Using assessment to improve learning: The BEAR Assessment System

Mark Wilson and Kathleen Scalise
University of California, Berkeley
March 2004

Abstract

This paper discusses how assessment practices in higher education can improve or hinder learning. Elements of effective learning environments that may better address underlying metacognitive issues are discussed. The principles of the Berkeley Evaluation & Assessment Research Assessment (BEAR) System are introduced, and their use to improve learning is described in the context of the UC Berkeley ChemQuery project.

Introduction

One of the surpassing mysteries in higher education is how — every semester and all around the country — substantial numbers of students come into class with all the right prerequisites and grades to prepare them to handle their new coursework — but in fact they seldom know what they are supposed to know. Why don't they know it? And furthermore, what does the instructor, especially in large lecture classes where the teaching load is already substantial, do about it?

In "From Naïve to Knowledgeable," Joseph Hesse (1989, p. 55), an instructor and conceptual change investigator in the sciences, said that a usual explanation is a "pass-the-buck" interpretation—somehow the student just didn't study enough, didn't remember enough, wasn't interested enough." Hesse illustrated this by quoting a colleague "who stated that, in his opinion, 90 percent of student mistakes could be attributed to a lack of study on their part. Blame the student! Period! End of discussion. (p. 55)"

Assumptions of lack of studying or insufficient engagement with the material are common explanations of student underperformance. Also, views of fixed intelligence are common (Dweck & Leggett, 1988), in which instructors and even students themselves sometimes take the view that at some point they have "topped out" in their ability to master the material. Throw in the complications of concept retention and knowledge transference, and it is perhaps too easy to justify the existence of underprepared students, and to support "natural" filtering mechanisms that eliminate students through attrition or failing grades.

However, this premise fails to address a set of important issues that we will take on in this paper: Whether the students really did know the material they were responsible for in the first place; how we know they knew it; and whether sound metacognitive principles are in place for instructors and students to monitor and improve student learning processes, optimizing their ability to construct, learn, retain and transfer knowledge.

In other words, is the problem really low ability or disengaged students, or are educational practices contributing to underpreparation and underperformance?

Current State of Formative Assessment Practices

To illustrate what some problematic practices might be, consider the role of formative assessment and feedback, as outlined by the recent National Research Council report, "Knowing What Students Know" (KWSK; Pellegrino, Chudowsky & Glaser, 2001, p. 87):

> "[A] major law of skill acquisition involves *knowledge of results*. Individuals acquire a skill much more rapidly if they receive feedback about the correctness of what they have done. If incorrect, they need to know the nature of their mistake. It was demonstrated long ago that practice without feedback produces little learning (Thorndike, 1931). One of the persistent dilemmas in education is that students spend time practicing incorrect skills with little or no feedback.

> Furthermore, the feedback they receive is often neither timely nor informative. For the less capable student, unguided practice can be practice in doing tasks incorrectly."

The use of homework, laboratories, papers, quizzes, and other activities by which students practice what they have learned is commonplace in education. Often some form of credit is given for the work, and feedback is offered to students, sometimes in the form of a grade, at other times with more extensive critiquing. Especially in large lecture courses, more detailed feedback is often limited by resource constraints.

A major literature survey of over 250 sources on formative assessment (Black & Wiliam, 1998) found that effective assessment practices can play a powerful role in the learning experience, moving an average student, for instance, to the top third of the class — but only if certain conditions are satisfied. Student tasks needed to be aligned, or on target, with learning goals, and students need to receive meaningful and timely feedback on their performance, as well as targeted follow-up work. To regulate learning effectively, students need to understand three things: (a) the measures on which they will be judged, (b) where they stand on these measures, and (c) how they can improve (Black & Wiliam, 1998, p. 143).

Elements of Effective Learning Environments

Clearly, a single summative score in the form of a grade can do little to inform mastery of complex material. Add to this the confounding effect of incorporating "effort" into this single grade, often on a basis that is not clearly defined for students, and one can see that the metacognitive "signal" by which students "tune" their performance has been weakened to the point of failure. The goal should be "rigorous and wise diagnostic information" (Wolf, Bixby, Glenn and Gardner, 1991), but this is seldom made available.

Approaches exist that might facilitate metacognition in large lecture classes. An "embedded assessment" system designed and used in assessment development at the University of California, Berkeley, called the BEAR Assessment System (BAS; Wilson & Sloane, 2000) is described in the following section. It consists of easy-to-use tools for generating solid diagnostic information and feedback, perhaps especially useful in large class settings. The BAS is a comprehensive, integrated system for assessing, interpreting, monitoring, and responding to student performance. It provides a set of tools for instructors and students to:

- reliably assess performance on central concepts and skills in curriculum,
- set standards of performance,
- validly track progress over the year on central concepts, and
- provide mechanisms for feedback and followup.

A note about embedded assessment: The term *embedded assessment* means just what it says: activities are "embedded," or become part of, class learning activities. Instructors do embedded assessment all the time: a homework assignment, a laboratory procedure, a classroom discussion, an essay. Any of these and many more can be considered embedded assessment activities if a student produces something that can be rated, or observed and assessed in some manner. The difference between these examples and what we discuss here as more formal embedded assessment is that the latter calls for attention to task design and formal "calibration" of assessment tasks in relationship to a framework that describes the learning to take place.

The Assessment Triangle and the BEAR Approach

Three broad elements on which every assessment should rest are described by the KWSK Assessment Triangle (Pellegrino et al, 2001, p. 296), shown in Figure 1.

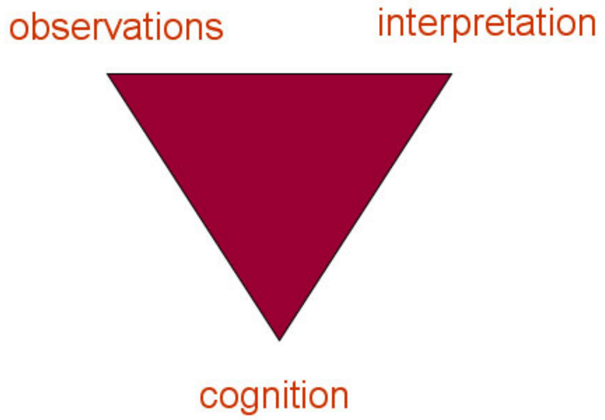

Figure 1. The KWSK assessment triangle.

According to the Committee Report, an effective assessment design requires:

- a model of student cognition and learning in the field of study;
- well-designed and tested assessment questions and tasks, often called items;
- and ways to make inferences about student competence for the particular context of use.

These elements are of course inextricably linked. Models of student learning should specify the most important aspects of student achievement to assess, and they provide clues about the types of tasks that will elicit evidence and the types of inferences that can relate observations back to learning models and ideas of cognition. To serve as quality evidence, items themselves need to be systematically developed with both the learning model and subsequent inferences in mind, and they need to be tried out and the results of the trials systematically examined. Finally, the inferences provide the "why" of it all — if we don't know what we want to do with the assessment information, then we can't figure out what the student model or the items should be. Of course, context determines many specifics of the assessment.

The BEAR Assessment System is based on the idea that good assessment addresses these considerations through four principles: (1) developmental perspective, (2) a match between instruction and assessment, (3) the generating of quality evidence, and (4) management by instructors to allow appropriate feedback, feed forward and follow-up. See Wilson (in press) for a detailed account of an instrument development process that works through these steps. Below we take up each of these issues in turn.

Principle 1: Developmental Perspective

A developmental perspective of student learning means assessing the development of student understanding of particular concepts and skills over time, as opposed to, for instance, making a single measurement at some final or supposedly significant time point. One strategy to address the developmental perspective is specifying a set of "progress variables" (Masters, Adams, and Wilson, 1990; Wilson, 1990). These variables define the most important student growth goals of the curriculum, and change from course to course as different areas of knowledge are the focus of interest and thus assessment.

With a progress variable approach, every instructional unit is seen as contributing to student progress on at least one of these variables, and every assessment is closely aligned with one or more variables. This alignment allows the creation of a calibrated and meaningful scale to map the growth of students, so that instructors can track the progress of individual students and groups of students as they engage in learning.

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
10-12 **Construction** Why composition, structure, properties, and amounts? (Using models)	The composition, structure, and properties of matter are explained by varying strengths of interactions between particles (electrons, nuclei, atoms, ions, molecules) and by the motions of these particles.	Students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models.	a) Composition: How can we account for composition? b) Structure: How can we account for 3-D structure? (e.g., crystal structure, formation of drops,) c) Properties: How can we account for variations in the properties of matter? (e.g., boiling point, viscosity, solubility, hardness, pH, etc.) d) Amount: What assumptions do we make when we measure the amount of matter? (e.g., non-ideal gas law, average mass)
7-9 **Formulation** How can we think about interactions between molecules? (Multirelational)	The composition, structure, and properties, of matter are related to how electrons are distributed among atoms.	Students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning is limited to causal instead of explanatory mechanisms. In their interpretations of new situations students may over-generalize as they try to relate multiple ideas and construct formulas.	a) Composition: Why is the periodic table a roadmap for chemists? (Why is it a "periodic" table?) How can we think about the arrangements of electrons in atoms? (e.g., shells, orbitals) How do the numbers of valence electrons relate to composition? (e.g., transfer or sharing) b) Structure: How do connections between atoms (bonds) and motions of atoms explain 3-D structure? (diamond rigid, water flows, air invisible) c) Properties: How can matter be classified according to bonds? (ionic solids dissolve in water, covalent solids hard, matter phase) d) Amount: How can one quantity of matter be related to another? (e.g., mass/mole/number, ideal gas law, Beer's law)
4-6 **Recognition** How do chemists describe matter? (Unirelational)	Matter is categorized and described by various types of subatomic particles, atoms, and molecules.	Students explore the language and symbols used by chemists to describe matter. They relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. Ways of thinking about matter are limited to relating one idea to another at a simplistic level of understanding.	a) Composition: How does the periodic table show trends? How are elements, compounds, and mixtures classified by letters and symbols? b) Structure: How do the arrangements and motions of atoms differ in solids, liquids, and gases? c) Properties: How can the periodic table be used to predict properties? d) Amount: How do chemists keep track of quantities of particles? (e.g., number, mass, volume, pressure, mole)
1-3 **Notions** What do you know about matter?	Matter has mass and takes up space. It can be classified according to how it occupies space.	Students articulate ideas about matter, and use experience, observation and logical reasoning to provide evidence. Focus is largely on macroscopic (not particulate).	a) Composition: How is matter distinct from energy, thoughts, feelings? b) Structure: How do solids, liquids, and gases differ from one another? c) Properties: How can you use properties to classify matter? d) Amount: How can you measure the amount of matter?

Figure 2: Perspectives of Chemists Framework, Matter Variable.

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 as deeper understandings are developed from and take the place of earlier understandings. To use the BEAR Assessment System 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 research into the underlying cognitive structure of the domain and 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 to instruction or perform in practice (Pellegrino et al, 2000). To more clearly understand what a progress variable is, let us consider an example. A university chemistry assessment project at UC Berkeley called *ChemQuery* recently developed a framework of progress variables called "Perspectives of Chemists" that attempts to embody understanding of chemistry from a novice to expert level of sophistication. Its three variables describe chemistry views regarding three "big ideas" in the discipline: Matter, change, and stability. The *Matter* strand is concerned with describing atomic and molecular views of matter. *Change* involves kinetic views of change and the conservation of matter during chemical change. *Stability* considers the network of relationships in conservation of energy. The Matter progress variable is shown in Figure 2. It describes how a student's view of matter progresses from a continuous, real-world view, to a particulate view, and then builds in sophistication.

Principle 2: Match between Instruction and Assessment

The match between the instruction and assessment in the BEAR Assessment System is established and maintained through two major parts of the system: progress variables, described above, and assessment tasks or activities, described in this section. The main motivation for the progress variables so far developed is that they serve as a framework for the assessments and a method of making measurement possible. However, this 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, nor that the curriculum description will entirely determine the assessment, but rather that the two, assessment and instruction, must be in step—they must both be designed to accomplish the same thing, the aims of learning, whatever those aims are determined to be.

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. In order 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 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, usually where instructors need to see how much progress their students have made on a specific topic. See Minstrell (1998) for an insightful account of such occasions).

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 activities. Doing so brings the richness and vibrancy of curriculum development into assessment, and also brings the discipline and hard-headedness of assessment data into the design of instruction.

By developing assessment tasks as part of curriculum materials, they can be made directly relevant to instruction. Assessment can become indistinguishable from other instructional activities, without precluding the generation of high-quality, comparative, and defensible assessment data on individual students and classes.

The variety of assessment tasks used by the BEAR Assessment System can range widely, including individual and group “challenges,” data interpretation questions, and tasks involving student reading, laboratory, or interactive exercises. In ChemQuery tasks, 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.[1]

Whatever the form of instruction, if student work is generated or students can be observed at work and this work can be scored and matched to progress variables, then it is possible to consider use of an assessment system such as BEAR and to clearly match the assessments to instruction.

Principle 3: Management by Instructors

For information from the assessment tasks and the BEAR analysis to be useful to instructors and students, it must be couched in terms that are directly related to the instructional goals behind the progress variables. Open-ended tasks, if used, must be quickly, readily, and reliably scorable. Our response to these two issues are scoring guides (for instance, rubrics), scorable by people, such as students themselves, by readers, teaching assistants and instructors, or scorable by machine, say, using web-based interfaces with real-time delivery of instructional material and feedback, or more traditional machine-readable answer sheets.

Note that decisions on the structure of tasks and deployment of scoring and guides can be made course by course, but there should be a balance between the time constraints and needs of instructors for automatic machine scoring or reader scoring against the metacognitive needs of students to have instructors understand, engage and react to student levels of performance.

When scoring guides are used, instructors and students need concrete examples—which we call “exemplars”—of the rating of student work. Exemplars provide concrete examples of what a instructor might expect from students at varying levels of development along each variable. They are also a resource to understand the rationale of the scoring guides. Actual samples of student work, scored and moderated by those who pilot-tested the BEAR Assessment System in ChemQuery, are available for each activity. These illustrate typical responses for each score level, as well as atypical responses that exercise the raters’ skills..

In addition to the scoring guides, the instructor needs a tool to indicate when assessments might take place, and what variables they pertain to. These are called Assessment Blueprints and are a valuable 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, as indicated in the Assessment Blueprints.

Principle 4: Evidence of Quality

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure along a progress variable as described above. 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 onto the progress variables, (c) describe the structural elements of the accountability system—tasks and raters—in terms of the achievement variables, and (d) establish acceptable levels of system functioning, in terms of quality control indices such as reliability. While this type of discussion can become very technical to consider, it is sufficient to keep in mind that the traditional elements of assessment standardization, such as validity and reliability studies and bias and equity studies, must be carried out to satisfy quality control and ensure that evidence can be relied upon.

Our approach on this technical end of measurement is to use item response modeling (also known as IRT), as described by Adams and Wilson (1992, 1996). These are examples of measurement models well-developed enough for use in classroom-based

[1] This is not a limitation of the BEAR system, but reflects the only practical way we then had available for instructors to attend to a full classroom of student work.

assessment in a fairly routine and feasible way. The output from these models can be used as quality control information to address the concerns above, and to determine where individual students fall on a progress variable such as ChemQuery's Matter variable, or any other progress variable that might be conceived and validated. Such output was used to validate and calibrate the Matter progress variable, and to create the map of the progress variable in Figure 3. The left side of this map shows the measured distribution of students who responded to the Matter items in 2001-02 trials, and the right side shows the calibrated difficulty of the tasks. Item response modeling can be used to locate a student or describe an entire class along a progress variable, as well as generate fit statistics and other indices for how well levels specified by the model fit classroom data. Tables of reliability coefficients and standard errors are generated, and inter-rater comparisons also can be made.

The formal nature of these models and their flexibility allows one to address technical challenges inherent in the classroom assessment situation, such as the maintenance of instructor rating consistency and the maintenance of a meaningful scale throughout the school year. This puts richer information into the hands of instructors in the classroom. The central feature is the *progress map*, which provides a graph of the progress that students are making through the curriculum. Maps are derived from empirical analyses of student data collected from coursework. [2]

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. By placing students' performance on the continuum defined by the map, instructors can demonstrate students' progress with respect to the goals and expectations of the course. The maps, therefore, are one tool to provide feedback on how students as a whole are progressing in the course.

Maps, as graphical representations of student performance on assessment tasks, can be used to show how students are developing on progress variables throughout the course. 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 instructor might feel the need to go back and re-address those concepts or issues reflected by the assessments. Additionally, during the development stage, unsatisfactory map results can indicate changes or additions to the curriculum.

Bringing It All Together: Assessment Moderation

The four principles of the BEAR system are not designed to operate in isolation. Each of the principles provides a unifying "thread" throughout the system, but their interrelationships also make the system more integrated. For example, the progress variables provide an initial unity to the curriculum materials, and define 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 perspective and used both to diagnose an individual's progress with respect to a given variable and 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 well-integrated system that addresses desired linkages among curriculum, instruction, and assessment.

---

[2] These maps were drawn in GradeMap, a software package developed at the University of California, Berkeley (Wilson, Kennedy & Draney, 2004). The analyses for these maps were performed using the ConQuest software (Wu, Adams and Wilson, 1998), which implements an EM algorithm for estimation of multidimensional Rasch-type models. For details on estimation and model-fitting, see Draney and Peres (1998).

logits

```
=====================================================================
   2            |Task 1, Level 2
                |
                |
                |
                |
                |
                |
                |
                |
                |
                |Task 4, Level 2
               X|
                |Task 3, Level 2
              XX|
   1          XX|
              XX|
             XXX|
             XXX|
            XXXX|Task 2, Level 2; Task 5, Level 2
            XXXX|
            XXXX|
          XXXXXX|
           XXXXX|
          XXXXXX|
        XXXXXXXX|
       XXXXXXXXX|
        XXXXXXXX|
      XXXXXXXXXX|
   0   XXXXXXXXX|
          XXXXXX|Task 6, Level 2!! Out of Level
        XXXXXXXX|
        XXXXXXXX|
        XXXXXXXX|
      XXXXXXXXXX|
         XXXXXXX|
            XXXX|
            XXXX|Task 1, Level 1
              XX|
              XX|Task 2, Level 1; Task 3, Level 1
             XXX|Task 5, Level 1
               X|
  -1           X|
               X|Task 6, Level 1
               X|
               X|
                |
                |
                |Task 4, Level 1
=====================================================================
```

*Numbers at left are in a measurement unit called the “logit,” or log of the odds, with higher numbers indicating better student performance.

Figure 3. Calibration of LBC Visualizing Matter Variable (Partial Credit, Generalised-Item Thresholds)

Proper operation of the BEAR Assessment System requires that instructors and students “take control” of essential parts of the assessment system, including the scoring process. We have devised the “assessment moderation meeting” as part of our staff and student development strategy to accomplish these goals.

Moderation is the process by which instructors, teaching assistants, readers, students and others involved in a course discuss student work and the scores for work, ensuring that scores are interpreted in a similar way by all in the moderation group.

In instructor moderation sessions, instructors 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. The moderation process gives instructors the responsibility of interpreting the scores given to students' work and allows them to set the standards for acceptable work. Instructors use moderation to adapt their judgments to local conditions. Upon reaching consensus on the interpretations of score levels, instructors can then adjust their individual scores to better reflect the instructor-adapted standards. The use of moderation allows instructors to make judgments about students' scores in a public way with respect to public standards and improves the fairness and consistency of the scores.

Moderation sessions also provide the opportunity for instructors to discuss implications of the assessment for their instruction, for example, by discussing ways to address common student mistakes or difficult concepts in instructional sequence. This last aspect of moderation is perhaps the strongest influence of moderation on instruction.

Moderation also can take place with students, so they better grasp what the instructor and course are valuing in terms of student learning. Students can score class work, if that is appropriate, or can score work provided as examples in the curriculum materials. They can map scores against progress variables and see more concretely the paths toward mastery of learning aims. (See video, "Moderation in All Things: A Class Act," Berkeley Evaluation & Assessment Research Center.)

Conclusion

It seems clear that going beyond grades to map individual trajectories of learning is feasible, especially where computers and data collection devices are readily available in higher education and when instructional materials are restructured to use these tools to better accommodate improved formative assessment. In practice, the benefit to students is promising. Wilson & Sloane (2000) have documented evidence of the educationally important and statistically significant effects that use of the system can have on student performance. Furthermore, it is interesting to consider how theories of learning and theories of instruction may change as a result of better data and a clearer understanding of learning trajectories. We hope to see many embedded assessment efforts unfold in coming years, especially in large lecture classes, and invite those interested to, of course, consider using the BEAR Assessment System.

But, moreover, it is important for educators to ponder the principles behind successful formative assessment. It is in satisfying these principles that the argument for formative assessment lies, and the metacognitive needs of students can be met.

“This kind of analysis gives me more than just a grade,” said one instructor using the BEAR system. “I can diagnose a problem and move forward with a greater number of students. I can see the amount of time it takes for my students to learn, and find out how much of something they know, or how well they know it, not just whether they have a fact in their heads or they don’t. It lets me value even wrong answers, because it shows me what in each answer I can value and support and work with. To me, it’s a whole different way to truly value student thinking.”

Acknowledgements

The work of the first author was supported by a National Science Foundation (NSF) grant to the CAESL Center, WestEd. The work of the second author and the ChemQuery examples cited in the text were produced and supported by an NSF grant to Professor Angelica Stacy, Department of Chemistry, University of California, Berkeley.

References

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

Adams, R. J., and M. Wilson. 1996. Formulating the Rasch model as a mixed coefficients multinomial logit. In *Objective measurement III: Theory into practice*, eds. G. Engelhard and M. Wilson. Norwood, NJ: Ablex.

Black, P. and D. Wiliam (1998). Inside the Black Box: Raising Standards through Classroom Assessment. Phi Delta Kappan 80(2): 139-148.

Claesgens, J., K. Scalise, K., Draney, M. Wilson, and A. Stacy. 2002. *Perspectives of chemists: A framework to promote conceptual understanding of chemistry.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Dweck, C. S. and E. L. Leggett (1988). *A social-cognitive approach to motivation and personality.* Psychological Review 95: 256-273.

Hesse, J. (1989). *From Naive to Knowledgeable.* The Science Teacher: 55-58.

Masters, G. N., R. A. Adams, and M. Wilson. 1990. Charting student progress. In *International encyclopedia of education: Research and studies. Supplementary volume 2,* eds. T. Husen and T. N. Postlethwaite, 628–34. Oxford: Pergamon Press.

Minstrell, J. 1998. *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 (Oct.).

Pellegrino, J., Chudowsky, N., Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. N. R. C. Center for Education. Washington, D.C., National Academy Press

Wilson, M. 1990. Measurement of developmental levels. In *International encyclopedia of education: Research and studies. Supplementary volume,* eds. T. Husen and T. N. Postlethwaite. Oxford: Pergamon Press.

Wilson, M. (in press). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wilson, M., C. Kennedy, & K. Draney. 2004. GradeMap (Version 4.0) [computer program]. Berkeley: University of California, BEAR Center.

Wilson, M., and K. Sloane. 2000. From principles to practice: An embedded assessment system. *Applied Measurement in Education* 13(2): 181–208.

Wolf, D., J. Bixby, J. Glenn III, and H. Gardner.1991. To use their minds well: Investigating new forms of student assessment. *Review of Research in Education* 17: 31–74.