Examining the Utility of Elementary Science Notebooks for Formative Assessment Purposes

Pamela Aschbacher and Alicia Alonzo

Caltech Pre-College Science Initiative (CAPSI)
California Institute of Technology

This article explores the potential of using students’ science notebooks to help elementary teachers monitor conceptual understanding. Data came from 25 teachers and their 4th- and 5th-grade students working on a unit on electricity. At the time data were collected, 8 of the teachers had received professional development on science content knowledge, unit learning goals, assessing student work, and feedback. Measures included conceptual understanding of circuits from students’ notebooks, a performance task, and a multiple-choice test. Classes of teachers with and without professional development are compared on these measures, and notebooks of students in trained teachers’ classes are examined in some detail. Qualitative data from teacher interviews and observations of professional development sessions help explain the findings and identify teacher support needs. Notebooks have potential for formative assessment, but this use is limited by the guidance teachers give students, itself a function of teachers’ knowledge of science content and learning goals.

In the last 20 years, national efforts to reform science education have advocated a significant shift away from memorization of factual knowledge toward deep conceptual understanding and the ability to carry out and appreciate scientific inquiry (American Association for the Advancement of Science, 1993; National Research Council, 1996). At the same time, policymakers have tried to improve student achievement through large-scale accountability testing, but a number of researchers and educators have raised concerns about the negative effects of this approach on schools and students (Aschbacher, 2002; Herman & Golan, 1991; Linn, 2005).

The success of reform efforts ultimately depends on what happens in the class-
room, and instruction can be very complex and difficult to change in significant ways (Aschbacher, 1994; Cuban, 1993). Some researchers have begun to advocate the use of formative assessment by classroom teachers as a potentially powerful method to improve teaching and learning (Black, 1998; Black & Wiliam, 1998; Harlen, 1990). This article explores the potential of a particular source of information about student learning—science notebooks—which could be used as a formative assessment tool to improve teaching and learning.

Formative assessment is an integral part of the optimal teaching/learning process (Bransford, Brown, & Cocking, 2000; Herman, Aschbacher, & Winters, 1992; Pellegrino, Chudowsky, & Glaser, 2001; Stiggins, 2002). Its purpose is to determine what the student understands and why, so that teaching and learning may be optimized. Formative assessment can be formal or informal and includes eliciting, analyzing, and responding to information about student understanding (Ruiz-Primo & Furtak, 2004). The response phase involves refinement of future teaching plans and supporting students’ use of assessment information, by providing feedback and supporting students’ response to that feedback to improve learning. Without information about student understanding, teachers’ efforts to improve their teaching and student learning are limited. Reflecting on practice without looking at student work is unlikely to help teachers know how to adjust their instruction, because teachers do not necessarily recognize important differences between their attempts at reform practice and what was intended by reformers (Spillane & Zeuli, 1999). Even if teachers were able to accurately self-assess, the ultimate measure of success for educational reforms is the effect on students. To maximize their own effectiveness and evaluate how well their efforts are succeeding in the classroom, teachers must consider what was actually understood by the students (Hein & Price, 1994).

Just as traditional forms of teaching are inadequate to address the full array of learning goals embodied in current reform efforts, traditional forms of assessment may not be sufficient to assess the types of understandings students are expected to demonstrate (Herman et al., 1992). This article explores the potential of students’ science notebooks as a tool to support elementary teachers’ formative assessment practices—that is, monitoring and facilitating their students’ complex understanding of science concepts and inquiry processes.

The use of notebooks (collections of student writing before, during, and after hands-on investigations) as a part of science instruction has been encouraged in many school districts in the last 20 years. This approach is supported by a number of researchers who advocate writing in science to enhance student understanding of scientific content and processes, as well as general writing skills (Bass, Baxter, & Glaser, 2001; Baxter, Bass, & Glaser, 2000; Keys, Hand, Prain, & Collins, 1999; Rivard & Straw, 2000; Shepardson & Britsch, 1997). Additional support for the instructional value of notebooks comes from the literature on expertise, which suggests that competence is developed through active construction of knowledge, in-
cluding explaining concepts to oneself or others (Chi, 2000). Given their role in instruction, science notebooks should have potential as a formative assessment tool for at least two reasons:

1. Notebooks are an embedded part of the curriculum, so they are a ready source of data for teachers, who can obtain information about student understanding at any point without needing additional time and expertise to create their own quizzes. In notebooks, teachers can find evidence about a particular student, a sample of students, or the whole class.

2. Because they are embedded in the curriculum, notebooks are a direct measure of student understanding of the implemented curriculum, and are thus particularly relevant for formative assessment purposes.

Despite notebooks’ potential as a formative assessment tool, however, research has indicated that teachers seldom use them that way (Alonzo & Aschbacher, 2005; Baxter et al., 2000; Ruiz-Primo, Li, & Shavelson, 2001). This is not surprising, given that few teachers have much, if any, background in assessment, let alone formative assessment (Stiggins, 1991). In addition, most districts do not have the expertise or resources to provide teachers more than very limited guidance in science education. Teachers are sometimes told that since scientists keep notebooks, students doing inquiry should do likewise, but teachers are seldom supported in learning how to use student notebooks as evidence in adjusting their teaching to improve learning (Aschbacher & Roth, 2002).

Several researchers have explored the value of science notebooks relative to other methods of assessment. For example, Baxter and Shavelson (1994) found that scores on science notebooks were more comparable to observations of students doing hands-on tasks (group means were similar and correlations were between .75 and .84) than were other assessment methods (correlations of simulations and multiple-choice tests with observations of hands-on tasks were only .28 to .53). They concluded that notebooks were useful tools and that each assessment method may measure slightly different aspects of achievement. Sugrue, Webb, and Schlackman (1998) suggested that for complex concepts, open-ended written responses can uncover more misconceptions than multiple-choice items reveal.

To improve students’ conceptual understanding, teachers need to be able to monitor and diagnose student understanding of specific concepts. This article examines the utility of notebook evidence of specific conceptual understanding for formative assessment. If teachers were to examine science notebooks, what could they infer about student understanding? We also consider how particular teacher practices may affect what students record and the inferences that can be drawn about what students know, which may, in turn, affect the utility of notebooks as a formative assessment tool. Finally, we explore some of the challenges that teachers face in using notebooks for this purpose and discuss implications for professional development.
Our Approach

This research was part of a 3-year design experiment to study how purposeful use of science notebooks might improve teaching practices and student achievement in elementary science. Four districts that had already begun to promote the use of science notebooks participated in the research. A collaborative team of science coordinators, teachers, scientists, and researchers developed and studied several iterations of a model for notebook design and implementation to accompany the fourth- and fifth-grade hands-on inquiry science curricula used in these districts.

In this work, we defined student notebooks as a set of student writings and drawings that describe and reflect inquiry experiences as they occur within the science classroom. In professional development for these teachers, we advocated the following notebook entries: a research question (also called focus question) for the inquiry, records of data collected during an investigation, organization of data to facilitate analysis, knowledge claims that address the original research question, and evidence that involves data and reasoning to support the knowledge claims. Notebooks may also include other entries, such as predictions, description of materials and procedures, critiques of methods, and questions for future investigation, but we judged the first set of entries to be the most important for the initial professional development on notebooks.

To analyze the formative assessment potential of science notebooks produced in this study, we posed three questions:

1. How well did notebook scores predict other measures of students’ conceptual understanding?
2. How did teachers’ patterns of notebook use affect the inferences teachers might draw from them—that is, their utility as a formative assessment tool?
3. What factors inhibited teachers’ use of notebooks to assess students’ conceptual understanding?

Method

This article is based on data from 1 year (2002–2003) of a 3-year study of elementary science notebook use. The data discussed here reflect work in a hands-on inquiry science unit about electrical circuits, Circuits and Pathways by Insights (Education Development Center, 1990), used in each district in the study.

Sites and Participants

Elementary teachers and classrooms in the study came from four districts in California that participated in a network of districts trying to establish and maintain
hands-on elementary science programs. According to district data, about two-thirds of the students were from low-income families and about one-third were English language learners.

Teachers and administrators were feeling the pressures of accountability testing in language arts, which resulted in reduced time and resources for science. The four districts had been encouraging, but not requiring, teachers to use science notebooks for a few years prior to this study. Districts were unable to provide much, if any, professional development for teachers or principals about why or how to use notebooks in science. All four districts were eager to join the study on how to use science notebooks more effectively, and they hoped that improved use could support science and literacy goals.

Two groups of teachers and their fourth- or fifth-grade students (depending on the grade at which the unit was used in the district) participated in this study. We refer to them here as (a) the Protocol Teachers (PTs), who had received about 25 hr of specially designed professional development on notebook use over 2 years, and (b) the Regular Teachers (RTs), who had not yet received any special professional development from the project.

The PTs \((n = 8)\) were selected by their district science coordinators and principals as experienced teachers with the potential to provide professional development to other teachers after the study. They averaged 15 years of experience \((\text{Range} = 7–25\) years\). Four of the PTs had already helped provide science kit training to other teachers in their districts. All but one had taught the target unit before, and all were experienced in teaching hands-on science at the target grade level and using notebooks according to their districts’ minimal recommendations. The PTs participated in our professional development, implemented project ideas about notebooks during the target unit, and participated in interviews about their work.

The RTs \((n = 17)\) provide a comparison group that represents baseline or typical district practice, without special project professional development at the time data were collected. All had received a few hours of district professional development on how to use the science kit, which typically involved limited modeling of notebook use. We included experienced and relatively novice teachers in the RT group, to see whether teaching experience influenced changes in practice or student work. Nine of the RTs were comparable to the PTs in experience, with an average of 13 years \((\text{Range} = 4–27\) years\). They had all taught the unit and grade level before. The other eight RTs were relatively new to the teaching profession (six had 0–3 years prior teaching experience) or novice with regard to the unit (two had 5 and 7 years teaching experience but had not taught the unit or grade level before). The RTs taught the target unit and used science notebooks as recommended by their districts, but had no special professional development from the project. They worked in different schools from the PTs and had little, if any, contact with them during this part of the project. There was no significant difference in achievement on the pretest between classes taught by the experienced and novice RTs, so analy-
sis of student achievement here combines these subgroups into one group of RT classes.

Professional Development

PTs participated in about 25 total hr of professional development provided by the project over 2 years. Each year, the PTs attended a 1-day workshop prior to teaching the target unit, another 1-day workshop toward the middle of the unit, and 2 to 3 hours in after-school study group meetings at their schools during the unit. Teachers were paid $150 per day stipends for the professional development workshops and $25/hr for the study groups, which occurred outside normal work hours.

The professional development experiences were conducted by a collaborative team of researchers, district science coordinators, and scientists. The workshops included hands-on learning experiences for teachers that articulated and taught:

- the science content of the unit;
- the learning goals of each lesson and how they relate to and build an understanding of big ideas;
- the purposes and content of science notebooks as they relate to the nature of scientific inquiry;
- analysis of student notebook entries to assess learning and revise instruction to help students achieve the learning goals; and
- productive feedback strategies and rationale for providing feedback.

We did not advocate any particular grading strategies because grading practices and policies varied across districts. Study group meetings, conducted by one or two researchers with one to three teachers, provided additional opportunities to examine students’ notebooks for evidence of conceptual understanding, to extend teachers’ own science content knowledge through hands-on experiences and discussion, and to obtain individualized help.

Data Sources

Student outcome data for this article came from ratings and analysis of students’ science notebooks, scores on a multiple-choice posttest of unit concepts, and scores on “What’s My Circuit?” (WMC), a performance assessment embedded in the unit. Each of these data sources is described in the following. We used the work of 10 students per class, randomly sampled from those who had completed all assessments and had parent permission to participate. A few students who were unable to read or write English were excluded from the sampling.

**Notebooks.** We developed a set of scoring rubrics to rate students’ conceptual understanding of three big ideas from the “Circuits and Pathways” unit: sim-
ple, series, and parallel circuits. Each concept scale addressed several subideas that related to the central concept (e.g., the effect of removing a bulb from a parallel circuit). For each concept, students were given a point if their notebooks contained drawings showing that they had successfully constructed the type of circuit in question. Additional points were awarded for each of the subideas that were evident in entries for relevant notebook lessons. Most subideas could be demonstrated through either text or labeled drawings. A high level of English language proficiency was not required to receive credit for understanding a subidea. The simple circuit scale has a total of 7 points (i.e., it contains 6 subideas), the series circuit scale has 5 points, and the parallel circuit scale has 4 points. To facilitate comparisons across scales, student performance is reported as a percentage of possible points for a given concept.

All notebooks were read by at least two raters. One of the scientist–researchers that helped create the scales trained two research assistants (RAs) with strong science knowledge to score the notebooks. All three raters scored the first four classes of notebooks and reached 80% or greater agreement on each one. Because it was easy to overlook evidence in student notebooks for a variety of reasons, both RAs read each of the sampled notebooks in the remaining classes and came to consensus on the scores. As they worked, the scientist–researcher scored 20% of the 10 sampled notebooks per class and met regularly with the RAs to check the reliability of their ratings and to recalibrate as necessary. The two RAs reached exact agreement 81.6% of the time on whether a student could build a given type of circuit and reached 82.2% ±1 agreement on whether a student showed evidence of content understanding of subideas over the three scales. Consensus of the RAs’ ratings with the scientist–researcher’s ratings was 88.9% (exact agreement) on circuits built and 81.3% (±1 agreement) on subideas across the three scales.

We also examined each class set of 10 sampled notebooks for the eight PT classes in a more qualitative and holistic way, to try to observe and infer several aspects of how these teachers had used notebooks:

- how much guidance they had likely provided students about recording data and writing claims;
- whether they had asked students to copy entries (i.e., all notebooks in a class looked virtually identical for certain entries and sometimes used terms that students were unlikely to have generated);
- the extent to which the writing was aligned with the focus question for a given lesson;
- whether the teacher provided written feedback; and
- whether students responded to this feedback in their notebooks.

Rubrics for the notebook study are available at www.capsi.caltech.edu/research/
We did not examine the baseline RT class notebooks in this manner because their entries tended to include mostly copied questions and predictions, lists of materials, and little, if any, careful data records or meaning-making.

**Multiple-choice test.** We developed a 14-item test of unit concepts, drawing on ideas from assessments used in research by Palincsar and Magnusson (2000) for their Guided Inquiry Supporting Multiple Literacies unit on electricity and in research by Sugrue et al. (1998), as well as the pre–post assessment included in the Insights curriculum. Of the items, 11 were relevant to the three concepts explored here: 5 items addressed knowledge of simple circuits, 4 items addressed series circuits, and 2 addressed differences between parallel and series circuits. The same test was given as a pretest just prior to the unit and as a posttest immediately after the last day of the unit.

**Performance assessment.** WMC was originally designed by Jerry Pine and Gail Baxter (J. Pine, personal communication, October 5, 2001) to accompany the target unit in one district and was later adopted as an embedded assessment in other districts. We adapted the assessment and administered it to students in the middle of the unit, after the lessons on series circuits that it addresses. In WMC, students work independently to figure out the configurations of two mystery circuits (each built inside a black box with only a lightbulb showing). They are given similar materials to make their own circuits for comparison purposes. The tasks related to series circuits used in this analysis are: drawing three circuits accurately (6 points), explaining their reasoning (3 points), and creating a generalized rule for the brightness of bulbs in a series circuit (3 points).

To better understand the ways that the project professional development had impacted PTs’ thinking about science notebooks, as well as the challenges they faced in using science notebooks in the ways we were advocating, we interviewed four of the PTs individually at the end of the year. The interviews included questions about whether/how they used student notebooks for formative assessment, challenges they faced in using science notebooks, and what they had gained from the project to date. Interviews were audiotaped, transcribed, and analyzed for themes related to the purpose of this article.

**RESULTS AND DISCUSSION**

We now present our findings and related discussion for each research question, and then discuss our overall conclusions in a subsequent section. There were no significant differences in achievement due to grade level, so we included fourth- and fifth-grade classes in the same analyses. Also, there were only very small differences
(1–6%) among the PT, experienced RT, and novice RT classes’ scores on the multiple-choice pretest, so they were all quite comparable at the beginning of the unit.

1. How Well Do Notebook Scores Predict Other Measures of Students’ Conceptual Understanding?

Linear regression analyses on all 25 classes combined (N = 245 students) confirmed that the notebook scores predicted performance on other measures (the posttest and WMC performance assessment), but they accounted for only a very small amount of the variance in each case (see Table 1). When regressions were run to compare the predictive value for PTs’ students (n = 77) and RTs’ students (n = 168), notebooks scores were consistently more predictive for students in PT classes than in RT classes, suggesting that professional development may have helped to produce notebooks that are a somewhat more accurate reflection of student knowledge.

There are several good reasons why notebooks might have rather limited correlations with the multiple-choice posttest yet still be useful for formative assessment. Notebook entries reflect learning in real time, whereas the posttest is a summative measure, completed at the end of the unit. Conceptual learning is not usually immediate—students do not usually grasp a concept fully as soon as it is introduced. In fact, one reason notebooks might be useful in formative assessment is that they provide a window on students’ emerging conceptions. Teachers can use them to help decide who needs extra help and when to revisit concepts or ideas in later lessons. And, whether or not teachers explicitly reteach concepts in later lessons, students themselves can have new insights about early concepts when they build new kinds of circuits in later lessons. This possibility is supported by the comparison of data from three sources for the series circuit concept. The notebook series scores and WMC series scores were obtained at about the same point in

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Dependent Variable</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB simple</td>
<td>Posttest simple</td>
<td>.058**</td>
</tr>
<tr>
<td>NB series</td>
<td>Posttest series</td>
<td>.036*</td>
</tr>
<tr>
<td></td>
<td>WMC series</td>
<td>.109**</td>
</tr>
<tr>
<td>NB parallel</td>
<td>Posttest parallel</td>
<td>.047**</td>
</tr>
</tbody>
</table>

Note. NB = notebook. WMC = “What’s My Circuit?” performance assessment.
*p < .05; **p < .01.
*N = 245.
time, and they are more highly correlated ($r = .33$) than the notebook series scores are to the posttest series scores ($r = .19$).

Another reason we might not expect notebooks to predict posttest scores well is that the two measures probably tap somewhat different abilities (Baxter & Shavelson, 1994). Indeed, notebook entries are meant to be constructed, generative responses, whereas the posttest is a multiple-choice test with selected responses. The test involves mostly recognition of drawings of three types of circuits, and of statements or drawings that represent ideas about the relative brightness of bulbs in different circuits and the effect of removing a bulb from a circuit. For example, it is quite possible a student could learn enough about circuits to recognize a series circuit drawing (as measured on the posttest) but not be able to accurately draw one himself (as measured by drawings in a notebook). In addition, the notebook rubrics are somewhat more comprehensive measures of the ideas that comprise two of the three concepts (series and parallel circuits) than is the posttest (or the WMC task for series circuits).

Figure 1 provides the PT and RT groups’ mean scores as a percentage of possible points for each measure for each concept. This figure illustrates several findings. Students in PT and RT classes received lower scores on the notebook scales than on the multiple-choice items assessing the same content; this is consistent with the idea that generative work is typically more complex and difficult than recognition of correct answers and may require more language skill. In addition, even though notebook scores were lower overall than multiple-choice test scores, the notebooks seemed slightly more sensitive to the effects of professional develop-

---

2Both two-tailed Pearson correlations are significant at $p < .003$, $N = 245$. 

---

FIGURE 1 Student scores on multiple measures of conceptual understanding in baseline (RT) and trained (PT) teachers’ classes.
ment than the multiple-choice test was. Students in PT classes scored 24% higher than those in RT classes, averaged across the three notebook scales, but only 17% averaged better over the three posttest scales. The disproportionately lower notebook scores among RT classes may reflect poorer student learning and/or poorer teacher guidance to students about what to record in their notebooks, a point we explore in the next section.

2. How Do the Guidelines That Teachers Give Students Affect Student Work and Notebook Utility as a Formative Assessment Tool?

To pursue the link between teachers’ directions for notebook entries and the usefulness of notebooks for formative assessment, we focus on the group of eight PTs who had professional development but, nonetheless, varied in the way they used the notebooks and in the success of their students. Table 2 provides PT class mean scores (based on a sample of 10 students per class) for each measure of the three concepts. As in the previous analysis, all the scores were normalized to 100, so they represent the percentage of possible points obtained. For each of the three concepts, we found considerable variation from class to class on the notebook scores but much less variation in posttest scores. Some classes’ notebook scores seemed more aligned with posttest scores than others.

It occurred to us that notebook entries could be poor predictors of posttest performance if they were ambiguous indicators—sometimes underestimating and sometimes overestimating—what students really know. It would be possible to

<table>
<thead>
<tr>
<th>PTs</th>
<th>Simple Circuits</th>
<th>Series Circuits</th>
<th>Parallel Circuits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>Post</td>
<td>NB</td>
</tr>
<tr>
<td>Perez</td>
<td>29</td>
<td>74</td>
<td>60</td>
</tr>
<tr>
<td>Frey</td>
<td>20</td>
<td>77</td>
<td>14</td>
</tr>
<tr>
<td>Dale</td>
<td>13</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>Warner</td>
<td>44</td>
<td>69</td>
<td>36</td>
</tr>
<tr>
<td>Anen</td>
<td>46</td>
<td>78</td>
<td>22</td>
</tr>
<tr>
<td>Cruz</td>
<td>74</td>
<td>75</td>
<td>66</td>
</tr>
<tr>
<td>Emery</td>
<td>44</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>Johnson</td>
<td>51</td>
<td>74</td>
<td>10</td>
</tr>
<tr>
<td>PT average</td>
<td>40</td>
<td>73</td>
<td>35</td>
</tr>
</tbody>
</table>

overestimate what students know if teachers or researchers assume that ideas copied from the board or class discussions accurately represent students’ own thinking. On the other hand, it would be possible to underestimate what students know if they were given vague directions or poor writing prompts (Aschbacher, 1999). In that case, they might write about how much they enjoy experimentation or use some vague language about what they learned without providing any specific ideas about a target concept. In both cases, we (and teachers) would have a difficult time telling what students actually understand.

To consider these possibilities and the implications for the utility of notebooks for assessment, we used our intimate knowledge of unit lessons and examined class sets of notebooks to consider how teachers had used them and the quality of student work produced under these circumstances. We observed or inferred the following aspects of notebook use because they seemed likely to shape the potential of notebooks as assessment tools:

- **Target**: the relationship of student writing to the initial inquiry focus question for each lesson;
- **Guidance**: the nature of likely directions, guidance, questions, or writing prompts used by the teacher to elicit student notebook entries, including what entries should contain or what format they should take;
- **Originality**: whether student responses were likely to have been generated by students (e.g., evidence of actually trying to create and record circuits that address the focus question) or merely copied (e.g., identically worded statements in most of the 10 notebooks in a class); and
- **Feedback**: whether there was any written feedback from the teacher to the student and any written evidence of a response by the student.

For this analysis, we were less interested in whether teachers actually used the notebooks for their own assessment purposes (most did not) than in the notebooks’ potential as an assessment tool.

Our examination of how the PTs used notebooks revealed four typical patterns of use, somewhat reminiscent of the Goldilocks story—too little, too much, and just right:

1. minimal guidance to students in what to write, which is often very vague and seldom conceptually focused;
2. low guidance that provides a little structure but still insufficient focus;
3. overly prescriptive guidance that promotes student copying of “the right” information rather than expressing their own ideas; and
4. moderate guidance, the “just right” type of questioning and directions that provide conceptual focus, allow students to do their own thinking related to
the key unit concepts, and encourage an accurate and complete record of data and thinking in the notebook.

These patterns did not always characterize a teacher’s use of notebooks across all lessons. For example, one teacher, Mrs. Perez, used minimal guidance in lesson 3 but more moderate guidance in lesson 10. Based on our experience providing professional development and observing classes, the variations in patterns of use seem to reflect content knowledge of a particular concept and knowledge of a lesson’s learning goals as well as other variables that make teaching complex and inconsistent. We describe these four observed patterns below, and provide some notebook examples to illustrate the student work produced and the inferences that might be drawn from it about student understanding.

**Minimal guidance.** In this approach to science writing, the teacher typically gave directions that were quite vague, such as “write what you learned today,” with no guidance to students about what was important to have learned, how to frame a knowledge claim, or how to select and provide evidence to adequately support knowledge claims. In this situation, some students wrote a fair amount but tended not to mention crucial conceptual information, such as in the claim and evidence statements in Figure 2.

With a class set of similar-level work, we inferred that in Mrs. Perez’s lesson 3, students had probably not been given much guidance about what they should write for their knowledge claim and evidence after an investigation. This resulted in claims like the one in Figure 2, where the student simply asserts that he “can light the bulb” rather than making a knowledge claim that demonstrates some understanding of the big ideas of the lesson. For example, a high-quality claim for this

![Figure 2](image-url)

**FIGURE 2** Poor student response due to minimal teacher guidance.
lesson might be: “To make the bulb light, both the side and the bottom must be connected to the battery.” Solid evidence for this claim might include:

In circuits #1 and #2, the side and the bottom of the light bulb were connected to the battery with wires. If I removed the wire from the side of the bulb, the circuit wouldn’t work. If I removed the wire from the bottom of the bulb, the circuit wouldn’t work.

If the teacher had realized that her students’ claims lacked enough detail to reveal understanding, she might have thought about her directions to the students. Did her students understand what it meant to make a knowledge claim? To provide evidence for that claim? For example, she might have to repeat the task and ask something like the following:

Can you make a claim that gives a rule for how to make a bulb light up with these materials?—Don’t just tell us you can do it or tell us about one way to light the bulb. We want to know what it takes to make the bulb light up every time. Tell us which drawings show what you tried, and what worked and what didn’t. Then we can see if your evidence supports the claim you are making about how to light a bulb.

There are indications from the entries that Mrs. Perez had taught key ideas of the lesson, but these did not always appear in her students’ writing. It is quite possible that this student understands much more than is revealed in Figure 2. Neither the teacher nor we can know what students really know when the prompt or guidance is too minimal to elicit important details of knowledge.

The absence of drawings and sloppy or vague student work that tended to appear in students’ notebooks with minimal guidance seem to reflect teachers’ tentative grasp of science content. We observed during professional development that teachers with poor content knowledge often could not identify the big ideas of the lesson and were not sure how the lessons in the unit related to one another. Knowing what student notebook entries should look like is often more than a matter of having generalized criteria for specific components. For example, in lesson 3, because students were expected to be making knowledge claims about “critical contact points” (the points on the battery and bulb that must be contacted to light the bulb), it was important that they represented the light bulbs in ways that the “critical contact points” (side and bottom of the bulb) were clearly visible. Thus, the teacher’s knowledge of what is important in the lesson is required to focus students’ recording on aspects of their work that are likely to provide useful evidence for meaningful knowledge claims. Without fully understanding the purpose of a given lesson, teachers could not easily support students’ drawings (and other entries), nor gauge from student notebook entries what learning had and had not occurred. Before the professional development, they tended to focus on features of the notebooks other than what they revealed about student understanding—for
example, neatness and the quantity of drawings or writing. When teachers were unclear about the difference between writing fluency and conceptual understanding, they tended to overestimate learning when an entry was vague but somewhat fluent, as in Figure 2.

There is evidence from the notebooks that Mrs. Perez looked at student work. If she had used this information formatively to shape her instruction, we would expect to see higher scores on the posttest (the scores were, in fact, moderately high) despite low notebook scores for simple circuits.

**Low guidance.** Some teachers appeared to provide a little more guidance than the very vague directions just described, but with mixed results in student work. For example, notebooks reveal that some teachers asked students to refer back to their focus question when writing about what they learned. But the success of this strategy depends on the quality of the focus question and students’ understanding of the question and what sort of response is required to fully answer it. Without some discussion of the expected type of response, students may be writing on-topic without demonstrating what they really understand about key ideas. For example, in Lesson 4, students examine the inside of a lightbulb to see the pathway that electricity takes there, and then relate this understanding to the ideas of critical contact points and complete circuits, introduced in Lesson 3. A good focus question for Lesson 4 might be: “How does the pathway of electricity get completed inside the bulb?” A student could answer this question by writing, “By the wires.” He is clearly writing something on-topic; but, his response does not reveal much about his understanding of the path electricity takes inside the bulb.

In Mrs. Emery’s class, for example, students were asked to write a letter to another class explaining how electricity travels through the bulb. As Figure 3 illustrates, one student answered the question as intended, but another student, although clearly trying to answer the question, did not talk about the inside of the lightbulb at all. In the notebook evidence of simple circuit knowledge, students in Mrs. Emery’s class did better than those in Mrs. Perez’s class, but they did not do better on the multiple-choice test at the end of the unit. Did Mrs. Emery’s class really know more about simple circuits, or did her slightly better notebook directions merely make it more likely that they could show what they know in the notebook? When teachers see a range in student performance across the class, they cannot assume it is only due to students’ actions or characteristics. They should question whether their directions, expectations, and examples were clear enough for all their students to know how to produce a good response. For teachers who are unclear, themselves, about the content knowledge and/or a lesson’s learning goals, it is difficult, if not impossible, to be very explicit about what a good response entails.

**Overly prescriptive guidance.** The guidance teachers give students can reflect their deep beliefs about teaching and learning. In the case of the pattern we call overly prescriptive guidance, all the notebook entries in the class have virtually
the same wording, indicating that students copied from the board. According to interviews and professional development discussions, some teachers believe the primary function of science notebooks (and the primary learning goal of these science lessons) is to provide a repository of correct information that students can remember and refer to in the future, rather than a record of students’ own actions and thinking at the time (which might be more useful to the student and the teacher).
Teachers feel pressured by the accountability system to be sure that their students know the information that might be on the test or the science standards that students are supposed to learn in this grade. These teachers thus tend to have students copy the important or “right” information that the teacher has written on the board, in the belief that the act of copying will result in understanding or at least memorization. Thus, they may tend to overestimate what students really understand.

The way that teachers develop the information to be copied varies. Based on interviews and observations, some teachers have a class discussion first, and then write what they consider to be the best version on the board for students to copy. Others do not bother with the discussion and merely present the teacher’s right answer directly. In either case, at least some of the students miss the opportunity to ponder and articulate their own ideas or to discuss them with others. Notebook entries do not necessarily reveal whether there has been such a class discussion. One way for an external rater to tell whether the statements were student-generated (i.e., via a class discussion) or teacher-generated is to determine whether the copied statement is in kid-friendly language. When the exact language that was provided to teachers in professional development is used, we can be fairly certain that it did not originate from a student. Occasionally, we also found that a teacher with poor content knowledge produced supposedly right answers for students to learn that were actually incorrect.

From an assessment point of view, the problem with student copying is that the teacher and an external rater cannot be certain whether students developed an understanding of what they copied. If students copy well, the teachers (and we) may tend to overestimate what they actually know, and particularly what they are able to generate on their own. If students copy badly, the teacher (and we) cannot always distinguish language or copying problems from poor understanding.

Because students in Mrs. Cruz’ class all used much the same wording, as illustrated by the two students’ entries in Figure 4, we inferred that she had her students copy the claim for Lesson 10 about parallel circuits. Although her class earned about half the possible points for parallel circuit scores in their notebooks, they only got about 15% correct on the posttest for this concept, suggesting that the copying did not help them learn the concept. If a teacher does not have test score data to compare to notebook entries, she may not realize that copying did not serve them well and can only use her students’ notebooks for information about compliance (whether the students were following directions and copying down the required information). If she found that most students copied when she did not intend it, she might reflect that future instruction should stress the importance of students writing a question or claim in their own words. She may also reflect on the scaffolding that she has provided and whether it is sufficient to allow students to generate the notebook components independently. Perhaps students are copying her examples because they do not understand enough about the component to write their own.
Moderate or “just right” guidance. This approach provides enough structure for writing so that students know what concepts are important for them to learn and what kind of response demonstrates understanding. At the same time, students have the opportunity to learn from the exercise of articulating their own ideas in writing, and the teacher (or others) can use this record to monitor student understanding. In this more optimal learning condition, we would expect moderate guidance to yield notebook scores that are more closely aligned with posttest scores than they are under low or minimal guidance conditions, and that is what we found.

An example of moderate guidance comes from Lesson 10 on parallel circuits, from Mrs. Perez’ class. She appears to have structured students’ work so that they could focus on a particular concept (comparison of bulb brightness in series and parallel circuits). Because the lesson had a clear focus, students understood what they should be learning and, therefore, what they should write about. The students seem to have written their own thinking about the big ideas of the lesson. There was little evidence from the notebooks that there was much attention paid to a second big idea in the lesson, the effects of removing a bulb from a parallel circuit compared to a series circuit. The notebook rubric for parallel circuits gives students a point for creating a parallel circuit, and if that is done, further points for indicating the multiple pathways through the circuit, and for knowledge of the brightness and effects of removing a bulb. Because Mrs. Perez’s class addressed at least half of what was addressed by each measure and did so with good focus, we would expect a moderate relationship between the notebook scores and the posttest
scores. Indeed, the results were as expected. Her class did better than others on the notebook score and moderately well on the posttest score for parallel circuits (see Table 2). Examples of two students’ work in Figure 5 illustrate different student voices in their notebooks and conceptually focused responses. Both individual voice and conceptual content in the writing help the teacher assess what the individual student is really thinking.

3. What Inhibited Teachers’ Use of Notebooks to Assess Students’ Conceptual Understanding?

Time is the reason most frequently cited by teachers for why they do not often use notebooks to assess students. This is certainly an issue—especially because science is only one of many subjects for which elementary school teachers are responsible. Teachers need strategies—such as sampling—to reduce the amount of time required to gain valuable information from students’ science notebooks. Notebooks also take more time to examine when teachers do not have clear criteria for what entries should look like. Thus, although time is the stated barrier, there may be other underlying factors that contribute to this perceived challenge.

Our data indicate that two additional factors represent fundamental challenges to teachers’ use of science notebooks to monitor students’ development of conceptual understanding: (a) lack of knowledge of the science content and learning goals of each lesson, and (b) the view that their primary responsibility is covering the curriculum rather than teaching for conceptual understanding.

Teachers’ understanding and confidence with lesson content and learning goals critically affect the entire process: their view of what the lesson is about, their choices and shaping of activities, and the guidance they give students about what is to be learned and what to record in notebooks. It also affects whether and how teachers look at student understanding in the notebooks and whether and how they use that information formatively, to give feedback to students and to revise their own practice as needed.

The curriculum did not make it easy. The big ideas of the lessons and how they relate to ideas in other lessons to build and deepen conceptual understanding were not made sufficiently explicit in this unit (or three others from different publishers in the larger study), particularly not to accountability-stressed elementary teachers with little science background, minimal district training, and limited principal support for science. Observations of professional development discussions and classroom practice in the first year revealed that many of the experienced teachers (who had even trained other teachers to use this unit) did not have sufficient understanding of lesson content and learning goals to make informed choices about what to teach and how to teach it, or to set priorities about what lessons to omit due to lack of time. Therefore, we decided to more explicitly teach the unit concepts and learning goals and to give teachers multiple opportunities to deepen their understanding of those ideas in hands-on workshops and study groups in the second year, but the
**FIGURE 5** Two examples of focused, original student work due to moderate guidance.

---

**Q:** How bright can the bulb light if you make a particular circuit? Series or parallel?

- **Series:**
  - Circuit was the dimmest of the group. We used two bulbs and one battery. It was dim.
  - My evidence is what came up on my brightness meter and my drawing.

- **Parallel:**
  - I claim that our series circuit was the dimmest of the group. We used two bulbs and one battery. It was dim.
  - My evidence is the same brightness. My evidence is what came up on my brightness meter and my drawing.

---

**Claims and Evidence:**

- I claim that I know the difference of the parallel circuit and series circuit because it when I tried it out, the parallel is brighter because I think the battery gives out more power to the bulbs. My evidence is my pictures.
time available (about 12 hr of training) was evidently less than needed for this unit. Teachers certainly increased their understanding, but there remained much room for improvement. In interviews, several teachers noted how their understanding of the science content and notebook components affected how they taught and how they used the notebooks to monitor student learning.

Mrs. Cruz: I know that last year I didn’t pay as close attention to their drawing … to make sure that their drawings were actually touching those contact points. So, for this year I really spent the time and made sure and made them go back and readjust their drawings if they weren’t quite as accurate … . Just my comfort level with the whole unit, I think, plays in it a lot.

Ms. Anen: I loved when you did the in-service on “purpose question” and what makes a good “purpose question.” That was an eye-opener.

Mrs. Cruz: [This year] I’m better able to focus them in on what I’m looking for. So they don’t wander as much … . I’m asking them for specific things in their notebook and it’s easier to look to see if they have those things.

Mrs. Emery: I was looking for whether they explained the concepts, could restate it in their own words, kind of. I was looking for that claim and evidence.

Observation of teachers during professional development activities revealed that their content knowledge was linked to their use of feedback. For example, without an understanding of what makes a circuit complete and what student drawings should include to show that they understand this concept, teachers did not realize that sloppy student drawings were not just untidy but also lacked important details that support conceptual understanding. Teachers did not feel confident—and rightly so—assessing student learning and giving feedback to students until they, themselves, better understood the ideas of the lesson. In addition, they were more motivated to make time for providing feedback to students after they learned in professional development some basic principles of good feedback and some research findings on its effects on learning. They began to experiment with different strategies for giving feedback, but most still struggled with incorporating feedback into their teaching and would profit from additional time to learn how to do it well.

Mrs. Cruz: I think as long as I was clear on what the concept of the lesson was, then my feedback was okay. And I think the more times I’ve done this unit … it makes it easier. … I’m a little more comfortable now in what I’m looking for.
Mrs. Emery: I think this next year I’ll actually make time before we start our lesson. Give it 5, 10 minutes for them to actually go back and look at what I wrote. And then if there was a correction, to try to get it done. … And try to do something where they’re a little more accountable. … Sometimes some of them would go and erase something and then write it in correctly … just to show that “I did it right.” But it wasn’t the same as thinking through and explaining.

The teachers described feeling responsible for covering the curriculum, but they did not mention teaching for understanding of key ideas. In fact, interview data suggested that teachers tended to have faith that covering the curriculum would result in desired student learning (whatever it might be). For example, a teacher who did not use notebooks for formative assessment thought all was well and continued to emphasize what she felt was important, the scientific terminology. When shown her students’ posttest results later at professional development, she was not prepared for how poorly her students had performed.

Ms. Anen: It really surprised me. You think that they’re doing much better than they actually do.

Another teacher also indicated her uncertainty about student learning while she was teaching the unit and was grateful for the validation she felt from our feedback on her students’ scores:

Mrs. Cruz: I still think it comes back to what my comfort level is as far as understanding the content and what to look for in science. I’m not a science person. It’s just what I’ve learned though this [project]. … I know I was pleased to see the pre- and post- [test] with the increase, so I knew that for the most part, they were getting it, which is nice to see that they’re on the right track when you’re not sure what they’re doing.

Unfortunately, there is no quick fix. She and many of her colleagues need more learning experiences and coaching to deepen their knowledge and confidence so they can use science notebook evidence to derive feedback on the effects of their teaching. With this knowledge, they can optimize their future instruction and thereby improve student achievement.

CONCLUSIONS

In summary, notebooks have potential as a tool for formative assessment in the classroom. They can reveal student thinking, which teachers can monitor to shape instruction and improve learning. Notebooks can predict achievement on an
end-of-unit test of the same concepts, although notebooks may measure somewhat different aspects of performance than multiple choice tests, due to their real-time and generative characteristics.

However, despite this potential, common patterns of current practice restrict the accuracy and usefulness of notebooks for formative assessment. Teachers’ use of notebooks, particularly the type of directions or guidance provided to students, influences what students write in the notebooks and how informative these entries are with respect to student understanding. When teachers have students copy the right answers, they (or external raters) can overestimate students’ conceptual understanding. Our data suggest that copying the right answers from the board did not help students do well on the posttest. Constructivist learning theories would suggest that this may be because the students missed the opportunity to fully develop and articulate their own ideas. When teachers provide students too little guidance about what is important to learn and allow sloppy data records and vague writing, the teacher (or an external rater) can underestimate what students may have actually learned.

The nature of guidance that teachers provide students for their notebook entries is influenced by teachers’ own knowledge of science content and unit learning goals, as well as whether they feel responsible for student attainment of specific goals rather than for just covering the curriculum. Without this knowledge and commitment to supporting student understanding, it is difficult to evaluate student work, give effective feedback, and revise teaching plans where needed. The fairly modest amount of professional development in this study seemed to help some teachers focus on important learning goals, better understand the content, and give clearer guidance to students about what to put in their notebooks. Student notebooks of some trained teachers provided evidence that seems to be a more accurate reflection of what the students know (i.e., aligns more closely with a posttest). The amount and type of professional development here was thus a good start, but we suspect that teachers could continue to make important progress if more support were available on a regular basis.

Our findings suggest that teacher professional development needs to address more than just the kinds of writing that a science notebook might contain, which is a common approach where notebook training exists. Professional development must also help teachers to develop a solid conceptual understanding of the big ideas and awareness of key learning goals, as well as how to evaluate student work against those goals, give feedback, and revise practice based on this evidence—in other words: how formative assessment nestles within and nurtures good instruction and successful learning.

ACKNOWLEDGMENTS

Alicia Alonzo is now at College of Education, University of Iowa.
This material is based on work supported by the National Science Foundation under Grant REC–0106994. Any opinions, findings and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We greatly appreciate the students, teachers, and district science coordinators who participated in this study, and we also acknowledge current and past Caltech Precollege Science Initiative (CAPSI) team members—including Ellen Roth, Melanie Jones, Cameron Baker, Laurie Thompson, Marie Rodriguez, Brian Foley, and Jerry Pine—for their contributions to the project. We are also grateful to Shirley Magnusson and Annemarie Palincsar for their generous advice to our project.

REFERENCES


Aschbacher, P. R., & Roth, E. J. (2002, April). What’s happening in the inquiry science elementary classroom and why? In L. Suter (Chair), Policy levers for urban systemic mathematics and science reform: Impact studies from four sites. Symposium presented at the annual meeting of the American Educational Research Association, New Orleans, LA.


