), 205–222.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, 21, 49–97.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.

Butler, R. (1987). Task involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology*, 79(4), 474–482.

Bybee, R. W. (1996). The contemporary reform of science education. In J. Rhoten and P. Bowers (Eds.), *Issues in science education* (Arlington, VA), 1–14.

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

Duschl, R. (2003). Assessment of inquiry. In M. Atkin and J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (Arlington, VA: NSTA press), 41–59.

Fellows, N. (1994). A window into thinking: using student writing to understand conceptual change in science learning. *Journal of Research in Science Teaching*, 31(9), 985–1001.

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

Halliday, M. A. K. and Martin, J. R. (1993) (Eds.). *Writing science: literacy and discursive power* (Pittsburgh, PA: University of Pittsburgh Press), 166–202.

Hanrahan, M. (1999). Rethinking science literacy: enhancing communication and participation in school science through affirmational dialogue journal writing. *Journal of Research in Science Teaching*, 36(6), 699–717.

Heinmiller, B. (2000). Assessing student learning–and my teaching–through student journals. Eisenhower National Clearinghouse web page: www.enc.org/focus/topics/assessment/articles/a06/index.htm.

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

KEYS, C. W. (1999). Language as an indicator of meaning generation: an analysis of middle school students' written discourse about scientific investigations. *Journal of Research in Science Teaching*, 36(9), 1044–1061.

LEMKE, J. L. (1990). *Talking science. Language, learning and values* (Norwood, NJ: Ablex Publishing Corporation).

LI, M., RUIZ-PRIMO, M. A., AYALA, C. and SHAVELSON, R. J. (2000, April). *Study of the reliability and validity of inferring students' understanding from their science journals.* Paper presented at the AERA Annual Meeting. New Orleans, LA.

LOZAUKAS, D., and BARELL, J. (1992). Reflective reading. A journal for inquiring minds. *The Science Teacher,* 59(8), 42–45.

LUCIDO, H. (1992). Physics for mastery. *The Physics Teacher*, 30, 96–101.

MARKS, G. and MOUSLEY, J. (1990). Mathematics education and genre: dare we make the process writing mistake again? *Language and Education*, 4(2), 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 and J.R. Martin (Eds.), *Writing science: literacy and discursive power* (Pittsburgh, PA: University of Pittsburgh Press), 166–202.

MCCOLSKEY, W. and 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 (1995). *National science education standards.* (Washington, DC: National Research Council).

PENROSE, A. and KATZ, S. (1998). *Writing in the sciences* (New York: St. Martin's Press).

RIVARD, L. P. (1994). A review of writing to learn in science: implications for practice and research. *Journal of Research in Science Teaching*, 31(9), 969–983.

RUIZ-PRIMO, M. A. (1998). *On the use of students' science journals as an assessment tool: a scoring approach* (Stanford University, CA: School of Education).

RUIZ-PRIMO, M. A. and SHAVELSON, R. J. (1996a). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6), 569–600.

RUIZ-PRIMO, M. A. and SHAVELSON, R. J. (1996b). Rhetoric and reality in science performance assessments: an update. *Journal of Research in Science Teaching*, 33(10), 1045–1063.

RUIZ-PRIMO, M. A., LI, M., AYALA, C. and SHAVELSON, R. J. (1999, March). *Student science journals and the evidence they provide: classroom learning and opportunity to learn.* Paper presented at the NARST Annual Meeting. Boston, MA.

RUIZ-PRIMO, M. A., LI, M. and SHAVELSON, R. J. (2002a). *Looking into students' science notebooks: what do teachers do with them?* Paper submitted for publication.

RUIZ-PRIMO, M. A., SHAVELSON, R. J., HAMILTON, L. and KLEIN, S. (2002b). On the evaluation of systemic education reform: searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 369–393.

SADLER, R. D. (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(1), 77–84.

SHEPARDSON, D. P. and BRITSCH, S. J. (1997). Children's science journals: tool for teaching, learning, and assessing. *Science and Children*, 34(5), 13–17, 46–47.

SOLANO-FLORES, G. and SHAVELSON, R. J. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practices*, 16, 16–25.

STECHER, B. M. and KLEIN, S. P. (1995). *Performance assessment in science: hands-on tasks and scoring guides* (Technical Report DRU-1087-NSF) (Santa Monica, CA: RAND Corporation).

WIGGINS, G. P. (1993). *Assessing student performance. Exploring the purpose and limits of testing* (San Francisco, CA: Jossey-Bass Publishers).

Appendix A: description of the performance assessments by proximity to the curriculum

To provide an idea of what close, proximal, and distal assessments are, we briefly describe the units and the assessments used. In the Variables unit students design and conduct experiments; describe the relationship between variables discovered through experimentation; record, graph and interpret data; and use these data to make predictions. Students identify, control variables, and conduct experiments using four multivariable systems, each corresponding to an activity: Swingers, Lifeboats, Plane Sense and Flippers. Students construct all of the system they use. For example, in the Swingers activity students construct a pendulum and in the Lifeboats, they construct boats of different capacity with paper cups.

In the Mixture unit students gain experience with the concepts of mixtures and solutions, saturation, concentration, and chemical reaction. One concept is the focus of each activity in the unit: Separating Mixtures, Reaching Saturation, Concentration and Fizz Quiz. During the unit students make up mixtures and solutions, use different methods to separate mixtures, determine the amount required to saturate certain volume of water, determine the relative concentrations of several solutions, and observe changes in substances by mixing solutions.

The table below provides a brief description of the assessments of different proximity by unit. To establish the proximity of the assessment tasks to the central characteristics of the curriculum, we judged their goals, content, and characteristics. For details about the assessments see Ruiz-Primo et al. (2002b).

Appendix A

FOSS Units	*Close*	*Proximal*	*Distal*
Variables.	*Pendulum* assessment asks to identify the variable that affects the time it takes a pendulum to complete 10 cycles. 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 focuses on the correctness of the variable identified, the accuracy of the students' measurements, and the correctness of the data interpretation, and the accuracy of the prediction required. This assessment was adapted from the one used by Stecher and Klein (1995).	*Bottles* assessment asks students to create a classification system that allows them to predict and explain whether an object will sink or float in tap water. The scoring system focuses on the accuracy of the observations and descriptions, and the accuracy of their predications and explanations. This assessment was adapted from the one developed by Solano-Flores and Shavelson (1997).	*Trash*, developed by the California Systemic Initiative Assessment Collaborative–CSIAC.* The instructional and assessment tasks differed in many ways, as is common in large-scale testing programs. The assessment task was sampled from a different domain, physical science. Few of the topics learned in the unit (for example, variables, systems, controlled experiment) were relevant to the assessment task. And the problem, procedures, materials, and measurement methods differed from those used in instructional activities.
Mixtures	*Saturated Solution* assessment asks students to find out which of three powders was the most and the least soluble in 20 ml. of water. They are asked to provide information about how they conducted the investigation, the results they obtained, and two other questions about solubility (for example, how they can dissolve the maximum possible powder in a saturated solution). The scoring system focuses on the accuracy of the results and the quality of the procedure used to solve the problem. This assessment was developed for the proximity study.	*Mystery Powders* assessment has two parts. In Part I students are asked to examine four powders using five tests (sight, touch, water, vinegar and iodine). In Part II, students are asked, based on their observations, to find the content of two mystery powders. The scoring system focuses on the accuracy of the observations and descriptions, the quality of the evidence provided (confirming, disconfirming, and other), and the accuracy of their answers. This assessment was developed for the proximity study.	*The CSIAC assessment is developed based on the standards proposed on the National Science Education Standards (National Research Council 1996) and the Benchmark for Science Literacy (AAAS 1993) and supports the learning goals of different systemic initiatives funded by NSF.

Appendix B: Number of entries by class and unit

Classroom Rank	*n*	*Unit*	
		Variables	*Mixtures*
Top			
Class 1	6	197	119
Class 2	6	209	78
Middle			
Class 3	6	72	119
Class 4	6	90	167
Low			
Class 5	6	52	65
Class 6	6	135	147
All	36	755	695