JIunlimited/Stockxpert

Assessment from the Ground Up

MARK WILSON

When assessments are based on teaching and curriculum content, testing can be an essential part of instruction rather than lost time for learning.

Schools are often faced with two choices of assessments: large-scale standardized or small-scale classroom.

Classroom assessments are the ones that teachers create — quizzes, tests, or assignments — to investigate and document student progress in the classroom. Large-scale summative assessments are the ones that policy makers (legislators, school board members, and district administrators) mandate for the purposes of accountability. Most often, these are norm- or criterion-referenced tests used to evaluate student achievement or program effectiveness across schools and districts.

It is a common assumption that large-scale assessments produced by testing companies are more trustworthy than classroom assessments. However, despite substantial progress in test design, standard administrations of norm- or criterion-referenced tests given at the end of a year or unit provide little guidance for classroom instruction. So, in order to make policy decisions, states, districts, and schools are offered one of two unsatisfactory choices: Use standardized assessments that are trusted by the public but pretty useless for making decisions about individual student progress, or use classroom assessments for which the teacher has little formal evidence about validity or reliability yet are informative for making instructional decisions.

Even though classroom and large-scale assessments need not be mutually exclusive tools for monitoring learning, it is extremely rare to find valid and reliable tests that can be linked directly to classroom practices and instructional activities.

The BEAR Essentials

At the University of California, Berkeley, we're helping to change that. Over the last 15 years, our research team at the Berkeley Evaluation and Assessment Research Center (we call it the BEAR Center) in the Graduate School of Education has been developing assessment systems that are both psychometrically sound and instructionally relevant.

Because of these two attributes, BEAR-designed assessments can be used within and across classrooms and even across schools. To say that they're instructionally relevant means they're embedded in the curriculum for which they've been designed to monitor student progress. This means assessments of student progress and performance are integrated into instructional materials and are virtually indistinguishable from day-to-day classroom activities. This offers the potential for great "ecological validity" because they're grounded in everyday teacher practices. We can assess a wider range of skills than institutional assessments. Furthermore, we avoid one-shot testing situations and focus instead on the process of learning and an individual's progress.

The BEAR Assessment System grew out of two early assessment projects our team worked on: 1) California's voluntary Golden State Exam, which debuted in 1985 and which met its demise in 2003 with implementation of the No Child Left Behind legislation, and SEPUP (Science Education for Public Understanding Program), an innovative science curriculum for grades 6–12 developed at UC Berkeley's Lawrence Hall of Science in 1987.

With my Graduate School of Education colleague, Kathryn Sloane, we reviewed the projects and their curricula to identify their core principles. Eventually, we formally described them in a journal article, "From Principles to Practice: An Embedded Assessment System," published in 2000 (also see Wilson 2005).

The Assessment Triangle

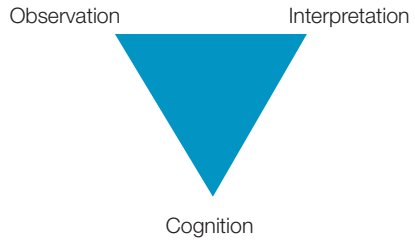
The National Research Council (2001) suggests that good assessment needs to ad-

R&D appears in each issue of *Kappan* with the assistance of the **Deans' Alliance**, which is composed of the deans of the education schools/colleges at the following universities: Harvard University, Michigan State University, Northwestern University, Stanford University, Teachers College Columbia University, University of California Berkeley, University of California Los Angeles, University of Michigan, University of Pennsylvania, and University of Wisconsin.

MARK WILSON, a psychometrician and education researcher, is a professor at University of California, Berkeley's Graduate School of Education and the founding director of the Berkeley Evaluation and Assessment Research (BEAR) Center.

dress the three inextricably linked parts of this triangle:

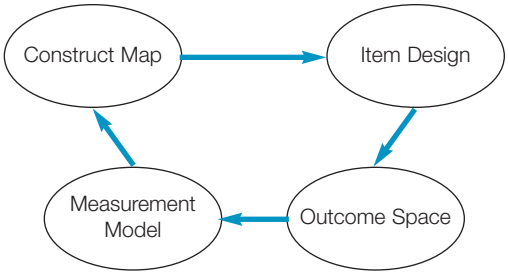
FIG. 1.
The National Research Council's Assessment Triangle



To address these components, the BEAR system employs four principles similar to those outlined by the National Research Council: 1) a developmental perspective on learning; 2) a tight link between instruction and assessment; 3) management by instructors to allow appropriate feedback, feed forward, and following up; and 4) the generation of quality evidence to make inferences (NRC 2001; Wilson 2005). (Note that we split the 3rd vertex, Interpretation, into two building blocks.)

In the BEAR Assessment System, these four principles are expressed as the four building blocks (see Figure 2) we follow to create any assessment: a construct map, an items design, an outcome space, and a model for making inferences about student performance (what we call a measurement model). The process is iterative, which means we're likely to move through all four building blocks several times as we design an assessment.

FIG. 2.
BEAR Assessment System



The result is a comprehensive, integrated system for assessing, interpreting, and monitoring student performance. The BEAR Assessment System provides tools to assess performance on central concepts in a curriculum, set standards, track progress over time, and provide feedback on student progress as well as the effectiveness of a curriculum (Wilson 2005).

The first step is to identify the construct(s) to be assessed, such as science knowledge. A **construct map** helps us outline the important levels in a continuum of developmental learning in the subject area. The construct map is based on a developmental perspective of student learning. Learning is conceptualized not as an acquisition of more knowledge, but as progress within a subject area (NRC 2001). For example, Figure 3 gives a shortened example of a construct map crafted around "Earth in the Solar System." BEAR partnered with WestEd as part of a project designed to conduct research in standards-based science assessment (Briggs et al. 2006; WestEd 2005). According to standards and an underlying research base, 8th-grade students are expected to understand different phenomena in this area, such as the day/night cycle and the phases of the moon. For this project, however, we needed to assess 5th graders as well, and it was important that we use the same continuum to describe the progression of learning for 5th and 8th grades. Consequently, we don't usually expect 5th graders to reach the top level or that many 8th graders will fall into the lowest levels of the continuum.

FIG. 3.
Modified Construct Map for Student Understanding of Earth in the Solar System

Level	Description
5 (8th grade)	Student is able to put the motions of the Earth and Moon into a complete description.
4 (5th grade)	Student is able to coordinate apparent and actual motion of objects in the sky.
3	Student knows that the Earth orbits the Sun, the Moon orbits the Earth, the Earth rotates on its axis; but the student has not put this knowledge together with an understanding of apparent motion to form an explanation. May not recognize that the Earth is both rotating and orbiting simultaneously.
2	Student recognizes that the Sun appears to move across the sky every day, the observational shape of the Moon changes every 28 days.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical.
0	No evidence or off-track.

JUunlimited/Stockxpert

The **items design** is a framework for designing the tasks and questions that will elicit specific kinds of evidence about student learning. The most fundamental element of this design is that the responses to the item can be mapped into the levels of the construct map. We aim to create items that tap into all levels of student knowledge. This is a basic tenet of content validity outlined by the American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education (1999): The items on a test are sampled appropriately from a wide range of student abilities. Traditional testing practices (in high stakes as well as standardized tests) have long been criticized for oversampling items that assess only basic knowledge and ignoring more complex levels of understanding. Matching items with the construct map ensures that we won't fall into that trap and also ensures that what's assessed is what's being taught in the curriculum. (See Figure 2.)

Matching items with the construct map ensures that we won't fall into the trap of oversampling items that assess only basic knowledge and ignore more complex levels of understanding.

The **outcome space** represents in detail the qualitatively different kinds of student responses elicited by the items. The outcome space represents the form of the student responses — that is, multiple choice, short answer, or essay. But it also represents the role of the teacher/scorer in interpreting responses and thus evaluating student learning. Central to the BEAR system is the creation of scoring guides that can be made concrete by including examples of scored student work. This helps teachers "see" progress in action and more deeply understand how to tailor their instruction accordingly. However, the construct map, items design, and outcome space aren't static; they may change once teachers collect empirical evidence (both qualitative and quantitative). Our theory of student learning must be tested, fleshed out, and revised in response to data.

For example, in the Earth in the Solar System project, we designed items to tap into 5th-grade student understanding. These items are "ordered multiple-choice" in that, unlike traditional multiple-choice questions, there is not just one 'right' answer and several wrong answers, but rather the choices are leveled in accordance with our construct map. Some represent more sophisticated understanding of the construct than others. (See Figure 4.)

FIG. 4.
5th-grade Sample Item from Earth in the Solar System

It is most likely colder at night because

A. the Earth is at the furthest point in its orbit around the Sun.	Level 3
B. the Sun has traveled to the other side of the Earth.	Level 2
C. the Sun is below the Earth and the Moon does not emit as much heat as the Sun.	Level 1
D. the place where it is night on Earth is rotated away from the Sun.	Level 4

BEAR uses many types of items (open-ended, traditional multiple choice, ordered multiple choice, portfolios, interviews) to elicit student responses. The design of these items is important, but most essential is the connection between the responses and how they're interpreted in light of the construct map.

Finally, **the measurement model** defines how inferences about student understandings can be drawn from the scores. Here, issues of technical quality are addressed. It's essential that assessments meet standards of fairness (such as consistency and lack of bias). For example, using open-ended scoring guides requires procedures for gathering, managing, and scoring student work. Raters must score the work, and issues of time, cost, fairness, and consistency arise. BEAR conducts pilot studies, trains raters, and examines reliability statistics to ensure that we have accurate estimates of student knowledge (Wilson and Hoskens 2001). We study the use of generalized forms of item response models for our research studies and choose measurement models appropriately (De Boeck and Wilson 2004; NRC 2001).

The Standards for Educational Testing (AERA, APA, and NCME 1999) describe different sources of validity evidence that need to be integrated to form a coherent validity argument. These include evidence based on test content, response process, test structure, relations to other variables, and testing consequences. In any assessment, be it classroom or large-scale, a necessary element of content validity is identifying what student progression

looks like within a curriculum and how that learning is expected to unfold. In the BEAR system, the first three elements of validity evidence are produced during the careful development and revision of a construct map.

Determining where students, classes, and even schools are "located" on the construct map can guide teacher or program planning.

Last but not least, the output from these models can be used to determine where students, classes, and even schools are "located" on the construct map; this information can guide teacher or program planning. This is necessary to ensure that the assessment is useful for instruction (what psychometricians call instructional or consequential validity). BEAR has gathered evidence for the usefulness of this approach by actively working with school districts and teachers (Sipusic 1999; Wilson and Sloane 2000) and has created software to facilitate the use of the BEAR system by teachers (Kennedy and Draney 2009; Kennedy, Wilson, and Draney 2006).

Maximizing the Power of Assessment

In traditional test design, as well as classroom assessments, we often presume that we know a construct implicitly, and we immediately begin writing items to assess it. The BEAR system circles through the assessment process more thoroughly — designing the construct, writing items, scoring the items, examining the scored data qualitatively and quantitatively, and then revisiting our construct, as well as our items. For every assessment task, our system seeks to answer four fundamental measurement questions: 1) What do we want to measure? 2) How are we going to observe it? 3) What sense are we going to make of the responses? and 4) How can we combine the item responses to get valid and reliable measures of student progress?

The power of assessment is in its ability to connect instruction to student learning. The key is to have a reasonably accurate and comprehensive idea of the typical ways in which students grow to understand a subject and to build assessments that can measure that growth. When assessments are built using this approach, testing doesn't have to be lost time for teaching and learning, but rather an essential part of both. K

Learn more about the BEAR Center and its work http://bearcenter.berkeley.edu/

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME). *Standards for Psychological and Educational Tests*. Washington D.C.: AERA, APA, and NCME, 1999.

Briggs, Derek C., Alicia C. Alonzo, Cheryl Schwab, and Mark Wilson. "Diagnostic Assessment with Ordered Multiple-Choice Items." *Educational Assessment* 11, no. 1 (2006): 33-63.

Briggs, Derek C., and Mark Wilson. "Generalizability in Item Response Modeling." *Journal of Educational Measurement* 44, no. 2 (2006): 131-155.

De Boeck, Paul, and Mark Wilson, eds. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York: Springer-Verlag, 2004.

Kennedy, Cathleen A., and Karen Draney. "Mapping Multiple Dimensions of Student Learning: The Construct Map Program." *Journal of Applied Measurement* 10, no. 1 (2009): 22-37.

Kennedy, Cathleen A., Mark Wilson, and Karen Draney. "ConstructMap." Computer program. Berkeley, Calif.: Berkeley Evaluation and Assessment Research Center, University of California, 2006.

National Research Council (NRC). *Knowing What Students Know: The Science and Design of Educational Assessment.* Washington, D.C.: National Academy Press, 2001.

Sipusic, Michael. *Moderation in All Things: A Class Act*. Video A 94720-1670. Berkeley, Calif.: Berkeley Evaluation and Assessment Research Center, Graduate School of Education, University of California, 1999.

WestEd. *Partnership for the Assessment of Science-Based Science: Frequently Asked Questions and Sample Assessment Items*. 2005. www.wested.org/cs/we/view/rs/612.

Wilson, Mark. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.

Wilson Mark, and Machteld Hoskens. "The Rater Bundle Model." *Journal of Educational and Behavioral Statistics* 26, no. 3 (2001): 283-306.

Wilson, Mark, and Kathryn Sloane. "From Principles to Practice: An Embedded Assessment System." *Applied Measurement in Education* 13, no. 2 (2000): 181-208.