Developing Expertise With Classroom Assessment in K-12 Science:
Learning to Interpret Student Work
Interim Findings From a 2-Year Study

CSE Technical Report 704

Maryl Gearhart [a], Sam Nagashima [b], Jennifer Pfotenhauer [a], Shaunna Clark [b], Cheryl Schwab [a], Terry Vendlinski [b], Ellen Osmundson [b], Joan Herman [b], Diana J. Bernbaum [a]

[a]Center for the Assessment and Evaluation of Student Learning (CAESL)/University of California, Berkeley

[b]Center for the Assessment and Evaluation of Student Learning (CAESL)/University of California, Los Angeles

December 2006

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by the National Science Foundation under grant ESI-0119790 to WestEd for the Center for Assessment and Evaluation of Student Learning (CAESL). Views expressed in the paper do not necessarily represent the views of the National Science Foundation.

DEVELOPING EXPERTISE WITH CLASSROOM ASSESSMENT IN K-12 SCIENCE:

LEARNING TO INTERPRET STUDENT WORK

INTERIM FINDINGS FROM A 2-YEAR STUDY

Maryl Gearhart [a], Sam Nagashima [b], Jennifer Pfotenhauer [a], Shaunna Clark [b], Cheryl Schwab [a], Terry Vendlinski [b], Ellen Osmundson [b], Joan Herman [b], Diana J. Bernbaum [a]

[a]Center for the Assessment and Evaluation of Student Learning (CAESL)/University of California, Berkeley

[b]Center for the Assessment and Evaluation of Student Learning (CAESL)/University of California, Los Angeles

Abstract

This article reports findings on growth in three science teachers' expertise with interpretation of student work over 1 year of participation in a program. The program was designed to strengthen classroom assessment in the middle grades. Using a framework for classroom assessment expertise, we analyzed patterns of teacher learning, and the roles of the professional program and the quality of the assessments provided with teachers' instructional materials.

The premise of this volume is that formative assessment is a critical component of effective instructional practice. To support student learning, teachers need to gather ongoing evidence of student progress, and use the information to give students helpful feedback and make appropriate adjustments in instruction. The importance of classroom assessment is represented in guides issued by professional organizations (National Research Council [NRC], 2001a, 2001b), standards for teacher practice (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990), and research on the effects of classroom assessment on student learning (Black & Wiliam, 1998; Brookhart, 2004; Crooks, 1988; Natriello, 1987; Shepard, 2001; Wiliam, Lee, Harrison, & Black, 2004). Clearly the need for teacher assessment expertise is well established, yet far less is known about the ways that teachers develop

the expertise. Our article addresses this gap with an analysis of the growth of three middle school science teachers during their first year in the Assessment Academy, a long-term program focused on classroom assessment.

The program's core strategy was an assessment portfolio that guided teachers in the design, implementation, and evaluation of assessments for curriculum units. In the first year, teachers completed two portfolios with support from facilitators, colleagues, and program readings and activities. We will focus on teachers' evolving expertise with interpretation of student work, and consider the supportive role of the portfolio, gaps in Academy resources, and the challenges teachers were faced with by weak assessment tasks and criteria. Our findings have implications for the design of professional development as well as the assessment resources provided in instructional materials.

The Academy Program: Context and Strategy

The National Science Education Standards envision classrooms where students engage in the processes and debates that constitute the professional work of scientists (NRC, 1996). It is a challenging vision, and classroom assessment is key (NRC, 2001a). To support the development of inquiry and conceptual understanding, teachers need to engage students with the conceptual content of inquiry activities, and scaffold learning by monitoring student understanding and refining instructional approaches based on sound information. Teachers should be using assessments that capture the full breadth of students' knowledge—of science concepts, inquiry processes, and science facts and vocabulary—as well as the alternative conceptions that students construct in the process of building understanding of complex science ideas (NRC, 2001b).

The vision places great demands on science teachers, who need deep understandings of science and conceptual development as well as expertise with assessment in order to develop quality assessment methods and interpret student understanding in sound and valid ways (Shepard, 2001). The premise of the Academy program that we followed was that teachers build this assessment expertise by working with colleagues and facilitators in a sustained program focused on teachers' instructional materials.

The Academy invited five districts to identify teams of three to four teachers (across grade levels) and one administrator to participate in a 3-year program. Participants were provided modest compensation, and districts approved release days. District teams convened several times a year to organize as cross-district grade-level

teams and collaborate on classroom assessments for curriculum units. The Academy's core strategy was an assessment portfolio that supported the grade-level teams as they designed, implemented, and evaluated assessments for units with the support of facilitators; teachers completed two portfolios their first year.[1,2] Developed in consultation with assessment specialists, the portfolio tasks and resources were the focus of institute activities, and guided teachers between institutes as they implemented assessments and reflected in writing on the quality and usefulness of the assessments.

The portfolio had three sections:

- Section I: With support from curriculum and assessment specialists, teams designed unit assessment plans. Teachers reflected on key assessment concepts as they evaluated the soundness and clarity of unit learning goals, the alignment of assessments with goals, and the capacity of assessments to elicit evidence of student understanding.
- Section II: In their classrooms, guided by portfolio prompts and resources (strategies for constructing criteria, models of whole class matrices), teachers analyzed the student work and used the information to revise instruction and provide students with feedback. Relevant assessment concepts included alignment of criteria with learning goals, capacity of criteria to capture the full range of student understanding, and reliability of scoring. While teachers worked on their own, facilitators made some classroom visits or met with district teams to discuss assessment issues.
- Section III: A follow-up institute was the context for grade-level teams to reconvene and critique their assessments with the support of Academy facilitators. The Academy framework of assessment concepts was a critical resource (Herman, 2005).

The Academy's premise that teachers can build knowledge and strengthen practice through collaboration on assessment has support in prior research (Aschbacher, 1999; Atkin & Coffey, 2003; Bell & Cowie, 2001; Black & Wiliam, 1998; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Falk & Ort, 1998; Gearhart & Saxe,

[1] Nolen and Taylor (Taylor, 1997) developed a somewhat similar portfolio-based approach to preservice preparation in classroom assessment, but their work was not a resource for the CAESL Academy.

[2] Further information on the portfolio is available from Kathy Diranna of WestEd at kdirann@wested.org

2004; Goldberg & Roswell, 2000; Laguarda & Anderson, 1998; Sato, 2003; Sheingold, Heller, & Paulukonis, 1995; Wilson & Sloane, 2000). As S. Wilson (2004) pointed out, teachers build common understandings of assessment tools, a professional language for assessment practice, and opportunities for reflection on practice and student learning. But many of these same studies point to challenges teachers face in understanding and applying assessment principles. For example, in one study, many teachers who scored Maryland performance assessments did not implement practices consistent with the state assessment framework despite explicit guidance; teachers created rubrics that were weakly aligned with the targeted learning goals, or vague and undifferentiated, or focused on additive or countable features of the product (Goldberg & Roswell, 2000). Laguarda and Anderson (1998) reported similar patterns from their evaluation of several Education Trust Standards in Practice projects. One factor that may have limited classroom impact in these studies was the weak relationship between the program and the curriculum that teachers were implementing (Shulha, 1999). The Academy strived to address this concern with a portfolio and institute model built around teachers' curriculum units.

Investigating the Development of Classroom Assessment Expertise: Provisional Framework

To guide our research on the role of the Academy portfolio in teachers' developing assessment expertise, we drafted a framework for expert practice and some guiding assumptions about teacher learning, and we shared drafts of the framework at institutes to build common understandings of assessment among all participants. Our framework for classroom assessment expertise integrates theory and research from both the psychometric and practitioner traditions (Shepard, 2001; Stiggins, 2005; Taylor & Nolen, 1996). Psychometric frameworks and standards provide important assessment constructs, though the challenge is to determine in what ways these are applicable to classroom practice (Brookhart, 2003; Taylor & Nolen, 1996). Practitioner research provides models of formative assessment practices (Black & Wiliam, 1998) and embedded assessment systems (e.g., Wilson & Sloane, 2000) that integrate assessment with reform practice. Drawing on these strands of theory and research, our framework captures relationships between understanding of *assessment concepts* and facility with *assessment practices.*

Understanding assessment concepts. There is a network of interconnected assessment concepts that we believe expert teachers need to understand. Quality

assessment requires clear and valued *goals* for student learning as the assessment targets, quality *tools* for gathering evidence of student learning, sound *interpretations* of the evidence, and quality *uses* of the information to guide instruction and provide students with useful feedback. (Herman, 2005, further specifies the elements of these core concepts of Goal, Tool, Interpretation, and Use.) Teachers must also recognize that all components of assessment must be *aligned,* and that any one assessment is embedded in a *system* of assessments to provide coordinated information.

Facility with assessment practices. Assessment is a critical component of a cycle of continuous instructional improvement. When expert teachers develop their plans for science units, they begin by identifying the learning goals and designing an integrated instructional and assessment plan. Implementation entails repeated cycles of instruction, assessment (using a variety of strategies ranging in formality and function), interpretation of evidence, and use of information to revise instruction or the assessment strategy. Alignment is critical; teachers need to revisit their learning goals and revise goals, instruction, or assessments periodically throughout the unit.

We think of the relationship between understanding and practice as the reflective questions teachers ask themselves in the context of assessment practice—the kinds of questions that Academy teachers considered with the support of their portfolios. When interpreting student work, for example, teachers need to consider such concepts as *developmental appropriateness* (can these criteria capture the full range of student conceptions?), *reliability* (how can I ensure that I score consistently?), and *fairness* (to what extent does my analysis of class performance reflect the opportunities of all my students to learn the curriculum?). "Looking at student work" is an assessment practice linked to a network of understandings and practices.

To understand how science teachers develop expertise with interpretation of student work, we draw on existing research on teacher learning. Constructivist accounts help us understand that teachers assimilate new ideas to prior knowledge and practices (Goldsmith & Schifter, 1997) and only gradually reconstruct what they know and do. To capture change, one approach is to track shifting relationships between teachers' practices and the functions they serve. In one such study, teachers who tried out new assessment tasks often graded student responses as correct/incorrect (an old function), while other teachers continued using old assessment tasks but invited students to explain their reasoning (a new function; Saxe, Gearhart, Franke, Howard, & Crockett, 1999). A second approach captures how teachers reorganize their understandings and practices. Davis (2003) showed how an elementary science teacher

gained expertise with inquiry teaching as she progressively added, linked, distinguished, and reconciled old and new ideas. Taking these research approaches together, our view is that teachers build expertise with interpretation of student work and the relevant assessment concepts when they repeatedly confront the need to align old and new understandings and practices (Goldsmith & Schifter, 1997).

The Academy's cycles of assessment implementation and reflection should help teachers integrate new assessment ideas, identify inconsistencies, and build greater coherence. In our analysis, we show how three teachers' growth was supported by repeated opportunities to revise methods of interpreting student works, but was impeded by gaps in the Academy program and by weak assessment resources in the instructional materials.

Methods

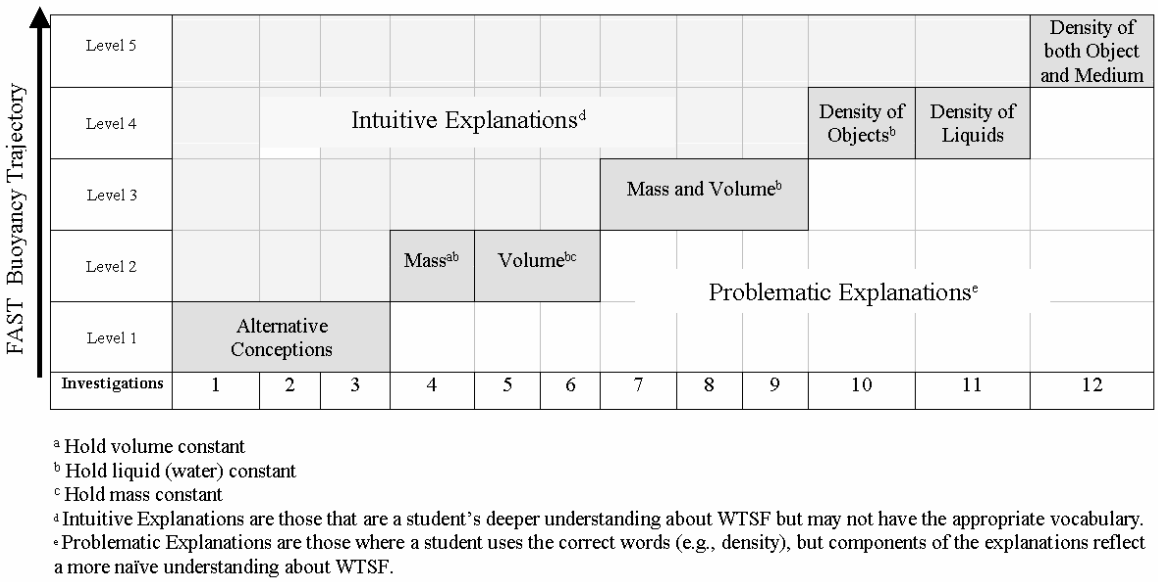
Participants

The case teachers were experienced professionals in middle-school science, comparable in science background, science specialization, years of teaching, and participation in science reform professional development programs. They differed, however, in the assessment resources provided with their instructional materials. In the fall of 2003, Yvette Jones and Carrie Green implemented off-the-shelf curriculum units, *Plate Tectonics* (Great Explorations in Mathematics and Science [GEMS]; Cuff, Willard, & Carmichael, 2001) and *Properties of Matter* (Science and Technology Concepts for Middle Schools [STC/MS], 2000).[3] Joan Reddy implemented a buoyancy unit from the Foundational Approaches to Science Teaching (FAST) series that included a new embedded assessment system designed by researchers to track the development of students' understanding of buoyancy and their reasoning ability.[4] She scored her assessments using rubrics reflecting two progress variables: The first reflects FAST's implicit developmental model for fostering student understanding of buoyancy (Figure

[3] All names are pseudonyms.

[4] Foundational Approaches in Science Teaching (FAST), a series for middle school science, was developed by the University of Hawaii Curriculum Research and Development Group (Pottenger & Young, 1992). The program is based on a constructivist philosophy of learning, is aligned with the national science standards (NRC, 1996), and uses carefully sequenced, student-conducted investigations to develop students' learning. The embedded assessments were designed by teams of researchers at Stanford and UC Berkeley with support from CAESL.

1); student explanations of buoyancy are scored at one of five levels. A second progress variable (not shown) was designed to measure growth in students' skill in reasoning from evidence. In the spring of 2004, all three case teachers implemented units of their own choosing, none with researcher-developed assessments.



[a] Hold volume constant
[b] Hold liquid (water) constant
[c] Hold mass constant
[d] Intuitive Explanations are those that are a student's deeper understanding about WTSF but may not have the appropriate vocabulary.
[e] Problematic Explanations are those where a student uses the correct words (e.g., density), but components of the explanations reflect a more naïve understanding about WTSF.

Figure 1. FAST Buoyancy Progress Variable: Developmental model describing students' evolving conceptions. From Stanford Education Assessment Laboratory (SEAL). (2005). Stanford, CA: Stanford, University. Reprinted with permission from Richard J. Shavelson, Stanford Education Assessment Laboratory, Stanford University.

Design and Instruments

This article draws on the teachers' portfolios and a series of interviews that we conducted throughout the first year of the Academy. The interviews targeted the teachers' initial assessment plans, their assessment implementation (which we observed), interpretations of student work, and final reflections after completion of each portfolio. To ensure depth in our understanding of each case teacher, one researcher followed the same teacher over time. To build common understandings and strengthen our capacity to challenge a case researcher's interpretations, most of us attended all institutes, and we debriefed our case work on a weekly basis. The first author has read all files.

Analysis

Analysis entailed iterative cycles of framework development and identification of both longitudinal and comparative patterns. Each case researcher reviewed portfolios, interviews, and observations longitudinally and drafted case memos, and secondary researchers and the first author discussed and challenged preliminary interpretations. We then entered all files and the portfolios into HyperRESEARCH databases [http://www.researchware.com/hr/index.html] (Hesse-Biber, Dupuis, & Kinder, 1991), and used descriptive coding to capture material relevant to interpretation of student work (e.g., criteria refinement, scoring, analysis of whole class patterns) as well as factors impacting or interfering with growth of expertise. HyperRESEARCH reports of the coded material were numbered consecutively by date to facilitate analysis of patterns of change.

Coding a given domain of knowledge or practice across time as if it were the same across time periods can be a problematic strategy. Material coded as "Criteria refinement," for example, might consist of a teacher's musings about how to score in August and the rubric she developed in November. Thus, while our descriptive coding has been useful in winnowing down material and examining patterns over time, we have read all material in preparing the analyses reported here.

Developing Expertise With Interpretation of Student Work

In both the fall of 2003 and the spring of 2004, Yvette Jones and Carrie Green represented the more typical Academy teacher, working with assessments that were generally weak in quality and usefulness; even the teachers' guides that provided some scoring schemes or tips on student conceptions did not provide methods for interpreting student progress. The Yvette and Carrie cases illustrate the kind of growth we can expect from teachers who are striving to analyze student work with limited resources. The third teacher, Joan Reddy, implemented a researcher-developed embedded assessment system for tracking students' understandings of buoyancy in the fall of 2003; she embraced the new approach thoughtfully but then confronted the need to develop her own assessments in the spring semester. Her case illustrates ways that quality materials support assessment practice but may not be sufficient as opportunities for teachers to develop more generalized assessment expertise. We argue that a synthesis of all three cases supports the need for both quality embedded assessment resources and professional development to support the growth of assessment expertise.

In the sections that follow, we present Yvette's case in greater detail and then use the other two cases to highlight contrasting patterns of learning.

Yvette Jones: Interpreting Student Work as the Impetus for Revising Many Components of Practice

Overview. During the 2003-2004 school year, Yvette taught math and science to three separate cores of sixth-grade students in a large, suburban middle school. Committed to improving her practice even after 15 years of teaching, she participated fully in the Academy and completed all portfolios in ways that were beyond what was expected, consistent with her detailed and reflective approach to teaching and assessment. Yvette viewed the Academy as an opportunity to strengthen the way she used assessment evidence to inform her teaching, provide feedback to students, and involve students in tracking their progress. Equity was a driving principle in her work, and indeed a member of the professional development team described Yvette as "the most equitable teacher I have ever observed." But in the first year of the Academy, Yvette confronted challenges in her efforts to use student work to track student progress. Her curriculum materials did not provide adequate assessment tools, and she had to devote long hours to developing and refining tasks and criteria aligned with her learning goals. Though she remained unsatisfied with her efforts during the first year, she strengthened her capacity to make more accurate inferences and responsive instructional decisions. Her greatest growth was her appreciation of the value of developmentally appropriate criteria for interpreting student work, as she expressed in her first year exit interview.

> The criteria for the assessment—it has to be based on concepts that you covered and allows the students to be detailed enough so that you're not making your own assumptions or inferences from the test. It should be used to encourage students . . . [I]f the test elicits students' understandings, then students can be encouraged or motivated or interested in what feedback you're going to give them specifically about where they're at in their learning. (7/04)

Fall 2003. When Yvette began her work with the Academy in August 2003, she adopted the *Plate Tectonics* unit (Cuff et al., 2001) and found it contained few strategies for assessment—whole class discussions and various lab investigations, but little that was identified as an assessment, and no resources to help her anticipate the range of student conceptions. Working with the materials and her previous experience teaching

plate tectonics, she chose activities that focused on key concepts and that she felt had assessment potential. Since these activities-turned-assessments did not have corresponding criteria in the curriculum materials, Yvette was forced to create her own in order to interpret student work, and she eventually revised these three times. The revisions represented progressive changes in several components of her assessment practices—goals, expectations, tasks, criteria—as she realigned these each time, gradually integrating new insights about assessment concepts such as developmental appropriateness, fairness, and accuracy.

The assessment that absorbed her attention was a GEMS group discussion that targeted relationships between lava viscosity, silica content, temperature, and volcano shape. In the summer of 2003, Yvette completed the planning matrix, identifying the concepts assessed and expected student responses (Table 1). At this time, Yvette interpreted the column for "expected student responses" (ESRs) as a request for greater specification of concepts, and did not use the column to reflect on student reasoning *about* the concepts; as a result, these ESRs were not resources for the construction of developmentally appropriate criteria. Two months later, she was still working on aligning goals and tasks; she replaced the group discussion questions with short-answer prompts for an individual assessment and identified the correct answers (Revision 1 in Table 1), but made limited progress toward criteria that capture student understandings of the targeted concepts. If a student wrote, for example, "When volcanoes have runny lava, there is a lot of lava, and that makes a bigger volcano," the student would simply get a zero, yet the response suggests a confusion between viscosity, velocity (speedy lava), and the relevance of the width versus height of the volcano.

Table 1

Yvette's Summer Plan and First Revision for the Volcano Assessment

	Summer 2003	Revision 1
Concepts assessed	Shield volcanoes are formed from low viscosity magma. Higher temperature magma comes from the deeper mantle of the Earth.	1. In the Hawaiian field study, all three lava flows the students investigate are derived from low viscosity (low silica) magmas, which indicates that the ultimate source of these magmas is the mantle. The least viscous flow represents magma from hotter, deeper areas of the mantle. 2. In Hawaii, material from the mantle is moving upward to form new islands within the Pacific Plate. Students plot other locations of shield volcanoes on the world map.

(*table continues*)

Table 1 (continued)

	Summer 2003	Revision 1
Expected student response What kinds of understandings do you expect from students? Will the students' response match the concept assessed? What are the criteria for good work? Are the criteria specified for the concept assessed?	Students understand that Hawaiian volcanoes are low viscosity, which causes the flow to spread out to form a shield. Students relate the magmas they experimented with to a low level of silica in the lava. Higher temperatures can cause magma to be less viscous.	Same

(*table continues*)

Table 1 (continued)

	Summer 2003	Revision 1
Criteria for evaluating responses What does the quality work look like?	None	The responses were scored correct/incorrect. Question 1. The source of magma comes from the mantle. It has low silica. Question 2. Temperatures tell how far or close it is from the core. Question 3. Batch 3 is like hottest magma. It flows the fastest. Question 4. The slowest batch 1 is cooler magma that is far away from the core.

Despite her persistent efforts to improve her assessments, like so many teachers with heavy teaching loads, Yvette still relied on informal evidence of student understanding. On the day she administered her volcano assessment, Yvette skimmed responses as students worked and as they turned in their papers. Immediately concerned that her students were challenged by the material, Yvette organized her students into small groups for a follow-up activity. She assigned each group one assessment question, and asked students to record their response on a whiteboard; as each group presented their work to the class, Yvette engaged the students in short discussion. This informal assessment had greater impact on her practice than any examination of the students' written responses; in her portfolio, Yvette wrote that there was "adequate evidence of students' understanding of the concept" because "the class seemed to have at least a ninety percent acquisition" based on whiteboards (YJ, Fall 2003 portfolio). Through her reflections in and about the portfolio, Yvette was deepening her commitment to using assessment information to interpret student learning and revise instruction. But she did not acknowledge possible bias in the group evidence she was using. Although the portfolio forms used in the fall of 2003 asked teachers to interpret student understanding, the forms did not require systematic analysis of written work, nor did they prompt reflection on possible sources of bias.

A month later at an Academy institute, Yvette met with colleagues and a facilitator to revise assessments, and Yvette chose to revise her volcano assessment based on analysis of her student work. Working to strengthen alignment of this assessment within her overall plan, she made no changes in the concepts assessed, but she did simplify the ESRs, included connections to classroom experiences, and modified the criteria by making them slightly more schematic and specifying point values (Table 2). These slight changes suggest her subtle shift away from focus on "the" correct answer toward identification of the core ideas underlying each item.

Table 2

Yvette's Second Revision of the Assessment

Expected student responses:

Higher temperatures can cause magma to be less viscous. Some familiar foods that become less viscous when heated are butter, cheese, chocolate, etc. Magma three is the least viscous because it traveled fastest down the slope of the volcano. Magma one is the most viscous because it had the slowest travel time.

Criteria for evaluating responses (attach rubrics, scales, etc. as needed):

Question 1. The low viscosity, low-silica content magmas come from the Earth's mantle. (1 point)

Question 2. Higher temperatures can cause magma to be less viscous. (1 point)

Question 3. Magma batch 3, because it traveled fastest down the slopes of the volcano. (1 point)

Question 4. Magma 1, the most viscous. (1 point)

She recognized that the Academy work of creating and using criteria was helping her to be more thoughtful and consistent when scoring student work. Reflecting on her

assessment practice prior to the Academy, she commented that she used to give credit in unsystematic ways.

> [I have] learned how crucial it is to have a rubric . . . I got better at using the rubric the more I created those—whether they were essays or whiteboards. Because in your mind you create a rubric and you are looking for some very specific answers . . . So having a rubric helped me to be consistent . . . I could see what the focused concepts were and give credit for that and not just, "Oh they tried to give their ideas." (11/03)

Her point scoring approach helped her give credit for the "focused concepts." But it also had limitations as she realized when she re-analyzed her student work with a facilitator and recognized that her criteria were not useful in identifying misconceptions.

> I looked at my assessment and found misconceptions, whereas when I was interviewed by you [earlier in the month], I didn't look for misconceptions. I was looking to see if they got the concept, yes or no. (11/03)

Yvette was growing in her ability to consider the reasoning underlying students' responses—the notion of "developmental appropriateness" in our framework. Unfortunately neither the Academy nor her instructional materials provided the guidance and resources she was seeking. "I wonder what's a strategy I could use to remedy some of those misconceptions. I want to pick someone's brain for awhile" (11/03). An Academy facilitator suggested she purchase the American Association for the Advancement of Science (AAAS) Benchmarks (AAAS, 1993) to help her think about reasonable expectations of student understanding in sixth grade. Unfortunately, while this resource was helpful in developing clearer, more appropriate learning goals, it did not help her to think about a range of student conceptions at her grade level.

Still not satisfied with her assessment, Yvette refined it yet again in her portfolio. She combined the content from the two previous ESRs, modified the tasks to request evidence from classroom activities, and refined her criteria for three items—1 point for a correct response and 1 point for evidence. This adjustment reflected her growing concern with differentiating subconcepts and capturing student understanding. If we trace her growth since the prior summer, we see that she moved from scoring and analyzing student work based on general impressions (or possibly extraneous factors) toward tools that targeted students' understandings of specific concepts. She was yearning for resources to support her continued learning.

Spring 2004. In the spring of 2004, Yvette taught another GEMS unit, *Hot Water, Warm Homes*. She was determined to devote more attention to interpretation of student understanding, but weaknesses in unit assessments continued to make demands on her time.

Yvette developed a pre-test consisting of a prestructured concept map that students completed by filling in selected vocabulary words. Yvette thought it would allow students to show how they understood the relationships between thermal energy, conduction, convection, temperature, and radiation. But in analyzing the student responses, Yvette found that 14 of 23 students scored high, and she worried that this pre-test task allowed for false positives: "A student who is good at matching . . . or good at looking at context clues would do well on this test. I don't think it assessed their understanding about these concepts directly" (7/04). She saw little value in closer analysis of the student work, so, just as she did in the fall of 2003, she supplemented the individual assessment with a group activity.

> After reviewing the pre-assessment for student understanding, I recognized that it did not tell me enough information about radiation and how it relates to thermal energy, infrared light and visible light. So during the first session of the unit, I asked students to white board "What connection does the increase in temperature relate to radiation, infrared light, visible light and thermal energy?" (Spring 2004 portfolio)

When she revised this pre-test later (for use the following year), she tossed her concept map approach and revised the whiteboard prompts as individual short-answer tasks.

Yvette was more satisfied with her juncture assessment, a short-answer essay given individually, and spent considerable time analyzing student responses. For each concept assessed, she charted the distribution of scores and then made additional notes on conceptual understanding—for example, students' assumptions that "the light cannot escape [the car]" or "no acknowledgment of equilibrium." Although her criteria were still point based, they had greater capacity to capture student thinking than her prior criteria because they were differentiated for four subconcepts. She also applied these criteria in new ways; guided by the portfolio forms, she transferred the scores to a matrix to identify the concepts students appeared to understand, and, when a response scored a zero, she revisited the response to alternative conceptions. Her interpretation of the student data led her to discover that one instructional activity had fostered

unexpected outcomes, and she revised the activity to "prevent the creation of alternative conceptions" (Spring 2004 portfolio). For the first time, her interpretation of a set of student work provided her with relevant and useful information that met the needs of her students, especially those struggling with the major concepts of the unit.

Yvette decided to use this juncture assessment as the posttest to help her track progress. At posttest, she found that fewer students still held alternative conceptions, though a sizable number of her students did not demonstrate understanding of two of the essential concepts. She was pleased that using similar tasks from juncture to post and consistent criteria helped her to evaluate how her students developed.

Summary. With the support of the Academy portfolio and facilitation, Yvette became more skilled in analyzing student work—developing criteria, charting scores and taking notes on alternative conceptions, and reflecting on patterns. She became increasingly aware that the soundness of her interpretations depended on strengthening a network of assessment practices, especially the clarity of her goals and the quality of her assessment tools. Yvette's efforts to develop her own assessments required persistence. What motivated her was her personal commitment to learn how to use assessment for learning as well as for grading.

> [Working on the portfolio] has given me an opportunity to look at how I teach and how I examine the student work—not just to assess them but to assess the lesson, and what's effective and what's not effective. And it's even helped me to look at what tools are the best tools to use. Which one is going to give me the information that I need at the time? Which one is more suitable for which purpose? . . . So it's given me the opportunity to look closely at assessment. Whereas before it was, "OK. I have a test. I need to give it." . . . It doesn't have to end in with a final grade—it can be something that helps me to structure my lessons. (YJ, 3/04)

Yvette was learning about myriad aspects of classroom assessment, but she was overly entangled in assessment development. If her tools had been of better quality and if the Academy had provided more resources on student understanding, Yvette could have devoted more attention to interpretation and to the decisions made from her analysis.

Joan Reddy: Using Research-Based Assessments as a Resource for Interpreting Student Work

Overview. The case of Joan Reddy illustrates how quality embedded assessments can support teachers' interpretation of student work, yet the insights teachers get from such assessments in one curriculum unit may be less useful when they are examining student work from other units. Joan taught eighth-grade science in a middle school situated in a small urban community bordered by farmland. Head of the science department, she embraced the opportunity to implement the FAST embedded assessment system in the fall of 2003, and used the assessments in thoughtful ways to monitor her students' progress with the concept of buoyancy. But spring 2004 was back to the status quo; like most Academy teachers, she implemented a published unit without high-quality assessments and faced the need to develop her own assessments and criteria for interpreting student work. Her efforts to extend the FAST model revealed its limited utility for curriculum units that target a broader array of standards, as well as some gaps in Joan's evolving understandings of assessment.

Fall 2003. Joan used the fall 2003 assessments (Figure 1) in careful and deliberate ways when interpreting student work. In her think-aloud below, we can hear how she teased out the misconception that object shape is a factor in flotation (without considering displaced volume).

> Student response: If there is a block of steel, and you put it in water, it sinks because it had more mass. If you put a hollow piece of steel of the same mass, and shaped like a banana, it would float because it was shaped different, so it could float. For example, a fish has a swim bladder. He can let air in and out, and that is for him to go up or down or sub-surface. Like this, in this diagram.
>
> Joan: So he is trying to show a solid piece of steel floating and a solid piece of steel sinking. I think here, he made a point of mass, didn't say anything about volume, talked about shape, but not really volume. It looks like this child is still at the mass level. He is trying to make a relationship between the shape and the mass. (10/28/03)

Joan considered the response in detail as she decided how to classify it on the "progress guide." In her portfolio, she scored all the papers in this manner and then entered her scores in a comprehensive matrix for each assessment throughout the unit.

This FAST strategy for tracking progress was afforded by the new system of comparable assessments and common progress guides.

Spring 2004. In the spring, working with a unit on mixtures and solutions without an embedded assessment system, Joan applied her understanding of the FAST model to the design of her own assessments and criteria. To interpret student work, Joan followed portfolio guidelines and developed criteria for low, medium, and high responses for several key assessments. Table 3 illustrates her approach. These criteria were more comprehensive in content than FAST progress guides, because Joan wanted to track students' progress on a broad range of dimensions aligned with state standards. Her capacity to do that with a holistic scale with multiple elements was limited, however, as the scale did not differentiate unit objectives (and included other proficiencies). Joan's view of student understanding was standards-based and additive, different from the qualitative stages in the FAST progress guides. She was interested in knowing whether students included all the knowledge "pieces" she regarded as critical: ". . . at medium they will say all of them, and low they will say some of them, but in high they should see most of the pieces—which means they should say 'a solution could be separated by filtration,' 'by chemical means,' [and] 'by chemical absorption'" (5/14/04).

Table 3
Example of Joan's Criteria for Spring 2004 Assessments

High	Medium	Low
Have 20 or more multiple-choice and fill-in-the-blank questions correct	Have 11-20 multiple-choice and fill-in-the-blank questions correct	Have 0-10 multiple-choice and fill-in-the-blank questions correct
In short answers, specify:	In short answers, specify:	In short answers, specify:
Have 5-7 questions correct	Have 3-4 questions correct	Have 1-2 questions correct
Mixtures are made up of two or more substances	Mixtures are made up of two or more substances or components or compounds	Mixtures are made up of two or more substances or elements
Mixtures can be separated by physical changes	Mixtures can be separated by physical changes	Mixtures can be separated by physical changes or chemical changes
Magnets attract iron	Magnets attract iron	Magnets attract iron or filtration or some other method
A mixture of pebbles, salt and pepper could be separated by straining, filtering and evaporation	Some of the concepts are placed correctly on the concept map	Few of the concepts are placed correctly on the concept map

(*table continues*)

Table 3 (continued)

High	Medium	Low
Most of the concepts are placed correctly on the concept map Different types of mixtures are solutions, suspensions, and colloids		

Joan was not satisfied with her criteria, but was unsure why not. She did not yet recognize that her standards-based criteria differed from the FAST developmental model, or that a holistic scale cannot provide evidence of students' understanding of any concept or proficiency with any skill. Joan appeared to be caught between competing views of assessment targets—standards coverage vs. conceptual understanding—without a firm understanding of assessment principles underlying design choices. Neither her experience with FAST nor Academy resources prepared her for a unit assessing students' progress with a broad range of concepts and skills.

As suggested in the portfolio, Joan entered her scores into a matrix, recording student scores of high, medium, low for all assessments to track student progress. The form of this matrix was similar to her matrices for FAST scores, but could not serve the same purpose since tasks and criteria were different across the assessments. Joan was aware that there was some discrepancy, but, influenced by FAST, she wanted to determine what she could learn from this matrix about class performance within and across assessments.

> Joan: [A]s a teacher, I want to see, like within this unit, if I'm just scoring them on high, medium, or low, is there a pattern? Are there students going up and down? Are they going from low to high, medium to high, high to high? Are there some who were high before but going down? I want to see that pattern and then I kind of want to go back and talk to the students and see if there are any misconceptions, or see if something happened where I need to go back and talk about those, either to whole class or to those specific students. So I felt this was a visual way for me to do it, just to see how that student is doing. By looking at this I can tell for each of the lessons that we have been working on the range. . . This student right here was low, low, low, medium, and then on separating a solution he was high, then medium, high, high, medium, high. So he's kind of staying in that range I'm going to watch. Those are the kind of students I'm going to watch closely and also the students who are jumping around. (5/14/04)

Her uncertainty about the matrix was evident in her comment that the scores could not provide her evidence of "misconceptions" and thus her decision to supplement with informal methods of gathering information ("go back and talk to the students"). Her strategy of using the chart to target certain students was her way to strengthen equitable attention to all students, but the meaning of her comparisons is

uncertain from a measurement perspective, and she seemed to understand that. Academy activities had not addressed measurement of progress in systematic ways.

Summary. With the support of the reflective portfolio, Joan used FAST assessments thoughtfully and systematically in the fall of 2003, but her extension of FAST progress guides was only partially appropriate to her spring 2004 unit with its greater range of learning goals. This pattern is no surprise given the complexity of the psychometric principles underlying the FAST assessments (which were unlikely to be transparent), and the challenge of aligning assessments with myriad state standards. The Academy did not anticipate the need to guide teachers in the relevance and usefulness of FAST to other units. As a result, Joan ended the first year with room to grow in her understanding of sound ways to build criteria that capture conceptual understanding and to analyze student progress. Yet she developed a deep commitment to clear learning goals and devised a coordinated set of formal and informal strategies for gathering information.

Carrie Green: Reliance on Informal Assessment and Frustration With Scoring Student Work

Overview. Carrie taught six periods of the same physical science course to eighth graders. Committed to inquiry teaching, Carrie was engaged with her students every moment, prompting students to reflect on critical concepts during whole class discussions and lab investigations. Unlike Yvette and Joan, Carrie scored student work infrequently, relying on informal assessment and look-throughs of student work. She found the paper load daunting, and she lacked confidence in her assessments, her criteria, and her methods of charting responses. Carrie completed some sections of her portfolios and, while she constructed reasonable inferences about patterns of student understanding, she tended to focus on what each class "got" or "didn't get" rather than systematic patterns of understanding based on scored student work. In the first year of the Academy, Carrie acknowledged that, "I've never looked at student work this closely before." She became more interested in criteria to capture levels of understanding and eventually developed criteria that were differentiated and aligned with her learning goals. Still she was frustrated with the work this required and would have preferred to implement units that provided her quality assessment tasks and criteria and feasible charting methods.

Fall 2003. Carrie chose a curriculum unit in the fall of 2003 that provided limited assessment resources, and she had to develop her own criteria for the assessments. Carrie and her team partner chose a lab as their pre-assessment to provide information about students' understandings of several concepts, including the distinction between observation and conclusion, skills with tools and measurement, and conceptions of matter. But when Carrie had her student work in front of her, she realized that the evidence was inadequate to make inferences about all of these assessment targets. She was stumped and delayed further analysis until she could meet with her partner a month later. Together they pared down the concepts to strengthen alignment and focused only on students' understandings of the distinction between observations and conclusions. That decided, they sorted the responses into "high" and "low" piles and drafted criteria as shown in Table 4.

Table 4

First Draft of Criteria for Carrie's Preassessment

High	Low
Specific	Observations were surface level
Used cause and effect	Didn't go beyond the obvious
More than one observation	One observation
Used data	Observations had nothing to do with
Observed what happened	activity

The dichotomous criteria in their first draft were more like correct/incorrect point schemes giving students credit for inclusion of certain answers. Their facilitator reminded them to sort papers into high, medium, and low piles to capture a range, and the team's second draft is shown in Table 5. This version contained an additional level, but the underlying dimensions had essentially the same content, and it introduced the

extraneous criteria of legibility and writing quality. In a final version (not shown), they removed the extraneous criteria, and their final product was a set of criteria that were fairly specified and reasonably aligned with the revised targeted concepts, but not likely to capture students' conceptions of evidence and inference.

Table 5

Second Draft of Criteria for Carrie's Preassessment

High	Medium	Low
Accurate example (e.g., temp. really did go down)	Accurate No data 1 or 2 valid observations	Inaccurate observations No or wrong data 1 observation/weak
Use data when appropriate	Object or environment	Unspecific
Exhaustive list of cause and effect	Not all items included Sloppy but legible	Unrelated WHAT [i.e., no idea what student means]
Specific object/environment		
Include all items in activity		
Clarity of writing		
Just facts		

Carrie's interest in student conceptions emerged only gradually over the first year as she began to reflect on the Academy notion of "developmental appropriateness." In a

February 2004 interview, Carrie noticed she had considered only correct answers in her original assessment plan in Section I: "We wrote 'expected student response,' what you're expecting the students to say. The one thing that we didn't do was come up with what students might say if they didn't get it" (2/04). She recognized that the teacher's guide was no help, contributing to a right/wrong approach in two ways. First, the guide suggested point schemes for correcting student work, "so it was really easy for me: 'did they talk about physical properties?' 'did they talk about solubility?' 'did they talk about density?' If they did, then they hit all the points according to this rubric" (2/04). Second, the guide repeatedly listed "misconceptions" as "incorrect." For example: "*Students incorrectly think that mass alone is the determining factor. . . ; students incorrectly think that because mass and volume are both used to measure 'quantities.'*"

Carrie began to wrestle with two new Academy ideas about criteria in February. One was the value of thinking through high, medium, and low responses when constructing criteria. The second was the importance of considering what students "know" as well as what students "say" or "do."

> We talked about writing out what the concepts were and what the expected student responses for high medium and low would be for this concept. . . . That proved to be really really helpful, because I was forced to think about, what I want my kids to *know,* every kid, in order to move on. . . . It made a huge difference on my perspective. . . . [our emphasis]

But she was still uncertain of the instructional value of a three-level rubric with qualitative levels of understanding.

> We did try really hard to talk about what they may say and sort of what levels would show more understanding than others. . . . but I didn't go into any kind of detail on what [students] would need to do in order to move up [from low to medium, or medium to high] . . . I didn't use it and [don't know] the benefit that it could have been.

Although Carrie questioned the value of rubrics and resisted scoring, her interest in student conceptions was clearly growing. For example, for one lab assessment, she reported that students were making progress in the quality of their observations, but were not yet constructing a deep understanding of density.

> **Got it:** -They were able to separate the substances via methods given to them to determine what the mixture was made of. -As far as general observations, they

had moved away from their incomplete observations in the first lesson to much more thorough observations that consisted of multiple parts.

Didn't get it: -Soluble/insoluble; most did sink/float, but many could not say whether it was more/less dense than water. - Patterns [nothing written]

Evidence [nothing written]

Although she framed her inferences as "got it" versus "didn't get it," and although she cited no evidence in the student work, her inferences were thoughtful and complex. In an interview about this student work, she commented on student conceptions.

> The trend that I noticed, was . . . that the hardest concept which was density . . . [T]hose that wrote about density could actually say "less dense than water" or "more dense than water," did all three [describe, solubility, density]. It was almost a building process: They could all describe; some could describe and say it was soluble in water or insoluble in water; and if they could do those two, they also seemed able to move to density. (2/04)

With the support of the portfolio, Carrie was increasingly intrigued by student conceptions, yet she was hampered by weak assessment tools and her limited capacity to develop adequate criteria and methods of analysis.

Spring 2004. In the spring, Carrie's criteria were more differentiated, reflecting her diligent work in sorting out the concepts she was targeting as her learning goals. Table 6 contains her rubric. Yet the criteria for students' understanding of forces were not yet capturing developmental levels of understanding as Carrie was keenly aware. In discussing the "high" and "low" levels on this rubric, Carrie rued that she was still scoring for right versus wrong, or, as she put it, "this whole yin yang thing . . . I mean, that's the easiest way to fall into writing a rubric." Carrie had made advances in her approach to specifying learning goals for core concepts and aligning her criteria, yet she could not figure out how to construct criteria that captured student understanding and progress. Without resources on conceptual development provided by either the Academy or her instructional materials, Carrie lacked the knowledge necessary to develop sound criteria.

Table 6

Carrie's Rubric for Forces

High	Low
Forces can be observed and measured.	
Accurately identifying forces Pick out force with given example Choose correct tool to measure force Use correct units Explain variations in capacity	Forces are mentioned along with other phenomena Unable to identify or lucky guess Cannot identify or use appropriate tool Missing units Does not understand need for multiple units
Forces can be a push or pull.	
Identify forces and explain why they chose it Explanation includes push or pull Ex: ___ is a force because it is pushing.	Many not be able to identify the force Can identify but not say why gives reason but reason does not include "push" or "pull"
Forces occur in pairs that act in opposite directions.	
Identify two forces acting together Explain why/how the two forces work together (some understanding of forces opposing one another) Tell the direction of forces	Only pick one force Cannot explain how they work together Cannot show direction of force

(*table continues*)

Table 6 (continued)

High	Low
Forces have varying strengths.	
Can predict and explain the diff. degrees of force using > < statements	Cannot correctly predict varying forces—what will happen
If forces are balanced, objects remain as they are.	
[blank]	[blank]

Summary. The Academy and the portfolio activities had a deep impact on Carrie's awareness of the information contained in student work and her interest in capturing student conceptions. But working without quality tools for interpreting student work, Carrie found developing criteria frustrating, and she preferred her methods of informal assessment. Carrie needed higher quality assessment tools and resources to guide her in interpretation of student understanding.

Discussion

Researchers have argued that effective professional development requires long-term engagement and commitment (Birman, Desimone, Garet, Porter, & Yoon, 2001; Hawley & Valli, 1999; Loucks-Horsley, Love, Stiles, Mundry, & Hewson, 2003), yet we understand little about how teachers learn over time in the context of a program like the Academy's. This article contributes insight into teacher learning by portraying the growth of three teachers who worked earnestly for a year to strengthen their expertise with classroom assessment, supported by the Academy portfolio and their colleagues and facilitators. We have shown how the portfolio repeatedly challenged teachers to wrestle with old and new ideas about interpretation of student work as they refined approaches to constructing criteria and analyzing student responses. All three teachers gradually recognized that strengthening their interpretation of student work required integrating new assessment concepts such as developmental appropriateness, as well as improving related components of assessment, especially the specificity of learning goals and the quality of assessment tasks. Our framework for classroom assessment expertise was an invaluable resource for this study. It represents expertise as a relationship between understandings of assessment concepts and facility with assessment practices; its complexities enabled us to capture complexities in the pathways that teachers took in their efforts to improve.

The teachers' growth was impressive, and the Academy portfolio and institutes played a critical role. Teachers were exposed to important assessment concepts and were asked to integrate these concepts in their practice through cycles of assessment planning, implementation, and evaluation. These repeated opportunities for learning, application, and portfolio reflection contributed to teachers' deepening interest in student work as evidence of student understanding and a resource for instructional improvement.

Yet Yvette, Joan, and Carrie grew only so far, and they were not satisfied with their methods of interpreting student work at the end of the first year of the Academy. A common pattern was their emerging appreciation of an important assessment concept without adequate capacity to implement it. For example, Yvette's and Carrie's interest in developmental appropriateness outstripped their use of that notion when refining criteria, because neither their materials nor the Academy provided them a framework for capturing conceptual change. Joan's interest in charting student progress was ahead of her understanding of the need for comparable assessments, because the FAST unit had not made that principle transparent, and the Academy did not address that measurement principle in the first year. These patterns have implications for improvement of both the Academy program and the assessments provided with instructional materials.

For its part, the Academy program was a generalized approach to assessment support, and as such, a valuable resource for our three case teachers. The Academy was a context for developing general assessment expertise that helped teachers working with a variety of units adapt and revise assessments for their purposes and contexts. But by the end of their first year in the Academy, Yvette, Joan, and Carrie were ready for more specialized support. They were seeking models of assessment practices directly pertinent to their units, insights about how students develop unit concepts, and an understanding of more advanced measurement principles such as comparability of measures for tracking progress. The implication is that long-term professional programs may need to differentiate their resources and strategies as teachers advance along different pathways to growth and improvement.

The assessments provided with the teachers' instructional materials played a different role in teachers' growth. When assessment tasks and criteria were missing or of poor quality, teachers were frustrated by the time required to improve them. They longed for higher quality assessments grounded in a developmental framework that would enable them to track student progress, and interpret everyday observations as well as more formal assessments. When tasks and criteria were of good quality, as they were for Joan, they were implemented thoughtfully, but the assessment system and its underlying principles were not readily applicable to other units or well understood. The synthesis here is that quality assessments and criteria do not ensure quality implementation and use. But quality assessment systems are crucial resources for busy teachers who cannot develop assessment systems for every unit they implement,

provided the design principles are shared in ways that build teachers' general assessment expertise.

Closing Remarks

Our focus on interpretation of student work proved to be a prudent choice for our research on teacher learning. Interpretation of student understanding has central importance in quality assessment practice, as supported by studies reporting that weaknesses in teachers' interpretation of evidence have consequences for equitable teaching and assessment (e.g., Morgan & Watson, 2002; Schafer, Swanson, Bene, & Newberry, 2001; Watson, 2000). Our study indicates that programs like the Academy as well as quality embedded assessments can help teachers learn to interpret student work in sound and equitable ways.

Of course, research is also needed on the development of other aspects of assessment expertise, including grading and informal assessment, as well as the assessment systems that teachers devise for a unit. Researchers are identifying important differences in teachers' approaches to gathering and coordinating information, including tendencies to rely on daily impressions versus more systematically collected evidence, and on formative versus summative assessment (Brown, 2004; Gipps, 1999). Though we documented some ways that our teachers were coordinating informal and formal assessment, it will be productive to make teachers' assessment systems a more intentional focus of analysis. Research is also needed on the ways that discipline and grade level shape classroom assessment and its dilemmas and challenges. In this report, our choice of three middle school teachers sidestepped comparisons of teacher learning at various grade levels, and we plan future studies that address elementary and secondary levels as well.

References

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. Oxford, UK: Oxford University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement and Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in educational assessment of students. Washington, DC: Author.

Aschbacher, P. (1999). Helping educators to develop and use alternative assessments: Barriers and facilitators. Educational Policy, *8,* 202–223.

Atkin, J. M., & Coffey, J. E. (Eds.). (2003). Everyday assessment in the science classroom. Arlington, VA: National Science Teachers Association Press.

Bell, B., & Cowie, B. (2001). Formative assessment and science education. Dordrecht: Kluwer Academic.

Birman, B. F., Desimone, L., Garet, M. S., Porter, A. C., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, *38,* 915–945.

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

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. Teaching and Teacher Education, *13,* 259–278.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice, *22*(4), 5–12.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. Teachers College Record, *106,* 429–458.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. Assessment in Education, *11,* 301–318.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, *58,* 438–481.

Cuff, K., Willard, C., & Carmichael, I. (2001). Plate tectonics. Berkeley: University of California, Lawrence Hall of Science.

Davis, E. (2003). Knowledge integration in science teaching: Analysing teachers' knowledge development. Research in Science Education, *34*(1), 21–53.

Falk, B., & Ort, S. (1998). Sitting down to score: Teacher learning through assessment. Phi Delta Kappan, *80,* 59–64.

Gearhart, M., & Saxe, G. B. (2004). When teachers know what students know: Integrating assessment in elementary mathematics. Theory Into Practice, *43,* 304–313.

Gipps, C. (1999). Socio-cultural aspects of assessment. Review of Research in Education, *24,* 355–392.

Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. Educational Assessment, *6,* 257–290.

Goldsmith, L., & Schifter, D. (1997). Understanding teachers in transition: Characteristics of a model for the development of mathematics teaching. In E. Fennema & B. S. Nelson (Eds.), Mathematics teachers in transition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development. In L. Darling-Hammond & G. Sykes (Eds.), Teaching as the learning profession: Handbook for policy and practice (pp. 127–150). San Francisco: Jossey-Bass.

Herman, J. A. (2005, September). Using assessment to improve school and classroom learning: Critical ingredients. Presentation at the annual conference of the Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Hesse-Biber, D., Dupuis, P., & Kinder, T. S. (1991). HyperRESEARCH: A computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis. Qualitative Sociology, *14,* 289–306.

Laguarda, K. G., & Anderson, L. M. (1998). Partnerships for standards-based professional development: Final report of the evaluation. Washington, DC: Policy Studies Associates, Inc.

Loucks-Horsley, S., Love, N., Stiles, K. E., Mundry, S., & Hewson, P. W. (2003). Designing professional development for teachers of science and mathematics. Thousand Oaks, CA: Sage.

Morgan, C., & Watson, A. (2002). The interpretative nature of teachers' assessment of students' mathematics: Issues for equity. Journal for Research in Mathematics Education, *33,* 78–110.

National Research Council. (1996). National science education standards. National Committee on Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2001a). Classroom assessment and the National Science Education Standards. Committee on Classroom Assessment and the National Science Education Standards. J. M. Atkin, P. Black, & J. Coffey (Eds.). Center for Education, Division of Behavior and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001b). Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavior and Social Sciences and Education. Washington, DC: National Academy Press.

Natriello, G. (1987). The impact of evaluation processes on students. Educational Psychologist, *22,* 155–175.

Pottenger, F., & Young, D. (1992). The local environment: FAST 1 foundational approaches in science teaching. Honolulu: University of Hawaii at Manoa, Curriculum Research and Development Group.

Sato, M. (2003). Working with teachers in assessment-related professional development. In J. M. Atkin & J. E. Coffey (Eds.), Everyday assessment in the science classroom (pp. 109–120). Arlington, VA: National Science Teachers Association Press.

Saxe, G. B., Gearhart, M., Franke, M. L., Howard, S., & Crockett, M. (1999). Teachers' shifting assessment practices in the context of educational reform in mathematics. Teaching and Teacher Education, *15,* 85–105.

Schafer, W. D., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. Applied Measurement in Education, *14*, 151–170.

Science and Technology Concepts for Middle Schools. (2000). Properties of matter. Burlington, NC: Carolina Biological Supply Company.

Shavelson, R., Stanford Education Assessment Laboratory (SEAL), & Curriculum Research & Development Group (CRDG). (2005). Embedding assessments in the FAST curriculum: The romance between curriculum and assessment. Final Report. Palo Alto: Stanford University.

Sheingold, K., Heller, J. I., & Paulukonis, S. T. (1995). Actively seeking evidence: Teacher change through assessment development (Rep. No. MS–94-04). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), Handbook of research on teaching, (4th ed., pp. 1066–1101). Washington, DC: American Educational Research Association.

Shulha, L. M. (1999). Understanding novice teachers' thinking about assessment. Alberta Journal of Educational Research, *45,* 288–303.

Stiggins, R. J. (2005). Student-involved assessment FOR learning (4th ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Taylor, C. S., (1997). Using portfolios to teach teachers about assessment: How to survive. Educational Assessment, *4*, 123–147.

Taylor, C. S., & Nolen, S. B. (1996). What does the psychometrician's classroom look like? Reframing assessment concepts in the context of learning. Educational Policy Analysis Archives, *4*(17). Retrieved 8 April 2006 from http://epaa.asu.edu/eepa/v4n17.html

Watson, A. (2000). Mathematics teachers acting as informal assessors: Practices, problems and recommendations. Educational Studies in Mathematics, *41,* 69–91.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. Assessment in Education, *11,* 49–65.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied Measurement in Education, *13,* 181–208.

Wilson, S. (2004). Student assessment as an opportunity to learn in and from one's teaching practice. In M. Wilson (Ed.), Towards coherence between classroom assessment and accountability (National Society for the Study of Education Yearbook, Vol. 103, Part 2, pp. 264–271). Chicago: University of Chicago Press.

Author Note

Diana Bernbaum, Maryl Gearhart, Jennifer Pfotenhauer, and Cheryl Schwab, School of Education, University of California, Berkeley; Shaunna Clark, Joan Herman, Sam Nagashima, Ellen Osmundson, and Terry Vendlinski, Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Jennifer Pfotenhauer is now at Malcolm X Elementary School, Berkeley, California.

We are grateful to the participating teachers who contributed their time and good will to our study, and to the professional development team for their collaboration and input: Kathy Diranna (Co-Director), Craig Strang (Co-Director), Lynn Barakos, Diane Carnahan, Karen Cerwin, and Jo Topps. We also thank the volume editors, Alison Bailey and Margaret Heritage, and an anonymous reviewer for their careful review and thoughtful feedback on previous versions of this report.

Correspondence may be sent to Maryl Gearhart, School of Education, University of California, Berkeley, California 94720-1670. E-mail: gearhart@berkeley.edu

Information regarding the Academy portfolio is available from Kathy DiRanna at kdirann@wested.org