APPLIED MEASUREMENT IN EDUCATION, *13*(2), 181–208


# From Principles to Practice: An Embedded Assessment System

Mark Wilson and Kathryn Sloane
*Graduate School of Education*
*University of California, Berkeley*

In this article, we describe the principles that guided the creation and implementation of a system of embedded assessments—the so-called BEAR (Berkeley Evaluation and Assessment Research) Assessment System. The assessment system was developed in the context of a specific curriculum in issues-oriented science for the middle grades but is designed generically to address the implementation of those principles. The assessment system builds on methodological advances in alternative assessment techniques and attempts to address salient issues in the integration of alternative assessment into the classroom teaching and learning context. The 4 principles are described, and we discuss how the application of these principles generates the component parts of the system and determines how the component parts work together. The use of teacher moderation to integrate the parts of the system in school and classroom is also discussed.

In recent years, "alternative assessment" has been a major topic of interest, debate, and experimentation in the nationwide efforts at educational reform. Initial hopes that alternative, authentic, or performance assessments of student achievement would drive (or at least facilitate) changes in what and how students are taught have been tempered by the realities of implementation. Efforts to introduce alternative assessments into large-scale, high-stakes state and district testing programs have met with mixed results due to high costs, logistical barriers, and political ramifications (e.g., Gipps, 1995; Rothman, 1995). For example, the demise of the California Learning Assessment System was due principally to the complications, technical, political, and financial, of using performance assessments for large-scale assessment. Efforts to introduce alternative assessments

Requests for reprints should be sent to Mark Wilson, Graduate School of Education, University of California, Berkeley, CA 94720. E-mail: mrwilson@socrates.berkeley.edu

into ongoing classroom practices have been less publicized but have also met with problems relating to costs (primarily in terms of time) and to teachers' level of preparation and acceptance (e.g., Chittenden, 1991; McCallum, Gipps, McAlister, & Brown, 1995; Shepard, 1995). The rationale for developing and using alternative assessment remains quite compelling, however. Alternative assessments, compared to traditional tests, offer the potential for greater "ecological validity" and relevance, assessment of a wider range of skills and knowledge, and adaptability to a variety of response modes (e.g., Baron, 1991; Gardner, 1992; Malcom, 1991; Wiggins, 1989, 1993).

In this article, we describe the principles behind the development of a generic embedded assessment system. We describe the component parts of the system, how they relate to the principles, and how they work together. The system, called the BEAR Assessment System because it was developed at the Berkeley Evaluation and Assessment Research Center at the University of California, Berkeley, builds on methodological advances in alternative assessment techniques. It attempts to address salient issues in the integration of alternative assessment into the classroom teaching and learning context. The BEAR Assessment System offers one model of how assessment can be incorporated into the classroom teaching and learning process.

The BEAR Assessment System is a comprehensive, integrated system for assessing, interpreting, and monitoring student performance. It provides a set of tools for teachers to use to do the following:

- Assess student performance on central concepts and skills in the curriculum.
- Set standards of student performance.
- Track student progress over the year on the central concepts.
- Provide feedback (to themselves, students, administrators, parents, or other audiences) on student progress and on the effectiveness of the instructional materials and the classroom instruction.

The approach used is that of *embedded assessment.* By using the term *embedded* we mean that opportunities to assess student progress and performance are integrated into the instructional materials and are virtually indistinguishable from the day-to-day classroom activities.

The BEAR Assessment System was first implemented for a specific middle school science curriculum: the Science Education for Public Understanding Project (SEPUP) at the Lawrence Hall of Science. SEPUP staff developed a year-long, issues-oriented science course for the middle school and junior high grades entitled: Issues, Evidence, and You. (IEY; SEPUP, 1995). This course focuses on environmentally and socially contextualized science content. Societal decision making is a central focus of IEY and, in many ways, distinguishes this course from other middle school science courses. The goal of issue-oriented sci-

ence is the development of an understanding of the science content and scientific problem-solving approaches related to social issues without promoting an advocacy position. The concepts and skills needed to understand the process of societal decision-making form the basis of the SEPUP curriculum. As part of the course, students are regularly required to recognize scientific evidence and weigh it against other community concerns with the goal of making informed choices about relevant contemporary issues or problems. IEY will be used to supply examples of the major parts of the BEAR Assessment System.

In designing the BEAR Assessment System we were guided by four principles. The principles represented the standards or ideals that should, we believe, be reflected in a technically sound, curriculum-embedded, classroom-based system of student assessment. The roots of these principles can be traced (in part) to recent work in measurement theory and the recent research literature on alternative assessment practices. However, the combination of these principles, and the relations among them, represent a new approach to classroom assessment. The principles are labeled (a) developmental perspective, (b) match between instruction and assessment, (c) teacher management and responsibility, and (d) quality evidence. In the next four sections, we discuss each of these principles, relate it to the BEAR Assessment System, and give examples of how it operates in the IEY example. Following that we discuss how the four principles are integrated in practice by using a part of the assessment system called *moderation*. Moderation is a process in which teachers discuss student work and the scores they have given that work, making sure that the scores are being interpreted in the same way by all teachers in the moderation group. We also briefly discuss an empirical evaluation that has been carried out. Finally, we suggest some interesting paths for future development and research.

## DEVELOPMENTAL PERSPECTIVE

The first principle is that an assessment system should be based on a developmental perspective of student learning. Assessing the development of students' understanding of particular concepts and skills (as opposed to current status only) requires a model of how student learning develops over a certain period of (instructional) time. A developmental perspective helps us move away from one-shot testing situations, and away from cross-sectional approaches to defining student performance—toward an approach that focuses on the process of learning and on an individual's progress through that process. Although it would be possible to consider the accretion of memorized facts as development, we have found that attention to development tends to move people beyond mere memorization. Clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material, are necessary elements in examining the construct validity of the

scores obtained from the assessment system. However, they are also necessary to ensure that the assessment system is useful for instructional purposes (see principle discussed next)—that is, for instructional validity.

Our strategy to address this issue is to develop a set of "progress variables" (Masters, Adams, & Wilson, 1990; Wilson, 1994b) that mediate between the level of detail that is present in the content of specific curricula and the necessarily more vague contents of state standards and curriculum framework documents. Such progress variables define the intended content of a specific curriculum up to a level of detail that would allow, say, biweekly tracking of student progress through the curriculum. Given the burdens of the typical teacher, this set might be composed of about 4 to 5 progress variables, these variables would define the most important student growth goals of the curriculum. Every instructional unit would be seen as contributing in some way to student progress on at least one of these variables—every assessment would be closely aligned with one (or more) of the variables. This alignment allows the creation of a calibrated scale to map the growth of students, so that teachers can track the progress of individual students and groups of students as they undergo instruction. This idea of a "crosswalk between standards and assessments" had also been suggested by Baker (as cited in Land, 1997, p. 6). These variables also create a conceptual basis for relating the curriculum to standards documents, to other curricula, and to assessments that are not specifically related to that curricula (discussed next; Wilson, in press).

In this approach, the idea of a progress variable is focused on the concept of progression or growth. Learning is conceptualized not simply as a matter of acquiring quantitatively more knowledge and skills, but as progress toward higher levels of competence as new knowledge is linked to existing knowledge, and deeper understandings are developed from and take the place of earlier understandings. The concepts of ordered levels of understanding and direction are fundamental: In any given area it is assumed that learning can be described and mapped as progress in the direction of qualitatively richer knowledge, higher order skills, and deeper understandings. Variables are derived in part from professional opinion about what constitutes higher and lower levels of performance or competence but are also informed by empirical research into how students respond or perform in practice. They provide qualitatively interpreted frames of reference for particular areas of learning, and permit students' levels of achievement to be interpreted in terms of the kinds of knowledge, skills, and understandings typically associated with those levels. They also permit individual and group achievements to be interpreted with respect to the achievements of other learners.

## The IEY Example

An example of such a set of progress variables is taken from the IEY middle school science curriculum. Following the developmental perspective principle we, along

with the SEPUP curriculum developers, devised a framework of progress variables that embody the learning that students are expected to experience in the IEY year. The five IEY variables are as follows:

1. *Understanding concepts* (UC): understanding scientific concepts (such as properties and interactions of materials, energy, or thresholds) to apply the relevant scientific concepts to the solution of problems. This variable is the IEY version of the traditional science content, although this content is not just "factoids."
2. *Designing and conducting investigations* (DCI): designing a scientific experiment, carrying through a complete scientific investigation, performing laboratory procedures to collect data, recording and organizing data, and analyzing and interpreting results of an experiment. This variable is the IEY version of the traditional science process.
3. *Evidence and tradeoffs* (ET): identifying objective scientific evidence as well as evaluating the advantages and disadvantages of different possible solutions to a problem based on the available evidence. This variable, and the two following are relatively new.
4. *Communicating scientific information* (CM): organizing and presenting results in a way that is free of technical errors and effectively communicates with the chosen audience.
5. *Group interaction* (GI): developing skills in working with teammates to complete a task (such as a lab experiment) and in sharing the work of the activity.

The first three variables—UC, DCI, and ET—are primary variables and are assessed most frequently. The traditional content of science tests has not been abandoned in this framework, traditional science content comes under the progress variable UC. Thus, teachers using this system do not lose anything compared to what they would get from a traditional approach, but they can gain. Students' performance on CM can be assessed in conjunction with almost any activity or assessment, depending on the teacher's interest in monitoring student progress on this variable. Opportunities in the course have been indicated for assessing students' skills in this area. The final variable, GI, is based on the SEPUP (1995) 4–2–1 model of instruction and can also be assessed throughout the year. We have developed a scoring guide for GI but have not as yet carried this variable through to complete development (this is the focus of a current research project).

Each of the five IEY variables is defined by a number of aspects that are called elements. The elements define how each variable is operationalized in the course. For example, DCI includes four elements: (a) designing investigations, (b) selecting and recording procedures, (c) organizing data, and (d) analyzing and interpreting data. Students can be assessed on one or several elements at a particular time.

The progress variables are the core of the generic assessment system that we have developed to help teachers make the best use of the information contained

TABLE 1
Principles for Assessment System and Their Relation to the Parts of the BEAR Assessment System

| | | *Implementation in IEY* | |
|---|---|---|---|
| *Principle* | *Parts of BEAR Assessment System* | *Component Parts* | *Integrative Part* |
| Developmental assessment | Progress variables | 5 IEY Variables UC, DCI, ET, CM, GI | Moderation |
| Match between instruction and assessment | Assessment tasks, linked to progress variables, different types for different purposes | Embedded tasks, link tests | Moderation |
| Teacher management and responsibility | Teachers judge student's work, teachers use results to plan instruction | Scoring guides, exemplars, assessment blueprint | Moderation |
| Quality evidence | Progress maps, reliability, SEM, and so forth | IEY maps for each variable | Moderation |

*Note.* IEY = Issues, Evidence, and You; UC = understanding concepts; DCI = designing and conducting investigations; ET = evidence and tradeoffs; CM = communicating scientific information; GI = group interaction; SEM = standard error of measurement.

in assessments. Table 1 lays out the different parts of the BEAR Assessment System, and relates them to the principle of developmental perspective and the other three principles described next. Note the prominent place that the progress variables occupy in Table 1, encapsulating as they do, the intent of the developmental perspective.

## MATCH BETWEEN INSTRUCTION AND ASSESSMENT

The need to integrate assessment into the curriculum and instruction process (i.e., the classroom context) is often emphasized in discussions of current assessment practices. A major part of the initial motivation for alternative assessment was to create a better match between desired instructional goals and actual assessment practices (e.g., Brown, Campione, Webber, & McGilly, 1992; Glaser, 1987; Resnick & Resnick, 1992). In other words, educational reformers have called for assessment techniques that better reflect (a) the problem solving and higher order thinking goals of the new curricula and (b) new instructional techniques. Currently, there is also the now standard call for assessment to be part of the teaching and learning process, that is, as a learning tool in and of itself. If assessment is also a learning event, then it does not take unnecessary time away from instruction, and the number of assessment tasks can be increased to improve the generalizability of the results (Linn & Baker, 1996). For assessment to become fully and meaningfully integrated into the teaching and learning process, however, the assessment must be

linked to a specific curriculum. That is, it must be curriculum dependent, not curriculum independent as must be the case in many high-stakes testing situations (Wolf & Reardon, 1996).

The second principle then, is that there must be a match between what is taught and what is assessed. This principle represents, of course, a basic tenet of content validity (American Psychological Association, 1985): that the items on a test are sampled appropriately from a domain that is defined by the content and the level of cognitive processing expected in a given body of instruction. Traditional testing practices—in high-stakes or standardized tests as well as in teacher-made tests—have long been criticized for oversampling items that assess only basic levels of knowledge of content topics and ignore more complex levels of understanding. The rationale for the development of authentic, alternative, or performance assessment techniques is based, at its heart, on the need for a better match between important learning objectives (e.g., problem solving) and the methods by which student performance on these objectives is assessed. As more attention is directed toward changing curricular materials and instructional methods, to reflect constructivist theories of learning or to reflect higher order learning objectives, the mismatch between curriculum, instruction, and assessment can become even more pronounced.

Concerns about the match between curriculum, instruction, and assessment have been discussed from both the curriculum development and assessment perspectives. From the curriculum perspective, efforts to emphasize new approaches to teaching and learning are inhibited by the form and content of accountability tests. Reports abound of teachers interrupting their use of their regular curricular materials to teach the material that students will encounter on the district or statewide tests. From an assessment perspective, advocates of assessment-driven reform hope to take advantage of the tendency to teach to the test by aligning high stakes testing procedures to the goals of curricular reform. As Resnick and Resnick (1992) argued in the following:

> Assessments must be designed so that when teachers do the natural thing—that is, prepare their students to perform well—they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform. (p. 59)

The match between the instruction and assessment in the BEAR Assessment System is established and maintained through two major parts of the system: the progress variables described earlier, and the assessment tasks, described next. In the previous section, the main motivation for the progress variables was that they serve as a framework for the assessments. However, the second principle makes clear that the framework for the assessments and the framework for the curriculum and instruction must be one and the same. This is not to imply that the needs of assessment must drive the curriculum, but rather that the two, assessment and in-

struction, must be in step—they drive one another. Using progress variables to structure both instruction and assessment is one way to make sure that the two are in alignment, at least at the planning level. To make this alignment concrete, however, the match must also exist at the level of classroom interaction and that is where the nature of the assessment tasks becomes so crucial.

Assessment tasks need to reflect the range and styles of the instructional practices in the curriculum. They must have a place in the "rhythm" of the instruction, occurring at places where it makes instructional sense to include them. This is usually at points where teachers need to see how much progress his or her students have made on a specific topic (for an elaboration of kinds of occasions like this, see Minstrell, 1998). One good way to achieve this is to develop both the instructional materials and the assessment tasks at the same time—adapting good instructional sequences to produce assessable responses and developing assessments into full-blown instructional events. Doing so will bring the richness and vibrancy of curriculum development into assessment and will also bring the discipline and hard-headedness of assessment into the design of instruction. It is much less satisfactory to try and do this in a post hoc way, putting the assessments in at the end of the development process almost certainly ensures that the progress variables will not accurately reflect the nature of the curriculum, and the assessments themselves will be less representative of the instructional style.

## The IEY Example Continued

The IEY progress variables formed the framework for the development of almost the entire IEY curriculum—both the IEY instructional materials and the assessments were built around a core set of progress variables. The only exception was that we needed to develop some of the curriculum materials (i.e., the first 12 activities or so) before there were enough concrete materials available for all the developers to engage fully in the debate about the progress variables. Once the five progress variables were established, all instructional objectives for each activity and all of the assessment tasks are linked to one (or more) of the five IEY variables (including the first 12 or so just mentioned). Following agreement on the progress variables, assessments were created that are an integral part of the instruction in IEY (Wilson, Thier, Sloane, & Nagle, 1996). The variety of assessment tasks used for assessment in IEY match in range the variety of instructional events: These include individual and group challenges, data processing questions, and questions following student readings. All assessment prompts are open-ended, requiring students to fully explain their responses. For the vast majority of assessment tasks, the student responses are in a written format, reflecting the only practical way we had available for teachers to attend to a classroom of student work.

Two examples of assessment prompts are shown in Figure 1. The first is taken from IEY Activity 19: "Is Neutralization the Solution to Pollution?" The second is

1. Based on everything you have learned about acid-base neutralization, do you think neutralization can be used as a solution to acid-base pollution? Explain your answer thoroughly. Describe both the advantages and disadvantages and other factors that must be considered before reaching a conclusion.

2. You are a public health official who works in the Water Department. Your supervisor has asked you to respond to the public's concern about water chlorination at the next City Council meeting. Prepare a written response explaining the issues raised in the newspaper articles. Be sure to discuss the advantages and disadvantages of chlorinating drinking water in your reponse, and then explain your recommendation about whether the water should be chlorinated.

FIGURE 1 Two Issues, Evidence, and You assessment tasks.

taken from IEY Activity 12: "The Peru Story." Both are typical in that they require students to integrate information from readings they did in previous activities and labs (the newspaper articles), and also asks them to explain their reasoning. They cannot be fully answered without access to the curricular materials that preceded them. Both are related to the ET variable. As with most IEY assessments, these prompts have multiple components that must be considered, and they outline for the students what is expected in their responses. There is no one correct answer; rather, students are required to make a statement or decision, and then justify it with the information and evidence they have learned through the activities. Their performance is judged by the validity of the arguments they present, not simply the conclusion that they draw.

To provide snapshots of student performance at certain points in the school year and to efficiently create the maps we describe later, we found that extra information was needed at regular points in the curriculum. Our response was to create

You run the shipping department of a company that makes glass kitchenware. You must decide what material to use for packing the glass so that it does not break when shipped to stores. You have narrowed the field to three materials: shredded newspaper, Styrofoam® pellets, and cornstarch foam pellets. Styrofoam® springs back to its original shape when squeezed, but newspaper and cornstarch foam do not. Both Styrofoam® and cornstarch foam float in water. Although Styrofoam® can be reused as a packing material, it will not break down in land fills. Newspaper can be recycled easily, and cornstarch easily dissolves in water.

Which material would you use? Discuss the advantages and disadvantages of each material. Be sure to describe the trade-offs made in your decision.

FIGURE 2 A link item associated with the evidence and tradeoffs variable.

"link tests" that are composed of fairly traditional looking items, each linked to at least one variable, that are not curriculum embedded like the assessment tasks, and that require short-to-moderate length written responses. An example is shown in Figure 2.

Link tests are a series of tests given at major transition points in the IEY course. Each test contains open-ended items related to the content of the course that further assess students' abilities with the IEY variables. Items on the link tests can also be used as an item bank for teachers to draw on in designing their own end of unit or other tests to be administered during the year. Teachers can use the link test items as models of variable-linked, open-ended questions, or they may select specific items to be included in other teacher-made tests. Teachers have also found the link tests useful for grading.

## TEACHER MANAGEMENT AND RESPONSIBILITY

Explanations of the distinctions between alternative assessment procedures and more traditional forms of testing frequently emphasize the importance of teachers' roles in mediating and interpreting the alternative assessment results within the classroom context (e.g., Chittenden, 1991; Wolf, Bixby, Glenn, & Gardner, 1991) as well as the more immediate and meaningful uses of the alternative assessment procedures in the ongoing instructional process (Cole, 1991; Darling-Hammond & Ancess, 1996). However, it is also recognized that such integration will require new views of the teaching and learning process, new roles for (and demands on) teachers, or even a new "assessment culture" in the classroom (Brown et al., 1992; Cole, 1991; Resnick & Resnick, 1992; Torrance, 1995a, 1995b; Zessoules & Gardner, 1991). Preparing teachers to use these types of assessments in their classroom teaching may be a difficult challenge. Teachers' understanding and acceptance of

innovations are crucial to the ultimate success of change (Airasian, 1988; Stake, 1991a, 1991b).

The third principle, therefore, that must be considered in building a classroom-based assessment system, is that teachers must be the managers of the system and hence, must have the tools to use it efficiently and use the assessment data effectively and appropriately. New forms of assessment, although perhaps more valid and more interesting and challenging to the student, make new demands on the teacher. For example, how can a teacher focus on rating one student's performance to the exclusion of monitoring other students' activities? How can teachers manage the additional time demands of scoring open-ended responses generated by 150 students? How can qualitative statements describing levels of performance be translated into letter grades, as required (and expected) by administrators, parents, and the students themselves? Any successful classroom-based system must take into account the demands placed on the teacher for administering, scoring, interpreting, and reporting student performance.

There are two broad issues involved in the teacher management and responsibility principle. First, it is the teachers who will use the assessment information to inform and guide the teaching and learning process. Alternative assessments conducted as part of district or statewide accountability programs, no matter how valid or appropriate to what is taught in the classroom, cannot provide the immediate feedback to teachers that is necessary for instructional management and monitoring (Haney, 1991; Resnick & Resnick, 1992). For this function of assessment teachers must be:

1. Involved in the process of collecting and selecting student work.
2. Able to score and use the results immediately—not wait for scores to be returned several months later.
3. Able to interpret the results in instructional terms.
4. Able to have a creative role in the way that the assessment system is realized in their classrooms.

Only then will teachers really be able to use the assessment system.

Second, issues of teacher professionalism and teacher accountability demand that teachers play a more central and active role in collecting and interpreting evidence of student progress and performance (Tucker, 1991). If they are to be held accountable for their students' performance, teachers need a good understanding of what students are expected to learn and of what counts as adequate evidence of student learning. They are then in a better position, and a more central and responsible position, for presenting, explaining, and defending their students' performances and the outcomes of their instruction.

Central to the incorporation of teacher management and responsibility into the BEAR Assessment System is the use of a single scoring guide for each variable,

which allows teachers to score student work on the assessment tasks efficiently, without having to learn a new set of criteria for each task. The scoring guides can be made concrete for each task by including examples of scored student work, which also helps teachers remain aligned with the general intentions of the assessment task developers.

## The IEY Example Continued

For the information from assessment tasks and link items to be useful to IEY teachers, it must be couched in terms that are directly interpretable with respect to the instructional goals of the IEY variables. Moreover, this must be done in a way that is intellectually and practically efficient. Our response to these two issues are the IEY Scoring Guides. IEY Scoring Guides define the elements of each variable and describe the performance criteria, or characteristics, for each score level of the element. There is one scoring guide for each of the five IEY variables, with each variable having between two and four elements (and the scoring guide is specific to each of these elements). A student's level of performance on an assessment task is determined by using the scoring guide(s) for the variable(s) being assessed. The guide is used throughout the entire course for all assessments relating to a particular variable. This means that there will inevitably be a need for interpretation of the scoring guide for any particular assessment. We found that the combination of a uniform scoring guide with examples for individual assessments was much more efficient for teachers than having different scoring guides for each assessment.

Each IEY Scoring Guide uses a general logic (adapted from the SOLO Taxonomy; Biggs & Collis, 1982) based on discerning what would be under most circumstances, a complete and correct response, this is coded "3." Below this a student might give a partially correct response that leaves out at least one essential element, this is coded a "2." Below this, a student might give a response that has only one correct aspect to it ,this is coded a "1." A response that has no aspects that are relevant is coded zero, and a response that goes beyond a "3" in some significant way is coded a "4." All IEY Scoring Guides share this structure but use specific criteria to adapt them uniquely to individual IEY variables and elements. The ET variable scoring guide is found in Figure 3.

To interpret each scoring guide, teachers need concrete examples—which we call *exemplars*—of the rating of student work. Exemplars provide concrete examples of what a teacher might expect from students at varying levels of development along each variable. They are also a resource, available to the teachers as part of the assessment system, which help them to understand the rationale of the scoring guides.

Actual samples of student work, scored and moderated by teachers who pilot-tested the BEAR Assessment System using IEY, are included with the docu-

**Evidence and Tradeoffs (ET) Variable**

| Score | *Using Evidence:* Response uses objective reason(s) based on relevant evidence to support choice. | *Using Evidence to Make Tradeoffs:* Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice. |
|---|---|---|
| 4 | Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence. | Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence & explaining how it influences choice. |
| 3 | Response provides major objective reasons AND supports each with relevant & accurate evidence. | Response discusses at least two perspectives of issue AND provides objective reasons, supported by relevant & accurate evidence, for each perspective. |
| 2 | Response provides some objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete. | Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete & accurate perspective has been provided. |
| 1 | Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity. | Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence. |
| 0 | No response; illegible response; response offers no reasons AND no evidence to support choice made. | No response; illegible response; response lacks reasons AND offers no evidence to support decision made. |
| X | Student had no opportunity to respond. | |

FIGURE 3 The evidence and tradeoffs scoring guide.

mentation for IEY. These illustrate typical responses for each score level for specific assessment activities. An example of a Level 3 response from Activity 12 is shown in Figure 4.

Teachers have found the exemplars to be very helpful in learning to use the scoring guides because they provide concrete examples of student work at each scoring level. Exemplars are available for all IEY variables except GI, which have not been collected to date for logistical reasons. With practice, teachers may not

| Level 3 | Uses relevant and accurate evidence to weigh the advantages and disadvantages of multiple options, and makes a choice supported by the evidence. |
|---|---|
| "As an edjucated employee of the Grizzelyville water company, I am well aware of the controversy surrounding the topic of the chlorination of our drinking water. I have read the two articals regarding the pro's and cons of chlorinated water. I have made an informed decision based on the evidence presented the articals entitled "The Peru Story" and "700 Extra People May bet Cancer in the US." It is my recommendation that our towns water be chlorin treated. The risks of infecting our citizens with a bacterial diseease such as cholera would be inevitable if we drink nontreated water. Our town should learn from the country of Peru. The artical "The Peru Story" reads thousands of inocent people die of cholera epidemic. In just months 3,500 people were killed and more infected with the diease. On the other hand if we do in fact chlorine treat our drinking water a risk is posed. An increase in bladder and rectal cancer is directly related to drinking chlorinated water. Specifically 700 more people in the US may get cancer. However, the cholera risk far outweighs the cancer risk for 2 very important reasons. Many more people will be effected by cholera where as the chance of one of our citizens getting cancer due to the water would be very minimal. Also cholera is a spreading diease where as cancer is not. If our town was infected with cholera we could pass it on to millions of others. And so, after careful consideration it is my opion that the citizens of Grizzelyville drink chlorine treated water." | Comment<br><br>Both sides of the chlorinating issue have been presented and supported. The choice to chlorinate was made. |

FIGURE 4 Exemplar for a Level 3 response on Activity 12, "The Peru Story" (element scored: using evidence to make tradeoffs).

need to refer to the exemplars; however, these are included as a resource for teachers' use whenever they find them helpful.

In addition to the scoring guides, the teacher needs a tool that indicates when assessments might take place and what variables they pertain to. The assessment blueprints are a valuable teacher tool for keeping track of when to assess students. Assessment tasks are distributed throughout the course at opportune points for checking and monitoring student performance, and these are indicated in the assessment blueprints. Teachers can use these blueprints to review and plan for the progression of assessment tasks relating to each variable.

The assessment blueprints (see Figure 5), designed to correspond with the four parts of the IEY course, list all the IEY activities. Assessments are indicated under the appropriate IEY variables for each activity in which there is an assessment. In addition, the main IEY content concepts addressed by each UC assessment task have also been identified and are listed in the blueprint.

## QUALITY EVIDENCE

The technical quality of performance assessments has been explored and debated primarily in the realm of high-stakes testing situations, such as statewide assessment systems. For classroom-based alternative assessment procedures to gain

| Activity | Variables and Elements | | | | |
|---|---|---|---|---|---|
| | **Designing and Conducting Investigations (DCI)**<br><br>* Designing Investigation<br>* Selecting & Recording Procedures<br>* Organizing Data<br>* Analyzing and Interpreting Data | **Evidence and Tradeoffs (ET)**<br><br>* Using Evidence<br>* Using Evidence to Make Tradeoffs | **Understanding Concepts (UC)**<br><br>* Recognizing Relevant Content<br>* Applying Relevant Content | **Communicating Scientific Information (CM)**<br><br>* Organization<br>* Technical Aspects | **Group Interaction (GI)**<br><br>* Time Management<br>* Role Performance/ Participation<br>* Shared Opportunity |
| 1 - Water Quality | | | | | |
| 2 - Exploring Sensory Thresholds | | | √: Both Elements *(Measurement and Scale)* | | |
| 3 - Concentration | | | √: Applying Relevant Content | | |
| 4 - Mapping Death | | | | | √: Time Management; Shared Opportunity |
| 5 - John Snow and Search for Evidence | | A: Using Evidence | | A: Both Elements | |
| 6 - Contaminated Water | √: Designing Investigations | | | | |
| 7 - Chlorination | A: All Elements | | | | |

FIGURE 5 Partial assessment blueprint.

"currency" in the assessment community, issues of technical quality will have to be addressed as well. Despite the plea of Wolf et al. (1991), the development of practical procedures for establishing the technical quality of classroom-based alternative assessments lags behind that for high-stakes assessment programs. One approach is called an *assessment net* (Wilson & Adams, 1992, 1996), which is composed of (a) a framework for describing and reporting the level of student performance along achievement continua, (b) the gathering of information through the use of diverse indicators based on observational practices that are consistent both with the educational variables to be measured and with the context in which that measurement is to take place, and (c) a measurement model that provides the opportunity for appropriate forms of quality control. The assessment net concept is the basis for the formal measurement approach used next. The description of the BEAR Assessment System in this article should make it clear that it has been planned an example of an assessment net.

It is not sufficient that alternative forms of assessment should express new ideas of instructional validity, they must also maintain the standards of fairness (such as consistency and unbiasedness) that have been accepted as standards for traditional assessments. Doing so involves many qualitative and technical challenges. On a logistical level, using open-ended or performance-based tasks require different procedures for collecting, managing, and scoring student work. Records of performances must be catalogued and stored. Responses can no longer be scanned by machine and entered directly into a statistical database. Raters must score the work, and to do so raises issues of time and cost as well as technical issues involved in rater fairness (e.g., consistency and reliability). There has been a tendency for the arguments surrounding new and conventional forms of assessment to be framed as a shift from an emphasis on reliability to a stronger focus on validity. This argument is bolstered, perhaps, by the long-accepted truism that teacher-made (i.e., classroom-based) student assessments have greater curricular or instructional validity in some sense, but will not have the strong technical properties of more carefully constructed standardized tests.

For classroom-based assessment to gain currency in educational reform, we contend that these assessments must be held to standards of fairness in terms of quality control. Teachers will continue to construct teacher-made tests and will rarely take the steps to establish the comparability or validity of these instruments. However, classroom-based assessment procedures can be developed for specific curricula and made available for teachers' use and adaptation. The evidence generated in the assessment process should be judged by its suitability for purposes of individual assessment and for purposes of evaluating student performance, instructional outcomes, or program effectiveness.

To ensure comparability of results across time and context, procedures are needed to (a) examine the coherence of information gathered using different formats, (b) map student performances on the progress variables, (c) describe the

structural elements of the accountability system—tasks and raters—in terms of the achievement variables, and (d) establish uniform levels of system functioning in terms of quality control indexes such as reliability. The traditional elements of test standardization, such as validity and reliability studies and bias and equity studies, must be carried out within the quality control procedure. To meet this need we propose the use of generalized forms of item response models. We believe that generalized item response models such as those described by Adams and Wilson (1992, 1996), Kelderman (1989), Linacre (1989), and Thissen and Steinberg (1986), have now reached levels of development that make their application to many forms of alternative assessment in a fairly routine way quite feasible. The output from these models can be used as quality control information and can be used to obtain student and school locations on the achievement variables, which may be interpreted both quantitatively and substantively. The formal nature of these models, and their flexibility, allow one to address technical challenges inherent in the classroom assessment situation, such as the maintenance of teacher rating consistency and the maintenance of a meaningful scale throughout the school year.

Apart from traditional quality control indexes such as tables of reliability coefficients and standard errors, the BEAR Assessment System incorporates advances from item response models that can put richer interpretational information into the hands of teachers in the classroom. The central feature of that is the so-called "progress map," which provides a criterion referenced graph of the progress that students are making through the curriculum. Many examples of such maps have been produced for tests over the last 20 or so years (for a large number of examples, see the "Practice" chapters in Engelhard & Wilson, 1996; Wilson, 1992, 1994c; Wilson & Engelhard, in press; Wilson, Engelhard, & Draney, 1997). We illustrate these maps in the discussion of the following IEY example.

## The IEY Example Continued

We have developed maps of the IEY variables. These are graphical representations of a variable, showing how it unfolds or evolves over the year in terms of student performance on assessment tasks. They are derived from empirical analyses of student data collected from IEY teachers' classrooms. The analyses for these maps were performed using ConQuest software (Wu, Adams, & Wilson, 1998), which implements an estimation–maximization algorithm for the estimation of multidimensional Rasch-type models (for a detailed account of the estimation and model-fitting, see Draney & Peres, 1998).

Once constructed, maps can be used to record and track student progress and to illustrate the skills a student has mastered and those that the student is working on. A map of student performance on the ET variable can be found in Figure 6. By placing students' performance on the continuum defined by the map, teachers can demonstrate students' progress with respect to the goals and expectations of the

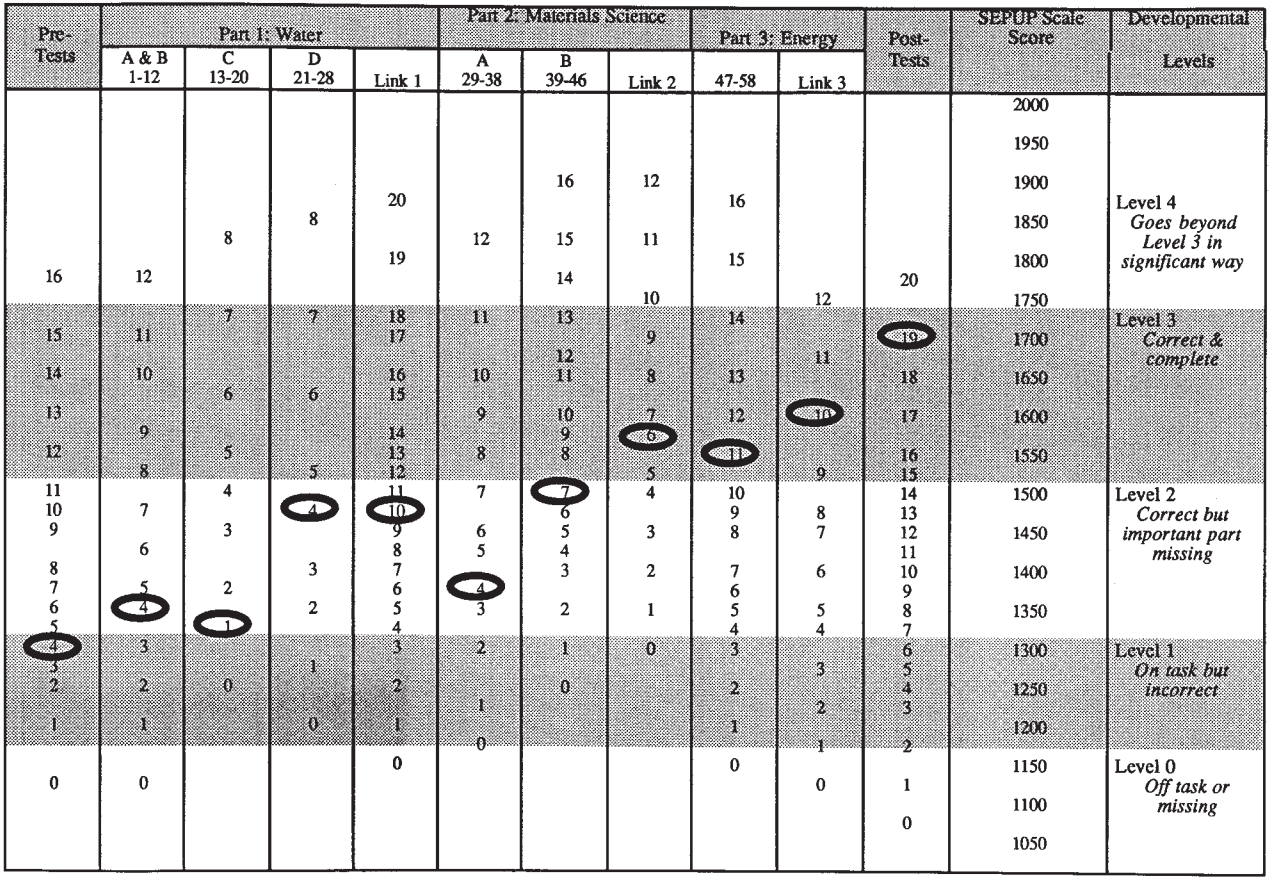


FIGURE 6 Evidence and tradeoffs variable map. SEPUP = Science Education for Public Understanding Project.

TABLE 2
Reliabilities of Link Tests for the Four IEY Progress Variables and the Total Composite

| | *Unidimensional* | | | | *Multidimensional* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Time* | *DCI* | *ET* | *UC* | *CM* | *DCI* | *ET* | *UC* | *CM* | *Total* |
| 0 | NA | .55 | .54 | .64 | NA | .73 | .69 | .65 | .79 |
| 1 | .74 | .80 | .68 | .63 | .83 | .83 | .82 | .83 | .88 |
| 2 | .80 | .80 | .84 | .63 | .90 | .83 | .85 | .78 | .90 |
| 3 | .79 | .73 | .76 | .80 | .85 | .80 | .85 | .80 | .91 |

*Note.* IEY = Issues, Evidence, and You; DCI = designing and conducting investigations; ET = evidence and tradeoffs; UC = understanding concepts; CM = communicating scientific information.

course. The maps, therefore, are one tool to provide feedback on how students as a whole are progressing in the course. They are also a source of information to use in providing feedback to individual students on their own performances in the course.

Maps, as graphical representations of student performance on assessment tasks, can be used by teachers for their own instructional planning and to show students, administrators, and parents how students are developing on the IEY variables over the year. As a result of teachers managing and using the BEAR Assessment System, maps can be produced that allow them to assess both individual and class progress. This can then be used to inform instructional planning. For instance, if the class as a whole has not performed well on a variable following a series of assessments, then the teacher might feel the need to go back and readdress those concepts or issues reflected by the assessments.

The traditional indexes of quality control for assessments are also available, (item response model) reliabilities for the link test for each part of the assessment system have been calculated and are given in Table 2. These have been estimated using a reliability approach suitable for marginal maximum likelihood modeling described in Mislevy, Beaton, Kaplan, and Sheehan (1992). Two values are given for each progress variable for each time period: The first (labeled "unidimensional") is based only on the link items pertaining to that variable. The second (labeled "multidimensional") is based on a multidimensional item response theory model and uses information from all the link items available in each period—effectively, this takes advantage of the correlations among all the progress variables. Note that there were insufficient DCI items in the Time 0 link test (pretest), so no reliabilities were calculated. The column labeled "Total" gives the reliabilities of a composite variable consisting of items from all four progress variables. Some overall patterns can be discerned. The initial period tends to show considerably lower reliability compared to the other periods across all progress variables: This may be due to the relative unfamiliarity of students with both the variables and the item formats at that time. As we might expect, the multidimensional reliabilities are noticeably higher than the unidimensional. This reflects (a) the greater information involved in the full vector

of link text results compared to the information from just the items for a specific variable and (b) the positive correlation among the four progress variables. These multidimensional reliabilities are almost as high as the reliabilities on the total test composite, indicating that this strategy allows one to gain greater interpretational power (i.e., by having four variables rather than one variable to interpret), at the cost of only slightly lower reliability.

A more useful indicator of quality control for individual assessments is the standard error of measurement. This can be expressed on the IEY maps by indicating 95% confidence intervals directly on the maps. For example, using only ET information, for the student location in the first column of Figure 6, the 95% confidence interval extends approximately 40 points on either side of the location (using the units on the "SEPUP Scale Score" on the right side). For the remainder of the locations on the map, the 95% confidence intervals fluctuate from approximately 50 points on either side to approximately 70 points on either side (see details in Wilson et al., in press). We can interpret that to mean, given the information we have, we can say, with 95% confidence, that the true location of the student is within that range and that the true growth curve for this student lies within a slightly wider band throughout the range of the school year (once again, with 95% confidence).

## BRINGING IT ALL TOGETHER

These principles are not designed to operate in isolation. Each of the principles provides a unifying "thread" throughout the system, but their relations also makes the system more integrated. For example, the developmental variables provide an initial unity to the curriculum materials. This framework defines not only the content of student learning but also the paths over which student learning develops throughout the year. The implication is that each assessment, then, has a designated place in the instructional flow, reflecting the type of learning that students are expected to demonstrate at that point in time. Hence, scores assigned to student work can then be linked back to the developmental framework and used both to diagnose an individual's progress with respect to a given variable but also to map student learning over time.

Adherence to each of the principles across each of the phases of the assessment process produces a coherence or internal consistency to the system. Adherence to each of the principles within each phase of the assessment process produces a comprehensive or well-integrated system that can address the complexity of the classroom context and the desired linkages among curriculum, instruction, and assessment (Wilson, 1994a).

Proper operation of the BEAR Assessment System requires that teachers take control of essential parts of the assessment system, including the scoring process and also demands that they grow professionally to master the system. We have de-

vised the assessment moderation meeting as part of our staff development strategy to accomplish these goals. Assessment moderation also plays a crucial role in achieving the quality evidence principle.

Moderation is the process in which teachers discuss student work and the scores they have given that work, making sure that the scores are being interpreted in the same way by all teachers in the moderation group. In moderation sessions, teachers discuss the scoring, interpretation, and use of student work and make decisions regarding standards of performance and methods for reliably judging student work related to those standards. Moderation sessions also provide the opportunity for teachers to discuss implications of the assessment for their instruction, for example, by discussing ways to address common student mistakes or difficult concepts in their subsequent instruction. The moderation process gives teachers the responsibility of interpreting the scores given to students' work and allows them to set the standards for acceptable work. Teachers use moderation to adapt their judgments to local conditions. On reaching consensus on the interpretations of score levels, teachers can then adjust their individual scores to better reflect the teacher adapted standards. The use of moderation allows teachers to make judgments about students' scores in a public way with respect to public standards, and improves the fairness and consistency of the scores across different teachers.

## EVALUATION

The important bottom-line question still remains: Does the assessment system make a difference to students' knowledge? We use evidence from the field test of the IEY curriculum as an example in the following paragraphs.

The IEY assessment system was field tested in the 1994–1995 school year, at six centers around the country known as Assessment Development Centers (ADC). The ADC teachers were required to use the IEY assessment system and to participate in moderation meetings throughout the year. In each of these classrooms, data were collected from the assessment activities and from the link tests. A set of link items was administered in the fall as a pretest and again in the spring as a posttest—Link Test 3I was given at approximately the same time as the posttests. In addition, there were Professional Development Centers (PDC), where teachers taught the IEY curriculum and were provided with the same assessment materials as the ADC group but were not required to use the assessment system. They did, however, participate in professional development activities of similar duration as the moderation activities (these were related to the curriculum rather than to the assessment system). Teachers in both the ADCs and the PDCs were volunteer teachers and hence, could be expected to be more committed and experienced than the average middle school teacher. The choice between ADC and PDC was made on the basis of success in returning data during the pilot test year. This we expect was

associated with the quality of the management at the centers but was not tied to the quality of teachers in each center. Each center, both ADC and PDC, was also asked to choose a comparison teacher, who was to be as similar as possible to the other teachers in the center but taught the regular middle school science curriculum. This process resulted in 26 ADC teachers, 25 PDC teachers, and 12 comparison teachers. Note that we did not specify anything concerning the selection of students of the ADC, PDC or comparison teachers, except that they be at an appropriate grade level (7 to 9).

To examine the effectiveness of the assessment system, the results of the pretests and posttests for the PDC and comparison groups, and the pretests, posttests, and link tests for the ADC group will be compared. The criterion-referenced description of the variable developed in the previous sections will allow these differences to be substantively interpreted.

For pretest and posttest group comparisons, the data were analyzed as follows: Only persons with data at both the pretest and the posttest were considered for analysis. The difficulties for all link items were computed using a rating scale model (Andrich, 1978) for each IEY variable to analyze data from a calibration sample equivalent to the pretest sample. Item difficulties for the posttest were then anchored to their pretest values, and person abilities computed for the posttest. This analysis was done for overall IEY scores (which we called IEY total). Figure 7 illustrates the progress of the ADC, PDC, and combined comparison groups during the 1994–1995 school year, and the relative sizes of the gains from pretest to posttest. The ADC group (the only group assessed throughout the year) made steady progress for the first three time periods. At the time of the posttest, however, there was a notable drop in student proficiency. There are several possible explanations for this. One is that the posttests were most often administered during the last week or two of the school year, which tends to be a busy time for both teachers and students. It is possible that the administration of some of the posttests was somewhat rushed, and that both students and teachers were distracted from the task. Also, as can be seen in the figure, much more time was spent on the first sections of the curriculum than on the last section. In particular, most ADC classes spent nearly half of the school year on Part 1 (water) of the curriculum, and many spent only a month or so on Part 3 (energy). Thus, one would expect that students would have felt much more secure with material from Part 1 than with material from Part 3. Because questions from Part 3 are included in both Link 3 and the posttest, this may adversely affect the proficiency estimates of students at the time of the posttest. Even with this drop in proficiency, however, it is clear that ADC students made substantial gains during the school year, especially as compared with PDC and comparison students. Regardless of the drop in performance at the end of the year, the ADC students made progress that is both statistically and educationally significant. In statistical terms, the difference between average gain for PDC and comparison groups was not statistically significant at the $\alpha = 0.05$ level.

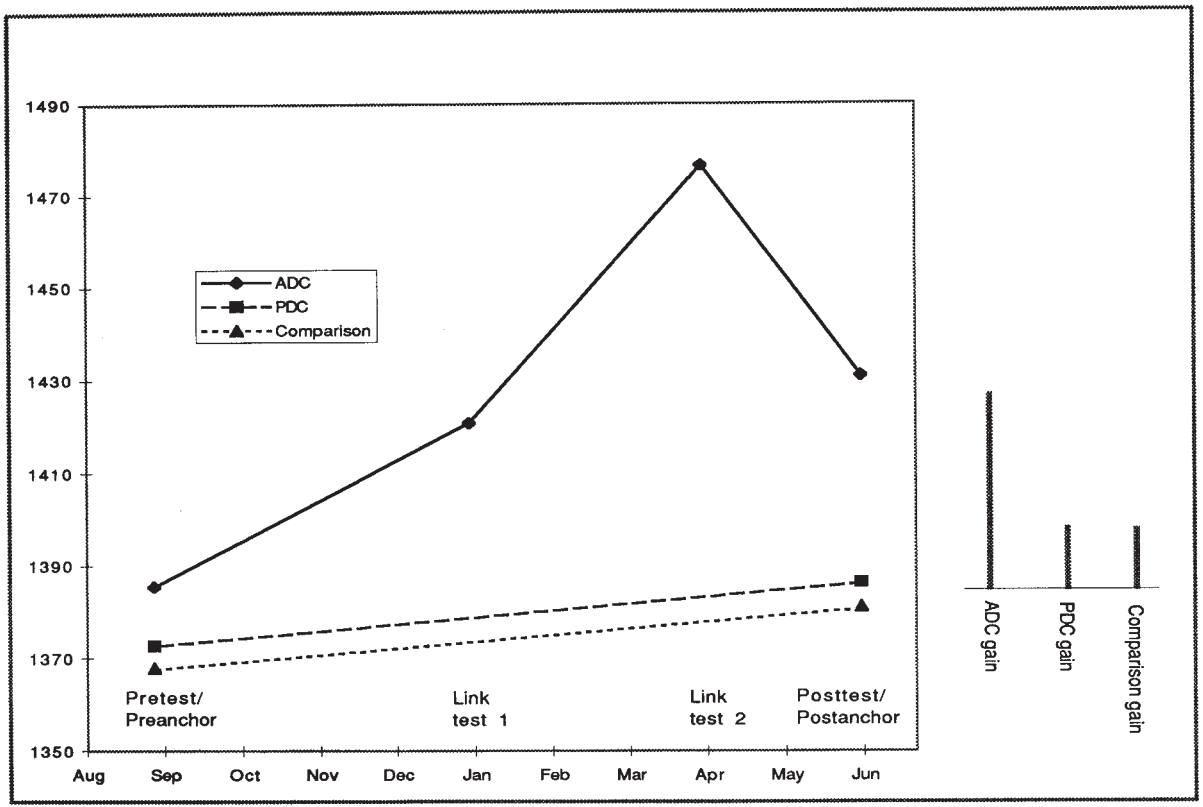


FIGURE 7 Gains for the Assessment Development Centers (ADC), Professional Development Centers (PDC), and comparison groups on the combined Issues, Evidence, and You variable during the 1994–1995 school year.

TABLE 3
Means and Standard Deviations for Three Groups at Pretest and Posttest

| | *ADC* | | *PDC* | | *Comparison* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Pretest | 1,386 | 50.0 | 1,373 | 36.9 | 1,368 | 60.0 |
| Posttest | 1,431 | 56.3 | 1,386 | 36.9 | 1,381 | 61.9 |
| Gain | 45 | | 13 | | 13 | |

*Note.* ADC = Assessment Development Centers; PDC = Professional Development Centers.

Differences between the average gain for the ADC group and either of the other two groups were statistically significant at the $\alpha = 0.05$ level. We can calculate effect sizes for this in the following way. The means and standard deviations for the three groups at the pretests and posttests are shown in Table 3 (in the rescaled IEY units). Assuming that the comparison group represents a set of typical classes, we can use it as a baseline against which to compare the ADC group. In those terms, the gain for the comparison group is 13/60 = 0.22 standard deviation units—that is, in the comparison group, the mean student at the 50th percentile, would move to the equivalent of the 59th percentile (assuming a normal distribution) over the course of the year. The gain for the ADC group (using the standard deviation of the comparison group as a yardstick), is 45/60 = 0.75 standard deviation units, which would move the mean student at the 50th percentile, to the equivalent of the 77th percentile (assuming a normal distribution) over the course of the year. This is a much larger gain than for the comparison group—it is a gain that is 3.46 times greater. This gain can also be put in educational terms: The gain for the ADC group is, for the average student, approximately a difference between a 2 and a 2.5 on the scoring guide. This is an educationally significant change, marking a difference between a student who typically achieves partial success, and one who achieves satisfactory completion about half the time.

## CONCLUSIONS

The comparison between the gains made by ADC students and the gains made by PDC students is of particular significance. Both groups used the same curriculum. The same instructional and assessment materials were provided to both groups. The only difference in treatment between the two groups was that teachers in the ADC group paid specific attention to the assessment system. They were trained in its use and had a professional development program specifically aimed at the assessment system. Thus, the curriculum by itself is not enough to produce the kinds of changes in student performance seen in the ADC group. It is specifically the focus on assessment that has produced this kind of educationally significant student growth. These

results make clear that there are considerable potential gains that could be realized by (a) closer attention to assessment concerns at the classroom level and (b) a more systematic approach to the gathering and interpretation of assessment information.

There are several major avenues of improvement in the implementation of an embedded assessment system that can be discerned. One anomaly noted previously in the IEY system is that one of the variables—GI—was not mapped because a systematic mode of data collection was not available. What is needed here is an effective means of recording teacher judgments of student performances, as they occur, live in the classroom. A second issue concerns the need for a standard set of assessment tasks within each section of the course. Although this was a feature of the field test (required for gathering sufficient calibration data), it is not generally necessary, teachers need to be able to choose among the possibilities inherent in the curriculum and hence, need to be able to choose among the assessment tasks. There is no theoretical barrier to this, but it is impractical to implement it without the availability of a microcomputer in the classroom for calculating and printing out custom versions of the maps. A personal digital assistant strategy for addressing the first problem, and computer program capable of carrying out the calculations necessary for the second, are currently under field testing.

The moderation meetings are a rich and valuable teacher development strategy that we have barely skimmed in our developmental activities so far. They may be useful in ways that go well beyond assessment and deserve considerable research attention in the coming years. The classroom assessment context provides a wealth of interesting issues to be addressed in the area of formal modeling. The issues of the importance of multidimensionality, modeling teacher rating behavior, exploring richer diagnostic models, and establishing reliable growth curves are only a few that might be mentioned. The work reported here went only a few short steps along each of these roads.

## ACKNOWLEDGMENTS

This article is dedicated to the memory of Chris White.

This project was supported by National Science Foundation Grants MDR9252906 and ESI–9553548. The project described here began in 1993 and has benefited from the contribution of all the members of the Science Education for Public Understanding Project (SEPUP) Development Team (at the Lawrence Hall of Science, University of California, Berkeley) and the SEPUP Assessment Team (at the Graduate School of Education, University of California, Berkeley). In particular, we would like to acknowledge the work of Herb Thier, Barbara Nagle, Mike Reeske, and Bob Horvat, in the development of the Issues, Evidence, and You (IEY) curriculum and in working with us to develop the embedded assessment materials. Members of the assessment team have also included Lily Roberts, Sara Samson, Robin Henke, Amy Jackson, Megan Martin, Chris White, Dick Duquin, Harry Case, Eric

Crane, Karen Draney, and Henry Stavinsky—all of whom made substantial contributions to the development and analyses of the assessment materials. And, we owe a special thanks to all of the teachers and administrators who participated in the field test of these materials and who provided the valuable feedback for improvements to the system and its implementation. Opinions reflect those of the authors and do not necessarily reflect those of the granting agency.

## REFERENCES

Adams, R. A., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 143–166). Norwood, NJ: Ablex.

Adams, R. J., & Wilson, M. (1992, April). *A random coefficients multinomial logit: Generalizing Rasch models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Airasian, P. W. (1988). Measurement-driven instruction: A closer look. *Educational Measurement: Issues and Practice, 7*(4), 6–11.

American Psychological Association, American Educational Association, & National Council for Measurement in Education. (1985). *Standards for psychological and educational tests*. Washington, DC: Authors.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Baron, J. B. (1991). Performance assessment: Blurring the edges of assessment, curriculum, and instruction. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 247–266). Washington, DC: American Association for the Advancement of Science.

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic.

Brown, A. L., Campione, J. C., Webber, L. S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 121–212). Boston: Kluwer Academic.

Chittenden, E. (1991). Authentic assessment, evaluation, and documentation of student performance. In V. Perrone (Ed.), *Expanding student assessment* (pp. 22–31). Alexandria, VA: Association for Supervision and Curriculum Development.

Cole, N. (1991). The impact of science assessment on classroom practice. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 97–106). Washington, DC: American Association for the Advancement of Science.

Darling-Hammond, L., & Ancess, J. (1996). Authentic assessment and school development. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education* (pp. 52–83). Chicago: University of Chicago Press.

Draney, K. D., & Peres, D. (1998). *Unidimensional and multidimensional modeling of complex science assessment data* (Tech. Rep. No. SA–98–1). University of California, Berkeley, BEAR Research.

Engelhard, G., & Wilson, M. (Eds.). (1996). *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 77–120). Boston: Kluwer Academic.

Gipps, C. (1995). Reliability, validity, and manageability in large-scale performance assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 105–123). Philadelphia: Open University Press.

Glaser, R. (1987). The integration of instruction and testing: Implications from the study of human cognition. In D. C. Berliner & B. V. Rosenshine (Eds.), *Talks to teachers: A festschrift for NL Gage* (pp. 329–341). New York: Random.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), *Expanding student assessment* (pp. 142–163). Alexandria, VA: Association for Supervision and Curriculum Development.

Kelderman, H. (1989, March). *Loglinear multidimensional IRT models for polytomously scored items*. Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, CA.

Land, R. (1997). Moving up to complex assessment systems. *Evaluation Comment, 7*(1), 1–21.

Linacre, J. M. (1989). *Many faceted Rasch measurement.* Unpublished doctoral dissertation, University of Chicago.

Linn, R., & Baker, E. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education* (pp. 84–103). Chicago: University of Chicago Press.

Malcom, S. M. (1991). Equity and excellence through authentic science assessment. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 313–330). Washington, DC: American Association for the Advancement of Science.

Masters, G. N., Adams, R. A., & Wilson, M. (1990). Charting student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (Supplementary Vol. 2, pp. 628–634). Oxford, England: Pergamon.

McCallum, B., Gipps, C., McAlister, S., & Brown, M. (1995). National curriculum assessment: Emerging models of teacher assessment in the classroom. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 88–104). Philadelphia: Open University Press.

Minstrell, J. (1998, October). *Student thinking and related instruction: Creating a facet-based learning environment.* Paper presented at the meeting of the Committee on Foundations of Assessment, Woods Hole, MA.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–162.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–76). Boston: Kluwer Academic.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Francisco: Jossey-Bass.

Science Education for Public Understanding Project. (1995). *Issues, evidence and you: Teacher's guide*. Berkeley: University of California, Lawrence Hall of Science.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership, 52*(5), 38–43.

Stake, R. (1991a). *Advances in program evaluation: Volume 1, part A. Using assessment policy to reform education*. Greenwich, CT: JAI.

Stake, R. (1991b). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan, 71,* 243–247.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 49,* 501–519.

Torrance, H. (1995a). The role of assessment in educational reform. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 144–156). Philadelphia: Open University Press.

Torrance, H. (1995b). Teacher involvement in new approaches to assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 44–56). Philadelphia: Open University Press.

Tucker, M. (1991). Why assessment is now issue number one. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 3–16). Washington, DC: American Association for the Advancement of Science.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41–47.
Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75,* 200–214.
Wilson, M. (Ed.). (1992). *Objective measurement: Theory into practice.* Norwood, NJ: Ablex.
Wilson, M. (1994a). Community of judgement: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), *Issues in educational accountability* (pp. 1–48). Washington, DC: Office of Technology Assessment.
Wilson, M. (1994b). Measurement of developmental levels. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 1508–1514). Oxford, England: Elsevier Science.
Wilson, M. (Ed.). (1994c). *Objective measurement II: Theory into practice.* Norwood, NJ: Ablex.
Wilson, M. (1999). Relating the National Science Education Standards to the Science for Public Understanding Program (SEPUP) assessment system. In K. Comfort (Ed.), *Advancing standards for science and mathematics education: Views from the field.* Washington, DC: American Association for the Advancement of Science. (Available: http://ehrweb.aaas.org/ehr/forum/)
Wilson, M., & Adams, R. J. (1992, June). *Evaluating progress with alternative assessments: A model for chapter 1.* Invited address to the conference on Curriculum and Assessment Reform, Boulder, CO.
Wilson, M., & Adams, R. J. (1996). Evaluating progress with alternative assessments: A model for chapter 1. In M. B. Kane (Ed.), *Implementing performance assessment: Promise, problems and challenges* (pp. 39–60). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Wilson, M., & Engelhard, G. (Eds.). (in press). *Objective measurement Vol. 5: Theory into practice.* Stamford, CT: Ablex.
Wilson, M., Engelhard, G., & Draney, K. (Eds.). (1997). *Objective measurement IV: Theory into practice*. Norwood, NJ: Ablex.
Wilson, M., Roberts, L., Draney, K., Samson, S., & Sloane, K. (in press). *SEPUP assessment resources handbook.* Berekely: University of California, SEPUP, Lawrence Hall of Science.
Wilson, M., Thier, H., Sloane, K., & Nagle, B. (1996, April). *What have we learned from developing an embedded assessment system?* Paper presented at the annual meeting of the American Educational Research Association, New York.
Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–74.
Wolf, D., & Reardon, S. (1996). Access to excellence through new forms of student assessment. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education* (Part 1, pp. 52–83). Chicago: University of Chicago Press.
Wu, M., Adams, R. J., & Wilson, M. (1998). ConQuest [Computer program]. Melbourne: Australian Council for Educational Research.
Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perrone (Ed.), *Expanding student assessment* (pp. 47–71). Alexandria, VA: Association for Supervision and Curriculum Development.