Higher Education (2006) 52: 635–663
DOI 10.1007/s10734-004-7263-y

Assessment to improve learning in higher education: The BEAR Assessment System

MARK WILSON & KATHLEEN SCALISE
University of California, Berkeley, USA

Abstract. This paper discusses how assessment practices in higher education can improve or hinder learning. An example is given to illustrate some common educational practices that may be contributing to underpreparation and underperformance of students. 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.

Keywords: assessment, BEAR Assessment System, chemistry education, diagnostic assessment, feedback, feed forward, formative, higher education, learning progressions, learning trajectories, metacognition, progress variables, science education.

Introduction

One of the deepest mysteries in 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 the new coursework – while really they do not 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. "This is best illustrated by quoting a colleague," Hesse said, "who stated that, in his opinion, 90% of student mistakes could be attributed to a lack of study on their part. Blame the student! Period! End of discussion."

Assumptions of lack of studying or insufficient engagement with the material are common explanations of student underperformance. Views of fixed intelligence also are common (Dweck and 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 consider a set of important issues that we will address 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 et al. 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 and 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 their 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 and Wiliam 1998, p. 143).

An example: in his own words

To illustrate potential metacognitive concerns when such conditions are not met, below we draw from a student web-log (blog) describing his experience in a large computer science class at a major US university. We will give selected passages from his blog over the period of the semester and comment on their relevance to our perspective. This transcript was not selected because this student's experience was unique – on the contrary, he was selected as being typical of many students in large and even smaller lecture classes in higher education. Excerpts drawn from his log are annotated for analysis. As background on the student, he is a skilled programmer, working on his campus as a computer consultant, and at the time he wrote his blog was a graduating senior with a solid GPA. The course he is taking is an upper division computer science course, typically oversubscribed at the beginning of each semester and taught by skilled senior faculty members, experts in their field.

Week 1.

> "Went to my first CS 173[1] discussion today. Went over what we are going to cover in the class. Sounds like some cool stuff… The only thing bugging me right now is that I am not officially enrolled in CS 173 yet. I am fourth on the waitlist."

Commentary: The student is engaged and interested in the course, which is not required for him. He perseveres three weeks on the waiting list before being admitted.

Week 2.

> "Had CS 173 today for the first time… Bryan and I have decided to each write our own Blackjack AI and to later pit them against each other online… I think it would be cool to use a genetic algorithm."

Commentary: He anticipates the programming activities in which he will engage.

Week 3.

> "Today I was not as productive as I had hoped... Did my CS reading. I finally got down to programming and got through the last two out of the three problems without any major issues. The first one, though, was giving me a huge amount of problems. I just couldn't figure out how to do the damn thing."

Commentary: First signs of a problem begin to appear. The homework projects are assigned about every two weeks and require perhaps 14–16 hours of effort per assignment. Many begin with an "easy" activity and build on this. However, many students incorrectly solve the initial problem and then propagate that mistake. This is an example of what Pellegrino et al. describe above as: "practicing incorrect skills with little or no feedback" (Pellegrino et al. 2001, p. 7).

Week 4.

> "Damn homework... I went to the lab to work on it. Everyone shortly there after came in and started working on their CS 173 homework. So I ended up staying around to help everyone the best I could since I was the only one to have finished the homework... I have no clue if it is correct, though."

Commentary: In his final comment in this week's blog, "I have no clue if it is correct," the student recognizes that the metacognitive information he needs to monitor his own learning is missing. Note that in this course, most student work was reviewed only by "readers", typically undergraduates who awarded a score and sometimes a brief comment (not more than one line per student per assignment), with no papers returned. Readers were instructed to be generous with credit for effort even on incorrect work. However, students were not made aware of the "effort" grading policy and considered high scores to mean correct answers. Typically, readers had no contact with the instructor and teaching assistants, meeting with the Head Teaching Assistant once per week but only to exchange papers and receive assignments.

Week 8.

> "Almost all of Monday was spent in the [computer lab] working on CS 173 homework again. What made this severely frustrating is that I was

> unable to solve the problem that I spent all day on; no one I know was able to solve that problem... Severely frustrated, I went on home."

Commentary: About halfway through the course, this student, while a talented programmer, begins to express rising levels of frustration. The instructor assumes students review model homework solutions after each assignment is completed and learn from these, but the solutions, written by the teaching assistants, who are graduate students specializing in this field, are "expert" solutions, beyond the grasp of many of the students and drawing on material beyond the scope of the class – a typical novice-expert learning issue (Pellegrino et al., 2001, pp. 72–77). Furthermore, homework problems often can be solved by multiple paths, but the model answers typically show only one. Finally, students receive high scores on the homework and many feel this indicates they do not need to review model solutions.

Week 9.

> "Printing out the CS 173 lecture notes (over 100 pages with 6 Powerpoint slides per page!)."

Commentary: A great deal of material is covered in this course very quickly. For instance, advanced probability models such as Bayesian estimation and hidden Markov models, which might deserve a course to themselves, are here covered in about 4 hours of lecture. As there is no probability or statistics prerequisite for the course, probability theory is covered in a single lecture and students who have not had statistics are at a great disadvantage for the rest of the course. This indicates that there is at least one separate dimension – understanding of probability theory – that needs to be assessed and scaffolded.

> "Went to the CS 173 review. Pretty much useless."

Commentary: There is a growing divergence between what students know in this class and what the teaching assistants think the students know. Students often privately grumble they find reviews and study sections useless, but opt to stop coming or sit quietly rather than ask questions.

> "I will keep this brief: that CS midterm required 120 minutes at least to fully do the exam; we had 80. Haven't heard that many people swearing after an exam in quite a long time."

Commentary: The exam presumes automaticity with material – many students are not at this level.

> “Midterm grade: 63/100 . The mean[2] was 55.5. Standard deviation was around 18. Would have liked mean + SD, but I will live. Still beat the mean which is what is really important.”

Commentary: The average student in the class misses nearly half the items on the midterm exam, many others miss more. The teaching assistants report that they are disappointed in student performance and the head teaching assistant concludes that the students “just didn’t study enough.” Our student focuses on “beating the mean” as his measure of success, rather than mastering the material. This is perhaps the only response he can have, as only normative feedback is given to him (i.e., his only point of comparison for his own work is the class mean).

Week 11.

> “I went into the [computer lab] and sat my ass down to study... I basically ended up printing out the homework solutions and going over them. They turned out to be of little help... Oh well. At least one of my solutions that I cared about turned out to be semantically[3] correct.”

Commentary: Our student selects the model homework solutions as a good study choice. However, as previously discussed, these model solutions often don’t help because the student can’t understand them and they do not trace out his problem solving path.

> “Much ado about crappy homework. After getting my very lame lab checked off today I went to the [computer lab]... to see if anyone needed help with CS 173 and to find out how I messed up on my homework.”

Commentary: Our student is frustrated that his effort isn’t paying off, and begins to turn to sarcasm and attacks on the assignments: “crappy homework,” “lame lab.” Some students in the course try to improve by doing “extra” homework assignments using text questions at the end of each chapter in the book – but no answers are available for these questions, so again, the students cannot evaluate the accuracy of their work.

Week 12.

> “I am actually slated to get an A in CS 154B.[4] CS 173 is completely in the air. If they do straight point conversion, I am slated to get a C+/B−. It will really depend on the final since it is worth 40% (too much in my book).”

Commentary: our student is showing success in another computer science class. But, he doesn't know where he stands in this class, or what the measures of success will be.

> "I finished reading my CS 173 reading in preparation for doing the homework, but I have yet to start it. The usual crap of having to figure out exactly what needs to be done has hit and now I am flipping out since it is not blatantly obvious. But I am sure it is just like all the other CS 173 homework and I am just complicating things in my head. Just need to take a step back and look at it anew."

Commentary: Our student "pep talks" to himself to convince himself that his difficulties are just in his head.

> "Went to FSC to work on CS 173. And kept working. And kept working until they closed at 2:00 a.m. Then went back to the lab with Andrew to finish the sucker. I ended up giving up on part of it."

Commentary: Our student appears to be challenged beyond mastery level but continues his efforts.

Week 14.

> "Went to CS 173… Covering vision and was interesting."

Commentary: Even at this point, our student still finds the course material interesting. The instructor is a top expert and a dynamic speaker. He develops detailed powerpoint slides and uses a text developed in-house. But class attendance is seriously slumping. At the beginning of the semester, even the aisles in the large lecture hall where the class is taught were filled with standing students. Now there are more seats open than filled.

> "After filling out the CS 173 class review forms (the most negative review for a class that I have ever written), I went to the lab."

Commentary: Here, a stellar instructor, an expert in his field who is presenting high-quality lectures with much support material, receives this student's most negative review ever, even though the course is of strong interest to the student.

Week 15.

> "Lectures for this semester ended today. No more CS 173!"

Commentary: The last lecture was attended by only a handful of students.

Week 16.

> "You know what, I love programming. Programming is my form of personal crack. I do it to relax. I do it in my free time. It even affects my reading; I always read computer mags and books. It really makes me think that if I had a choice about losing a hand or a foot, I would go with the foot so that I can still type."

Commentary: Our student continues to program and to dwell on his love of the content area.

Week 17.

> "Went and took the exam. Kind of stupid. Bunch of short answer with some other stuff that was never covered in the homework. I found out afterwards that about a third to half of the test lifted from last year's final. So everyone who had it or had read it knew how to answer those questions perfectly. Of course I had not seen it, let alone had a copy for the test… Needless to say the curve will be skewed."

Commentary: The final exam was perceived as aligned with a prior final, rather than with the homework,. The final exam was subsequently reviewed and indeed it did include many similar problems as a prior exam, which had been made available by the instructor through the computer science honor society web site. Thus, "item contamination" between the two tests becomes a factor. Such item contamination may mask for the instructor how well some students, specifically those previously exposed to the questions, have mastered the material.

Week 18.

> "Found out all of my grades today. For CS 173 I got a B- overall with a 66.67 on the final…. I am just glad that damn class is over."

Commentary: The instructor assigns the grades and our student moves on. He remains fully engaged in computer science and cannot be considered a low potential student. Shortly after the end of the course, he had examples of his code published in computer books and accepted into standard libraries, and subsequently went on to get a successful job in programming.

Elements of effective learning environments

These transcripts from a student web-log illustrate a common problem in educational settings. Instructors go to great efforts to design effective learning experiences, paying careful attention to providing interesting, well-informed lectures, readings, and other aspects of the learning experience, but neglect to implement effective formative assessment practices to support metacognition, relegating assessment to rank-ordering of students for the purposes of grading rather than using it to scaffold learning.

This can be a key problem for students. According to the National Research Council report "How People Learn" (Bransford et al. 2000), timely feedback and revision, on activities congruent with learning goals, is "extremely important" for developing adaptive expertise, learning, transfer and development.

"In many classrooms," according to this report (pp. 140–141), "opportunities for feedback appear to occur relatively infrequently. Most teacher feedback – grades on tests, papers, worksheets, homework ... – represents summative assessments that are intended to measure the results of learning. After receiving grades, students typically move on to a new topic and work for another set of grades. (But) feedback is most valuable when students have the opportunity to use it to revise their thinking as they are working."

This point gives a somewhat ironic perspective on the particular course discussed above, as it examined computer-based artificial intelligence learning algorithms that use "training data" to assess and "learn" whether the response to a given task was adequate, and to adjust or "tune" subsequent performance based on feedback. In some senses, the instructor for this course was teaching his students about an artificial intelligence version of metacognition – while not fully considering the "training data" needs of his own students to also receive feedback.

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 (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 et al. 1991), but this is seldom made available.

Approaches exist that might facilitate metacognition in large lecture classes. In this particular course, for example, the flow of "feed forward" – information to instructors about how their students are doing – came exclusively to the readers, rarely reaching the instructor or teaching assistants who might have used the information to adjust and strengthen the course. The flow of "feedback" was appropriately directed to the individual students – but it was a mere trickle, a single grade and at most a line of commentary on the whole homework that had required the student to expend many hours. Furthermore, the feedback signal was highly "noisy" since accuracy and effort were confounded. Moreover, a further source of potential feedback, the model solutions, were constructed as "expert" solutions, uninterpretable to many of the "novice" students. Fixes here might include (a) having the instructor and teaching assistants participate in sufficient grading for effective feed forward, (b) having readers more extensively mark and return papers to increase the feedback flow, and (c) offering model solutions drawn from student answers (including perhaps faulty solutions, so noted), which would be pitched at a more appropriate level of discourse, as well as expert-constructed solutions. Also, rather than grading for "effort," (d) students with incorrect solutions could be offered increased credit on homework if they revise their work, encouraging them to follow-up on their mistakes and thus giving them an "effort" boost to their grades when their effort has improved their performance.

Suggestions such as these can go a long way toward helping out in specific situations, but a true solution requires a more comprehensive and robust approach. An "embedded assessment" system designed and used in assessment development at the University of California, Berkeley, called the BEAR Assessment System (BAS; Wilson and 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 system was named for its origin at the Berkeley Evaluation and Assessment Research (BEAR) Center and 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 word 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 framework is used to generate interpretable, valid and reliable diagnostic information.

Embedded assessment is desirable because when a task is also a learning activity, it does not take time away from instruction, *and* the number of tasks can be increased to improve measurement, diagnostics, and accountability (Linn and Baker 1996).

The potential usefulness of embedded assessments can be greatly enhanced when the framework on which they are based is consistent with that for the more formal assessments used in accountability assessments, such as campus, school district or state assessments. This potentially enhances the value of formal assessments (for a discussion of this point, under the topic of "assessment nets," see Wilson and Adams 1996).

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.

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;*
- ways to make *inferences about student competence* for the *particular context of use*.

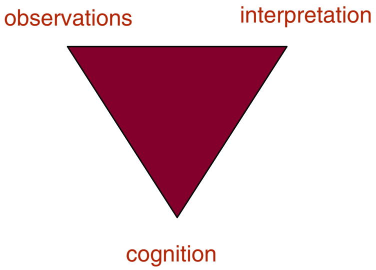


Figure 1. The KWSK assessment triangle.

These elements are of course inextricably linked, and reflect similar concerns as addressed in the conception of constructive alignment (Biggs 1999), regarding the desirability of achieving goodness-of-fit among learning outcomes, instructional approach and assessment.

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 (2005) 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

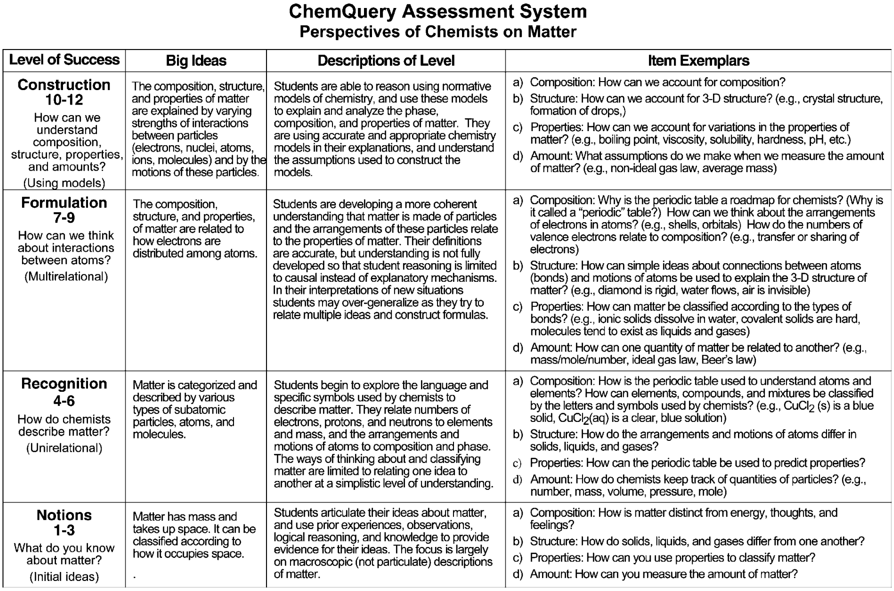
A "developmental perspective" regarding student learning means assessing the development of student understanding of particular concepts and skills over time, as opposed to, for instance, making a single measurement at some final or supposedly significant time point. Criteria for developmental perspectives have been challenging goals for educators for many years. What to assess and how to assess it, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of a variety of teaching and learning theories all impact what approaches might best inform developmental assessment. From Bruner's nine tenets of hermeneutic learning (Bruner 1996) to considerations of Empirical, Constructivist and Sociocultural schools of thought (Olson and Torrance, 1996) to the recent report, "How People Learn" (Bransford et al. 2000), broad sweeps of what might be considered in a developmental perspective have been posited and discussed. Cognitive taxonomies such as Bloom's Taxonomy of Educational Objectives (Bloom 1956), Haladyna's Cognitive Operations Dimensions (Haladyna 1994) and the Structure of the Observed Learning Outcome (SOLO) Taxonomy (Biggs and Collis 1982) are among many attempts to concretely identify generalizable frameworks. One issue is that as learning situations vary, and their goals and philosophical underpinnings take different forms, a "one-size-fits-all" development assessment approach rarely satisfies course needs. Much of the strength of the BEAR Assessment System comes in providing tools to model *many different kinds of learning theories and learning domains*. What is to be measured and how it is to be valued in each BEAR assessment application is drawn from the expertise and learning theories of the instructors and/or course developers, who address the developmental perspective of their applications by specifying a set of "progress variables" (Masters et al. 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 and learning theories are the focus of interest and thus assessment. A key point, however, is that such theoretical structures of developmental learning specified by experts are not accepted *a priori* but are subjected to rigorous comparison with empirical data on actual student learning patterns in the course or courses of interest, which can help support the theoretical learning structure specified or show where it might be improved. Revised frameworks can then be compared again with empirical data, in an iterative approach to refinement that moves between theory and practice.

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.

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 has 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. Three variables, or strands, have been designed to describe chemistry views regarding three "big ideas" in the discipline: Matter, change, and energy. 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. *Energy* 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 accounting for existence of atoms and molecule, and then builds in sophistication beyond that.

Of course, the rich content of the field of chemistry is not so simple as to be easily described in a few variables, and even in capturing the "big ideas" there are multiple ways this could be approached. The Perspectives framework was designed as one potential way, among the many that might exist. The Perspectives variables were derived, as discussed above, by the course instructors and content experts involved with the project considering their learning objectives and learning theories regarding their field, building a theoretical framework, and then testing



ChemQuery Assessment System
Perspectives of Chemists on Matter

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Construction **10-12** How can we understand 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)
Formulation **7-9** How can we think about interactions between atoms? (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 called 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 of electrons) b) Structure: How can simple ideas about connections between atoms (bonds) and motions of atoms be used to explain the 3-D structure of matter? (e.g., diamond is rigid, water flows, air is invisible) c) Properties: How can matter be classified according to the types of bonds? (e.g., ionic solids dissolve in water, covalent solids are hard, molecules tend to exist as liquids and gases) d) Amount: How can one quantity of matter be related to another? (e.g., mass/mole/number, ideal gas law, Beer's law)
Recognition **4-6** How do chemists describe matter? (Unirelational)	Matter is categorized and described by various types of subatomic particles, atoms, and molecules.	Students begin to explore the language and specific 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. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding.	a) Composition: How is the periodic table used to understand atoms and elements? How can elements, compounds, and mixtures be classified by the letters and symbols used by chemists? (e.g., $CuCl_2$ (s) is a blue solid, $CuCl_2$(aq) is a clear, blue solution) 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)
Notions **1-3** What do you know about matter? (Initial ideas)	Matter has mass and takes up space. It can be classified according to how it occupies space. .	Students articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas. The focus is largely on macroscopic (not particulate) descriptions of matter.	a) Composition: How is matter distinct from energy, thoughts, and 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. ChemQuery Assessment System perspective of chemists of matter.

and iteratively improving this framework with empirical studies of actual student work in the field (Claesgens et al. 2002; Scalise et al. 2004). This chemistry framework currently has been used to assess chemistry students in five course contexts, over a range of levels. These include university students in the San Francisco Bay Area at the completion of second-year organic chemistry at the university level, first-year general chemistry at the university level, and first-semester inorganic chemistry, and US secondary students (high school, ages about 14–17) at the beginning, middle and end of first-year studies. Offline paper-and-pencil instruments have been used in a variety of constructed and selected response formats, and an adaptive computer interface, BEAR CAT, has been developed.

Our assessments with pilot studies of this variable show that a student's atomic views of matter begin with having no atomic view at all, but simply the ability to describe some characteristics of matter, such as differentiating between a gas and a solid on the basis of real-world knowledge of boiling solutions such as might be encountered in food preparation, for instance, or bringing logic and patterning skills to bear on a question of why a salt dissolves. This then became the lowest level of the Matter variable, "Notions."

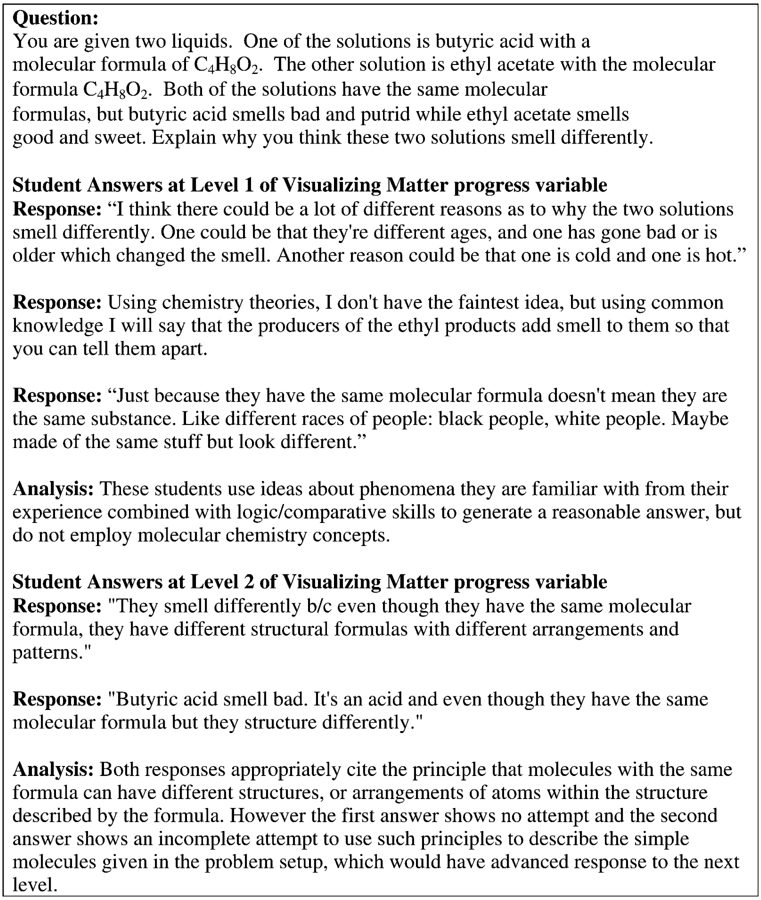
At this most novice level of sophistication, students employ no accurate molecular models of chemistry, but a progression in sophistication can be seen from (i) those unable or unwilling to make any relevant observation at all during an assessment task on matter, to (ii) those who can make an observation and then follow it with logical reasoning, to (iii) those who can extend this reasoning in an attempt to employ actual chemistry knowledge (although they will typically be done incorrectly at first attempts). All these behaviors fall into Level 1, the Notions level, and sublevels are assigned incremental 1–3 scores, which for simplicity of presentation are not shown in detail in this version of the framework.

When students begin to make the transition to accurately using simple molecular chemistry concepts, what we call the "Recognition" level begins. It also has three subscore levels, 4–6, that represent increasingly sophisticated "Recognition" answers. Across the Recognition levels of the Matter progress variable, we see students using very one-dimensional models of chemistry: A simple representation, or a single definition, will be used broadly to account for and interpret chemical phenomena. Students show little ability to combine these ideas.

An example of a ChemQuery assessment prompt and actual student answers at Levels 1 and 2 is shown in Figure 3, along with interpretation. Note that this example item is a "partial credit" item, and spans multiple levels of measurement with the awarding of varying degrees of credit. BEAR assessment items can take many formats and can be designed to span multiple levels (polytomous) or can act as "quick check" items that measure at one cut score (dichotomous), depending on the desires of the course instructors and developers.

When students can begin to combine and relate patterns to account for, for instance, the contribution of valence electrons and molecular geometry to dissolving, they are considered to have moved to the next framework level, Formulating (7–9). Remaining levels of the framework, Construction and Generation, represent further extensions and refinements and are not expected to be mastered at the introductory undergraduate levels, so are not addressed here.

This example shows how a progress variable can generate information on student mastery. Creating the developmental progress variables is not a trivial task; ChemQuery has been working on this for 2 years, at this point. But having succeeded in adapting this approach to a given curriculum, the instructor will be well situated to address many of the issues raised in the first part of this paper. This



Question:
You are given two liquids. One of the solutions is butyric acid with a molecular formula of $C_4H_8O_2$. The other solution is ethyl acetate with the molecular formula $C_4H_8O_2$. Both of the solutions have the same molecular formulas, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why you think these two solutions smell differently.

Student Answers at Level 1 of Visualizing Matter progress variable
Response: "I think there could be a lot of different reasons as to why the two solutions smell differently. One could be that they're different ages, and one has gone bad or is older which changed the smell. Another reason could be that one is cold and one is hot."

Response: Using chemistry theories, I don't have the faintest idea, but using common knowledge I will say that the producers of the ethyl products add smell to them so that you can tell them apart.

Response: "Just because they have the same molecular formula doesn't mean they are the same substance. Like different races of people: black people, white people. Maybe made of the same stuff but look different."

Analysis: These students use ideas about phenomena they are familiar with from their experience combined with logic/comparative skills to generate a reasonable answer, but do not employ molecular chemistry concepts.

Student Answers at Level 2 of Visualizing Matter progress variable
Response: "They smell differently b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns."

Response: "Butyric acid smell bad. It's an acid and even though they have the same molecular formula but they structure differently."

Analysis: Both responses appropriately cite the principle that molecules with the same formula can have different structures, or arrangements of atoms within the structure described by the formula. However the first answer shows no attempt and the second answer shows an incomplete attempt to use such principles to describe the simple molecules given in the problem setup, which would have advanced response to the next level.

Figure 3. To match instruction and assessment, this LBC assessment question followed a laboratory project in which students explored chemicals that had different smells.

approach assumes a match between instruction and assessment, which we address next.

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.

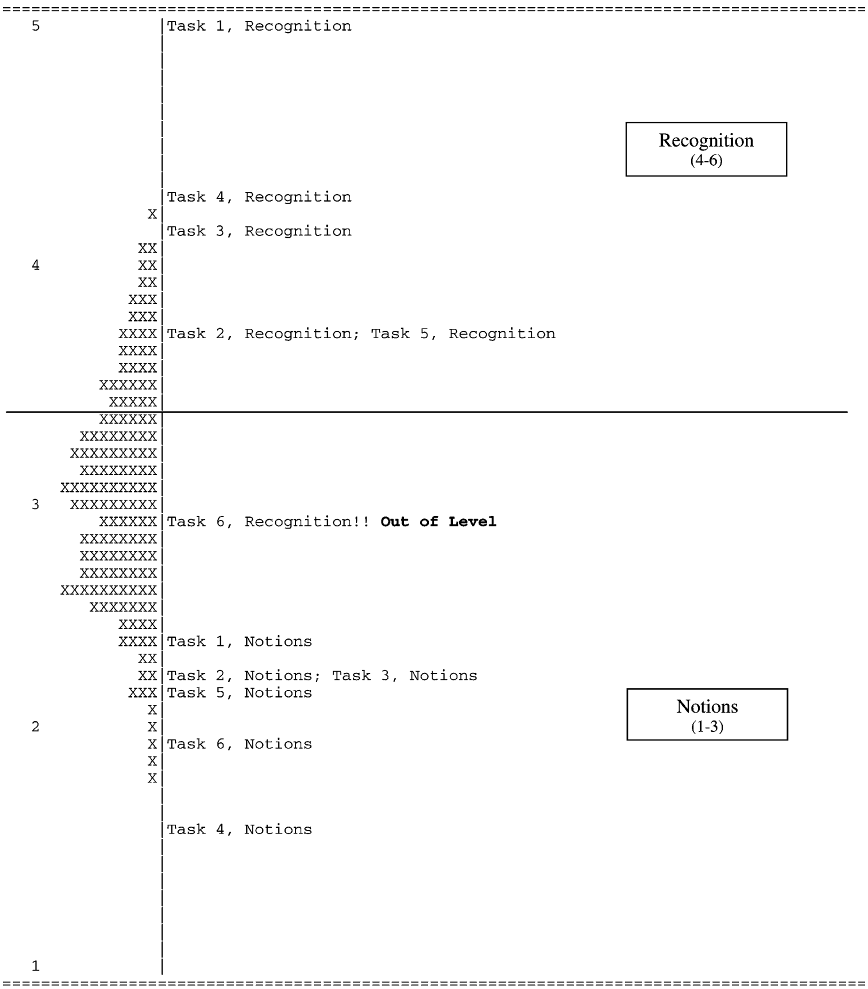
The variety of assessment tasks used by the BEAR Assessment System can range widely, including individual and group "challenges," data interpretation questions, and tasks involving student reading, laboratory, or interactive exercises. In ChemQuery tasks, all assessment prompts are open-ended, requiring students to fully explain their responses. For the vast majority of assessment tasks, the student responses are in a written format.[5]

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: Quality evidence

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure along a progress variable as described above, or even to develop a reasonable framework that can be supported by evidence. 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 uniform 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 measurement models now well-developed enough for use in classroom-based 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 4. Maps such as this can be very technical to consider, and usually require some training for correct interpretation (Wilson and Sloane, 2000). At a glance, the map on the left side shows the measured distribution of students who responded to the Matter items in 2001–2002 trials, and on the right side shows the measured 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



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

Figure 4. Statistical map of the Matter variable, generated from empirical data. Item response maps such as this can be very technical to consider, and usually require some training for correct interpretation (Wilson and Sloane 2000). At a glance, the X's on the left side of the map show the distribution of students in a course, with one X for each student (the array of X's can be viewed as a histogram on its side). The right side of the map shows the measured difficulty of the tasks, with easier tasks at the bottom of the map and harder tasks up higher on the Perspectives scale. 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. Here, six tasks specified by experts as measuring in the Notions level do come in at that difficulty when calibrated with student data. However one of the "Recognition" tasks proves too easy (Task 6, Recognition) and falls below the cut-off for the transition from Notions to Recognition. This task requires revision or reconsideration as to level.

data. Tables of reliability coefficients and standard errors are generated, and inter-rater comparisons also can be made.

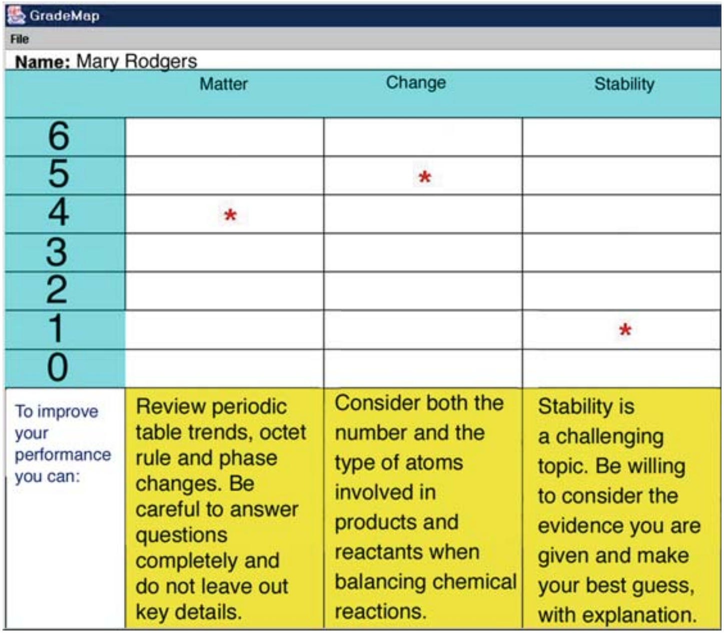
The formal nature of these models and their flexibility allows one to address technical challenges inherent in the classroom assessment situation, such as the maintenance of instructor rating consistency and the maintenance of a meaningful scale throughout the school year. This puts richer information 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. Figure 5 shows where one student stands on several progress variables. Maps are derived from empirical analyses of student data collected from coursework, and are available in many formats.[6]

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. 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.

Principle 4: 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, using web-based interfaces with real-time delivery of instructional material and feedback, or more traditional machine-readable answer sheets.



Name: Mary Rodgers

	Matter	Change	Stability
6			
5		*	
4	*		
3			
2			
1			*
0			
To improve your performance you can:	Review periodic table trends, octet rule and phase changes. Be careful to answer questions completely and do not leave out key details.	Consider both the number and the type of atoms involved in products and reactants when balancing chemical reactions.	Stability is a challenging topic. Be willing to consider the evidence you are given and make your best guess, with explanation.

Figure 5. This is a "conference map" for a single student. It helps the student know where he/she measures on each progress variable and makes suggestions for how the student can improve. Many other maps are also available.

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.

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, and these are indicated in the Assessment Blueprints. Instructors can use these blueprints to review and plan assessment tasks relating to each variable, and to modify the assessments to their own needs.

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 the complexity of the classroom and desired linkages among curriculum, instruction, and assessment.

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 similarly 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 the moderation process is perhaps the strongest influence of moderation on instruction.

Moderation also can take place with student groups, so students can better grasp and determine for themselves 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.)

How the BEAR Assessment System addresses the earlier student example

This paper began with a short case study of student learning issues in a large lecture class, where much attention was paid to supporting the learning environment in many ways but the neglect of sufficient metacognitive prompts interfered with student learning. In this particular course, some of the concerns included insufficient feedback to students on measures on which their performance would ultimately be judged; a feedback signal that was highly "noisy" since accuracy and effort were confounded in the grade; directing the flow of "feed forward" to readers rather than instructors or teaching assistants who could intervene in the learning process; and model solutions that were less useful to students because they were pitched at an expert level of discourse.

To place these concerns in a larger context, it can be seen that the important aspects of effective formative assessment as described by Black and Wiliam (1998, p. 143) – that students understand the measures on which they will be judged, where they stand on these measures, and how they can improve (Black & Wiliam 1998, p. 143) – were not met in this course.

Implementation of the BEAR Assessment System provides tools to help address these problems. Description of the student learning and cognition in this area of computer science through progress variables that describe expected learning patterns can (a) help students understand the measures on which they would be judged. Examining the assessment tasks to ensure they are aligned with learning goals and desired interpretations, and scoring students according to progress variables can help (b) eliminate the "effort" noise in the metacognitive signal and given students a much clearer idea of where they stood on these key measures. (Credit for effort can still be awarded by allowing students to revise assignments after correct solutions are discussed, strengthening follow-up reinforcement.) Even if readers continued to assess the bulk of the assignments, teaching assistants and instructors can have access to a much clearer picture of student progress, individually and in the class aggregate, through (c) progress maps based on quality data and inferences generated via the BEAR assessment system. Additionally, encouraging teaching assistants and instructors to evaluate some student work together and to discuss their scores in moderation sessions would bring to bear (d) a much strengthened feed forward signal for the instructor to tailor future lectures and help teaching assistants examine their plans for discussion and review sessions (and extending this to student moderation also might have a strong effect). Finally, (e) exemplary student answers at different levels of the progress variables, which could be selected in the moderation process, could be useful supplements to expert solutions.

Conclusion

It seems clear that going beyond grades to map individual trajectories of learning is feasible, especially as computers and data collection devices are readily available in higher education and as instructional materials are restructured to use these tools to better accommodate improved formative assessment. In practice, the benefit to students is promising. Wilson and Sloane (2000) have documented evidence of the large and very statistically significant effects that use of the system can have on student performance. Furthermore, it is fascinating 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.

Notes

1. Not the real course number.
2. Out of 100.
3. Semantically correct here means symbols and operations right, even if not correctly assembled according to the syntax, or rules governing construction, in this language.
4. Another course, again with a fictional course number.
5. 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.
6. These maps were drawn in GradeMap, a software package developed at the University of California, Berkeley (Wilson et al. 2004). The analyses for these maps were performed using the ConQuest software (Wu et al. 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).

References

Adams, R.J., and Wilson, M. (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 Wilson, M. (1996). 'Formulating the Rasch model as a mixed coefficients multinomial logit,' in Engelhard, G. and Wilson, M. (eds.), *Objective Measurement III: Theory into Practice* Norwood, NJ: Ablex.

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

Berry, B. (1997). 'New ways of testing and grading can help students learn and teachers teach, *Reforming Middle Schools and School Systems* 1(2). *http://www.middleweb.com/CSLB2testing.html.*

Biggs, J. (1999). *Teaching for Quality Learning at University*. Buckingham: SRHE and Open University Press.

Biggs, J.B. and Collis, K.F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy*. New York: Academic Press.

Black, P. and Wiliam, D. (1998). 'Inside the black box: Raising standards through classroom assessment', *Phi Delta Kappan* 80(2), 139–148.

Bloom, B.S. (ed.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I, Cognitive Domain*. New York, Toronto: Longmans, Green.

Bransford, J.D., Brown, A.L. and Cocking (2000). *The Design of Learning Environments: Assessment-Centered Environments. How People Learn: Brain, Mind, Experience, and School.* Washington, DC, National Academy Press, pp. 131–154.

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

Bruner, J. (1996). *The Culture of Education*. Cambridge, Mass: Harvard University Press.

Claesgens, J., Scalise, K. Draney, K., Wilson, M. and Stacy, A. (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.

Cole, N. (1991). 'The impact of science assessment on classroom practice', In Kulm, G. and Malcom, S. (eds.), *Science Assessment in the Service of Reform,* Washington, DC: American Association for the Advancement of Science, pp. 97–106.

Draney, K.D. and Peres, D. (1998). *Unidimensional and multidimensional modeling of complex science assessment data*. BEAR Research Report SA-98-1. Berkeley: University of California.

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

Haladyna, T.M. (1994). 'Cognitive taxonomies', In *Developing and Validating Multiple-Choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, (pp. 104–110).

Haney, W. (1991). 'We must take care: Fitting assessments to functions', in Perrone V. (eds.) '*Expanding Student Assessment,* Alexandria, VA: Association for Supervision and Curriculum Development, pp. 142–163.

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

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

Linn, R. and Baker, E. (1996). 'Can performance-based student assessments be psychometrically sound?', in Baron, J.B. and Wolf, D.P. *Performance-Based Student*

Assessment: Challenges and Possibilities. Ninety-fifth yearbook of the National Society for the Study of Education, Chicago: University of Chicago Press, pp. 84–103.

Masters, G.N., Adams, R.A. and Wilson, M. (1990). 'Charting student progress', in Husen, T. and Postlethwaite, T.N. Oxford: Pergamon Press, International Encyclopedia of Education: Research and Studies. Supplementary vol. 2, pp. 628–634.

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 (October).

Olson, D.R. and Torrance, N. (eds.) (1996). *Handbook of Education and Human Development: New Models of Learning, Teaching and Schooling*. Oxford: Blackwell.

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, DC, National Academy Press.

Resnick, L.B. and Resnick, D.P. (1992). 'Assessing the thinking curriculum: New tools for educational reform', in Gifford, B.R. and O'Connor, M.C. (eds.), *Changing Assessments,* Boston: Kluwer, pp. 37–76.

Scalise, K., Claesgens, J., Krystyniak, R., Mebane, S., Wilson, M. and Stacy, A. (2004). 'Perspectives of Chemists: Tracking conceptual understanding of student learning in chemistry at the secondary and university levels,' *Paper Presented at the Enhancing the Visibility and Credibility of Educational Research*, American Educational Research Association Annual Meeting, San Diego, CA.

SEPUP. (1995). *Issues, Evidence and You: Teacher's Guide*. Berkeley, CA: Lawrence Hall of Science.

Stake, R. (1991). *Advances in Program Evaluation: Volume 1, Part A: Using Assessment Policy to Reform Education*. Greenwich, CT: JAI Press.

Torrance, H. (1995a). 'The role of assessment in educational reform', in Torrance, H. (eds.), *Evaluating Authentic Assessment,* Philadelphia: Open University Press, pp. 144–156.

Torrance, H. (1995b). 'Teacher involvement in new approaches to assessment', in Torrance, H. (ed.), *Evaluating Authentic Assessment,* Philadelphia: Open University Press, pp. 44–56.

Tucker, M. (1991). 'Why assessment is now issue number one', in Kulm, G. and Malcom, S. (eds.), *Science Assessment in the Service of Reform,* Washington, DC: American Association for the Advancement of Science, pp. 3–16.

Wilson, M. (1990). 'Measurement of developmental levels', in Husen, T. and Postlethwaite, T.N. (eds.), *International Encyclopedia of Education: Research and Studies. Supplementary Volume,* Oxford: Pergamon Press.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Assoc.

Wilson, M. and Adams, R.J. (1996). 'Evaluating progress with alternative assessments: A model for Chapter 1, in Kane, M.B. (eds.,) *Implementing Performance Assessment: Promise, Problems and Challenges*, Hillsdale, NJ: Lawrence Erlbaum.

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

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

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

Wu, M., Adams, R.J. and Wilson, M. (1998). ACER*ConQuest* [computer program]. Melbourne, Australia: ACER Press.

Zessoules, R. and Gardner, H. (1991) 'Authentic assessment: Beyond the buzzword and into the classroom', in Perrone, V. (ed.), *Expanding Student Assessment*, Alexandria, VA: Association for Supervision and Curriculum Development, pp. 47–71.