

Making Benchmark Testing Work

Joan L. Herman and Eva L. Baker

Six criteria can help educators use benchmark tests to judge student skills and to target areas for improvement.

Many schools are developing assessment systems to monitor their students' progress toward state standards throughout the academic year. Educators in these schools wisely recognize that information from annual state tests is often too little, too late. State tests can be powerful motivators, communicating expectations and focusing curriculum and instruction. But they rarely provide the ongoing information that schools need to guide instructional programs and address the learning problems of students who might otherwise be left behind.

Vendors and service providers have jumped in to fill this gap with a variety of products and services, known by such names as *benchmark tests*, *progress monitoring systems*, and *formative assessments*. These vendor-developed products and locally developed testing systems are designed to coordinate with state standards and assessments and are administered regularly—often quarterly—to gauge student progress. Available options include customized testing programs, off-the-shelf products aligned with existing state tests, and CDs and Web portals containing item banks from which educators can construct their own tests. Services include rapid, automated scoring and elaborate reporting systems for multiple audiences and purposes. Not uncommon, for example, are separate reports for administrators, teachers, parents, and students, providing information on class, group, and individual performance on specific grade-level standards and for overall grade-level proficiency.

Despite the glitz and gee-whiz appeal of such products, information about their effectiveness in improving student learning is generally hard to come by. Yet the quality of the assessment is essential: There is little sense in spending the time and money for elaborate testing systems if the tests do not yield accurate, useful information. If the information is flawed or erroneous, it is unlikely to provide good guidance for instruction or to support better decision making. The whole rationale for conducting the assessment falls apart; it merely creates the illusion that something is being done and people are paying attention.

The *validity*, or quality, of an assessment is derived from an array of evidence showing the extent to which that assessment provides sound information for particular purposes. The purpose of benchmark testing is to provide both accurate information about how well students are progressing toward mastery of standards and useful diagnostic feedback to

guide instruction and improve learning. Here we discuss six criteria that determine the validity of benchmark tests: alignment, diagnostic value, fairness, technical quality, utility, and feasibility (Linn, Baker, & Dunbar, 1991). “Recommendations for Benchmark Tests” summarizes the implications of these criteria for educators.

Alignment

Alignment is the linchpin of standards-based reform. Unless benchmark tests reflect state standards and assessments, their results tell us little about whether students are making adequate progress toward achieving the standards and performing well on the assessment. The term *alignment*, however, has many potential meanings.

Aligning benchmark tests with state standards does *not* mean creating formative tests that mimic the content and format of annual state tests as specifically as possible. Although a strategy of strict test preparation may boost state test scores in the short term, available evidence suggests that early gains achieved in this way are not sustained in the long run (Herman, 2005; Hoff, 2000; Linn, 2000).

Further, aligning benchmark tests too closely with a state's tests gives short shrift to the state's standards. Annual tests of an hour or two's duration cannot address all the curriculum standards that a state has deemed essential knowledge for students. Because educators and students tend to focus on what will be tested, benchmark testing that covers only what appears on the state tests may accelerate curriculum narrowing. In contrast, good benchmark testing can encourage instruction on the full depth and breadth of the standards and give students opportunities to apply their knowledge and skills in a variety of contexts and formats.

For example, knowledge of Newton's laws is included in most states' physics standards. The typical state test may address this knowledge with one or two multiple-choice items or perhaps a short-answer item. In contrast, well-developed benchmark tests use not only multiple-choice and open-ended items but also performance tasks and laboratory experiments to delve deeper into students' understanding of Newton's laws. A test might ask students to explain the underlying principles of force and motion at work in a car crash, for instance, or to design a roller coaster that makes optimal use of physics principles.

Mapping Content

The alternative to aligning benchmark tests with the specific content and format of state assessments is to align them with priority content and performance expectations *implicit* in state standards. Alignment researchers and learning theorists have suggested that in establishing such priorities, educators must define both the major knowledge and skills to be addressed and the expected intellectual level of the performance (Porter, 2002).

The matrix in Figure 1 shows how a school might map expectations for a state's grade 6 mathematics standards. This matrix lays out the substance of standards in terms of the

specific content knowledge that students need to acquire and includes four cognitive levels suggested by Norman Webb (1997): *recall* (knowledge of specific facts, definitions, simple procedures, formulas, and so on); *conceptual understanding* (knowledge of principles and the ability to apply them in relatively routine situations); *problem solving* (the ability to reason, plan, use evidence, and apply abstract thinking in novel situations); and *extended and strategic thinking* (the ability to apply significant conceptual understanding to extended, novel tasks requiring complex reasoning; to make connections among different content areas; and to devise creative solutions).

Figure 1. A Sample Matrix to Map Expectations

Assessment Plan—6th Grade Math Standards Carver School District Standard: <i>Number Sense</i>		
Math Reasoning	<p>1.0</p> <p>Students compare and order positive and negative fractions and decimals. Students solve problems involving fractions, proportions, ratios, and percentages.</p>	1.1 Compare and order positive and negative fractions and decimals them on a number line.
Statistics & Analysis		1.2 Interpret and use ratios in different contexts (e.g., batting averages) show the relative sizes of two quantities.
Measurement & Geometry	<p>2.0</p> <p>Students calculate and solve problems involving addition, subtraction, multiplication, and division.</p>	1.3 Use proportions and cross-multiplication to solve problems.
Algebra & Functions		1.4 Calculate given percentages of quantities and solve problems involving discounts at sales, interest earned, and tips.
Number Sense		2.1 Solve problems involving addition, subtraction, multiplication, and division of positive fractions and explain why a particular operation was used in a given situation.
		2.2 Explain the meaning of multiplication and division of positive fractions and perform the calculations.
		2.3 Solve addition, subtraction, multiplication, and division problems in concrete situations that use positive and negative integers.
		2.4 Determine the least common multiple and the greatest common factor of whole numbers; use them to solve problems with fractions.

Figure 1 provides a starting place for identifying what content the school should teach and assess. But even when educators develop such a matrix, they have not yet resolved an important tension: Like the annual state tests, benchmark tests cannot possibly address all of a state's standards. Imagine a test that included items for every cell in Figure 1, with every standard implying a myriad of important objectives and topics that could be assessed at every cognitive level. Testing time would be endless. Instead, educators need to decide in advance what content is most important to assess, and at what levels.

Focusing on Big Ideas

By incorporating the key principles that underlie state or district standards into benchmark assessments, educators have a reasonable strategy for addressing the breadth of these standards. Cognitive research across many different subject areas suggests the power of focusing on the key principles underlying a content domain rather than on the specific topics within the domain (Ball & Bass, 2001; Carpenter & Franke, 2001; diSessa & Minstrell, 1998; Ericsson, 2002). Encompassing specific topics, the principles help students to organize and use their knowledge. For example, understanding the principle of *equivalence* can help students balance mathematical equations.

Research also demonstrates the power of engaging students in applying and explaining key principles (see, for example, Chi, 2000; VanLehn, 1996). Incorporating this idea into assessments can help increase their learning value. Therefore, despite the ease of scoring multiple-choice items, benchmark tests should employ many different formats to enable students to reveal the depth of their understanding.

For example, the two test items below call for students to give short answers, choose multiple-choice options, offer extended explanations, and draw pictures to demonstrate their understanding of fractions:

Six people are going to share five chocolate bars. Write the fraction that shows how much chocolate each person gets: _____

Then, explain what you did to find this answer. You can draw a picture of the chocolate bars to help explain your answer.

Which of the following fractions is between $2\frac{1}{2}$ and $2\frac{3}{4}$?

- A. $2\frac{1}{4}$
- B. $2\frac{5}{6}$
- C. $2\frac{2}{3}$
- D. $2\frac{1}{3}$

Explain how you found the answer to this problem. Draw a picture that shows your answer is correct.

Diagnostic Value

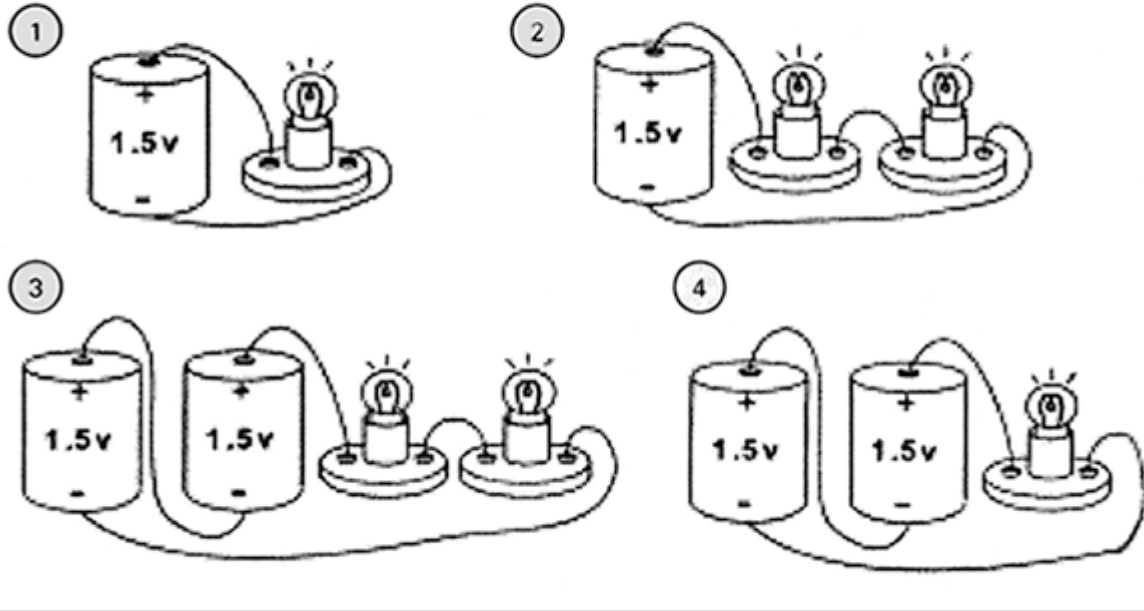
A test has diagnostic value to the extent that it provides useful feedback for instructional planning for individuals and groups. A test with high diagnostic value will tell us not only whether students are performing well but also why students are performing at certain levels and what to do about it.

Open-ended test items that ask students to explain their answers increase the diagnostic value of benchmark tests. Students' responses reveal their thinking, helping teachers to refine their instructional strategies and design targeted instruction for individual students.

Multiple-choice items can also yield important diagnostic information, particularly when they are purposely designed so that distracters—the incorrect answer options—reflect common student misunderstandings. Consider the science test item shown in Figure 2 (p. 53), which is designed to assess students' understanding of electricity principles. Number 4 is the correct choice; it has the highest current because the voltage is the largest and the resistance is the smallest of all the circuits. Students who incorrectly choose number 2 show the weakest understanding of current because the voltage is the smallest and the resistance is the largest of all four circuits. We may infer that students who circle number 1 realize that smaller resistance is often associated with larger current but do not understand the role of voltage; conversely, students who circle number 3 may understand that larger voltage is often associated with larger current but do not understand the importance of resistance. Our inferences would be stronger, of course, if the assessment also asked students to explain their reasoning for their choices.

Figure 2. Sample Test Item, Electricity Principles

Circle the circuit that has the highest current. Assume that the circuits are properly connected, all bulbs are identical, and wiring does not contribute a significant amount of resistance to the circuit.



This test item is intentionally designed so that each incorrect answer suggests a different misconception or common student error (see pp. 51–52).

To provide good diagnostic information on where and how students are experiencing difficulties, benchmark tests must include enough items on each potential topic to render a reliable diagnosis. Drawing inferences from performance on only one or two items or from unreliable subscales may result in faulty conclusions.

Fairness

Fair benchmark tests provide an accurate assessment of diverse subgroups. Biased test items, in contrast, introduce unnecessary complexities that systematically hamper access for particular subgroups. For example, when test items use complex language to assess students' science knowledge, English language learners or poor readers may suffer an unfair disadvantage because they are unable to demonstrate their actual skill in science. Similarly, setting problems in contexts that are less familiar to some subgroups can impede those groups' ability to apply their knowledge and skill. A mathematics problem that asks students to compute the best route for a subway trip may be clear to students from New York City but may confuse students who have never been on a subway, even if they know what it is.

Fairness also is a prime issue in testing students with disabilities. For example, students with specific reading disabilities may need more time to process text. Although details on accommodations are beyond our scope here, two points are worth underscoring: (1)

Accommodations offered to students in benchmark testing should mirror those documented in their individualized education plans and offered to them on annual state tests, and (2) the design of benchmark tests should make these tests accessible for as many students as possible. As with any standardized assessment, benchmark test items should be thoroughly reviewed for possible bias by representatives of diverse communities, as well as tested empirically to identify any items that have aberrant results for particular subgroups.

Technical Quality

Tests with high technical quality provide accurate and reliable information about student performance. Those of us who are not psychometricians tend to zone out when talk turns to technical indices of test quality. But item and test quality provide important information about whether we can trust the scores. If a test is weak on this characteristic, plans and decisions made on the basis of test data are likely to be faulty.

For example, *reliability*—determined by internal consistency, item response theory, inter-rater agreement, and a number of other indices—stands for the consistency of a measure and the extent to which scores on that measure represent some stable and coherent core. When a measure is highly reliable, the items within it operate similarly. Reliability problems arise if a student's performance varies significantly across items, within a short period of time, or under a whole host of other conditions (during the stress of an exam, when the student is tired, when the testing room is uncomfortable, and so on).

Imagine, for instance, a test of archery skill in which hitting the bull's-eye represents high levels of performance. One individual shoots five arrows that all hit the bull's-eye or very close to it. A second individual also shoots five arrows, but they all land in the outer ring of the target. Although the first archer's performance is more accurate, both performances are *reliable*. Looking at the results, we can be confident that we have an accurate measure of each individual's archery skills—at least on that particular day, under those conditions. A third archer's performance, however, is inconsistent, or unreliable: One arrow hits the bull's-eye; one lands near it; another hits the outermost ring; and two miss the target altogether. Assessing that performance, we would find it difficult to judge the archer's skill level, predict performance on the next shot, or devise corrective action—all inferences we want to draw from benchmark testing.

Inter-rater reliability enters the picture for open-ended and performance items—for example, a district writing assessment—which must be scored by human judges. Low inter-rater reliability often means that raters have not been trained well enough to agree on the meaning of high-quality performance. Districts should carefully monitor inter-rater reliability and take action to improve it if needed.

For benchmark tests to have diagnostic value, we must ensure the reliability of the diagnosis. We can easily create subscales that look useful—for example, aggregating the results from four items that appear to measure students' understanding of rational numbers. But if those items do not result in a reliable scale—if a student's performance

on them varies widely—then the results do not provide good information for our instructional decisions.

Reliability and accuracy are necessary but not sufficient prerequisites to *validity* (that is, the extent to which a test accomplishes its intended purposes). A prime purpose of benchmark testing is to show whether students are progressing toward achieving proficiency on state tests; therefore, if the benchmark tests are doing their job, there should be a strong predictive relationship between students' performance on the benchmark tests and students' performance on the state assessments.

Educators should plan to document the reliability and validity of their benchmark tests on an ongoing basis. Good tests aren't magically created by simply assembling test items that seem reasonable—even if the tests are aligned with priority standards and teachers and psychometricians developed them collaboratively. Schools need to pilot-test and revise their items and their test forms on the basis of the technical data, ideally before a test becomes operational. Devoting sufficient time for development will yield better information from a benchmark test in the long run.

Utility

Utility represents the extent to which intended users find the test results meaningful and are able to use them to improve teaching and learning. Benchmark tests with high utility provide information that administrators, teachers, and students can use to monitor student progress and take appropriate action. District administrators, for example, may use the data to identify schools that need immediate help in particular subjects. School principals may use the data to identify students for special after-school tutoring. And teachers may use the information to modify their teaching and to regroup students for supplementary instruction.

To make benchmark tests useful, schools must put the results in intended users' hands quickly and train them to interpret the information correctly. In addition, schools must administer assessments and provide feedback when such guidance can be most useful—that is, around the time when teachers address the test content in classroom instruction. If teachers in different classrooms or schools use different curriculum materials or take more or less time teaching the topic, then finding common testing times may be an issue. For example, if one school covers Newton's laws in the fall and another covers this topic in the spring, a fall benchmark test on the topic will be of little use to the second school. To address the problem, some school districts give teachers flexibility in determining what content to assess during each testing period.

Schools can also increase the effectiveness of benchmark tests by helping teachers use the results. Teachers who lack such support may not know what to do when assessment results show that students are struggling; they may hesitate to go back and reteach because they feel pressure to move on and “cover” the curriculum. Even if they do go back, they may replicate the same strategies that were unsuccessful in the first place.

In addition to giving teachers the data, schools must ensure that they have the pedagogical knowledge and access to alternative materials that they need to bridge identified learning gaps. Some districts and schools help teachers by establishing grade-level or subject-matter teams, including content and curriculum experts, to meet regularly to analyze student work, discuss strengths and weaknesses in learning, and formulate next steps for individual students and subgroups representing various learning needs.

Some of the benchmark testing products available address this need by producing test score reports that explicitly identify student strengths and weaknesses, make suggestions for teaching, and direct teachers to useful materials. However, the value of such approaches depends on the validity of score interpretations and the actual learning benefits of the suggested next steps. Educators considering the purchase of such products should look for evidence to support both of these components.

Feasibility

Benchmark testing should be worth the time and money that schools invest in it. Well-designed benchmark tests can contribute to as well as measure student learning. But if such tests are not well designed, they can waste students' and teachers' valuable time and energy, ultimately detracting from good teaching and meaningful learning.

Of course, to determine whether benchmark testing is worth the effort, educators ultimately need to look at the results. Are benchmark tests focusing attention on student performance and providing solid information on which to base improvement efforts? Are they actually improving student learning? The history of testing is fraught with good intentions that have gone awry. Like state assessments, benchmark tests will fulfill their promise only if we monitor their consequences and continually improve their quality.

Recommendations for Benchmark Tests

1. *Align standards and benchmark assessments from the beginning of test development.* Decide what specific content to assess and at what level of intellectual demand. Include the application of complex learning. To create benchmark tests that enrich student learning opportunities, focus on the big ideas of a content area and counteract curriculum narrowing by designing benchmark tests that allow students to apply their knowledge and skills in a variety of contexts and formats.
2. *Enhance the diagnostic value of assessment results through initial item and test structure design.* Use extended-response items to reveal student

thinking and potential misconceptions. Build distracters into multiple-choice items that reveal common student misunderstandings.

3. *Ensure the fairness of benchmark assessments for all students, including English language learners and students with disabilities.* Avoid unnecessarily complex language or specific contexts that could unfairly confound some students' ability to show what they know.
4. *Insist on data showing tests' technical quality.* Study psychometric indices to determine the reliability of assessments.
5. *Build in utility.* Design reports of test results to be user-friendly and to provide guidance on how to appropriately interpret and use the results.
6. *Hold benchmark testing accountable for meeting its purposes.* Crafting good benchmark tests and ensuring their wise use for improving student learning requires systematic design and continual evaluation.

References

Ball, D. L., & Bass, H. (2001). What mathematical knowledge is entailed in teaching children to reason mathematically? In National Research Council, *Knowing and learning mathematics for teaching: Proceedings of a workshop* (pp. 26–34). Washington, DC: National Academy Press.

Carpenter, T., & Franke, M. (2001). Developing algebraic reasoning in the elementary school. In H. Chick, K. Stacey, J. Vincent, & J. Vincent (Eds.), *Proceedings of the 12th ICMI Study Conference* (Vol. 1, pp. 155–162). Melbourne, Australia: University of Melbourne.

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instruction psychology* (Vol. 5, pp. 161–238). Mahwah, NJ: Erlbaum.

diSessa, A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. Goldman (Eds.), *Thinking practices in learning and teaching science and mathematics* (pp. 155–187). Mahwah, NJ: Erlbaum.

Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education* (pp. 21–55). Hillsdale, NJ: Erlbaum.

Herman, J. (2005). *Making accountability work to improve student learning*. (CSE Technical Report #649). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Hoff, D. (2000, Jan. 26). Testing ups and downs predictable. *Education Week*, pp. 1, 12–13.

Linn, R. (2000). Assessment and accountability. *Educational Research*, 29, 2.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21. (ERIC Document Reproduction Service No. EJ 436 999)

Porter, A. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31, 3–14.

VanLehn, K. (1996). Cognitive skill acquisition. In J. Spence, J. Darly, & D. J. Foss (Eds.), *Annual review of psychology* (Vol. 42, pp. 513–539). Palo Alto, CA: Annual Reviews.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Madison, WI: University of Wisconsin, National Institute for Science Education.

Joan L. Herman (herman@cse.ucla.edu) and **Eva L. Baker** (eva@ucla.edu) are Codirectors of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California-Los Angeles, 300 Charles Young Dr. North, Los Angeles, CA 90095.