Joan L. Herman
Ellen Osmundson
Ronald Dietel

# Benchmark Assessment for Improved Learning

Assessment and Accountability Comprehensive Center

AACC • A WestEd and CRESST partnership

AACC: Assessment and Accountability Comprehensive Center: A WestEd and CRESST partnership.
aacompcenter.org



The work reported herein was supported by WestEd, grant number 4956 s05-093, as administered by the U.S. Department of Education. The findings and opinions expressed herein are those of the author(s) and do not necessarily reflect the positionsor policies of AACC, WestEd, or the U.S. Department of Education.

To cite from this report, please use the following as your APA reference:
Herman, J. L., Osmundson, E., & Dietel, R. (2010). *Benchmark assessments for improved learning* (AACC Report). Los Angeles, CA: University of California.

The authors thank the following for reviewing this policy brief and providing feedback and recommendations: Margaret Heritage, (CRESST); and for editorial and design support: Judy K. Lee and Amy Otteson (CRESST).

# Benchmark Assessments for Improved Learning

Joan L. Herman, Ellen Osmundson, & Ronald Dietel

## ABSTRACT

This report describes the purposes of benchmark assessments and provides recommendations for selecting and using benchmark assessments—addressing validity, alignment, reliability, fairness and bias and accessibility, instructional sensitivity, utility, and reporting issues. We also present recommendations on building capacity to support schools' and districts' use of benchmark assessments for improved learning.

## INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB, 2002) has produced an explosion of interest in the use of assessment to measure and improve student learning. Initially focused on annual state tests, educators quickly learned that results came too little and too late to identify students who were falling behind. At the same time, evidence from the other end of the assessment spectrum was clear: teachers' ongoing use of assessment to guide and inform instruction—classroom formative assessment—can lead to statistically significant gains in student learning (Black & Wiliam, 1998).

Between state and formative assessment is benchmark assessment1, defined as follows:

> *Benchmark assessments are assessments administered periodically throughout the school year, at specified times during a curriculum sequence, to evaluate students' knowledge and skills relative to an explicit set of longer-term learning goals. The design and choice of benchmark assessments is driven by the purpose, intended users, and uses of the instruments. Benchmark assessment can inform policy, instructional planning, and decision-making at the classroom, school and/or district levels.*

In the following sections, we describe the role of benchmark assessment in a balanced system of assessment, establish purposes and criteria for selecting or developing benchmark assessments, and consider organizational capacities needed to support sound use.

1 We consider the terms interim assessment, quarterly assessment, progress monitoring, medium-cycle, and medium-scale assessment interchangeable with benchmark assessment.

## A BALANCED ASSESSMENT SYSTEM

Benchmark assessment is one component of a balanced assessment system explicitly designed to provide the ongoing data needed by teachers or administrators to serve district, school, and classroom improvement needs. The National Research Council (NRC) defines a quality assessment system as one that is (a) *coherent*, (b) *comprehensive*, and (c) *continuous* (NRC, 2001).

Components of a coherent system are aligned with the same significant, agreed-upon goals for student learning—that is, important learning standards. A comprehensive system addresses the full range of knowledge and skills expected by standards. It is comprehensive in covering the important content described by the standards and the range of kinds of thinking and applications—or cognitive demands—expected of students. A comprehensive system also provides users at multiple levels of the system (district, school, class) with appropriate data, at suitable levels of detail, to meet their decision-making needs. A system that is continuous provides continuous streams of data about student learning throughout the year, thus providing district and school decision-makers with periodic information for monitoring student learning, establishing a rich and productive foundation for understanding student achievement.

Where do benchmark assessments fit in a balanced assessment system? While annual state assessments provide a general indicator of how students are doing relative to annual learning standards, and while formative assessment is embedded in ongoing classroom instruction to inform *immediate* teaching and learning goals, benchmark assessments occupy a middle position strategically located and administered outside daily classroom use but inside the school and/or district curriculum. Often uniform in timing and content across classrooms and schools, benchmark assessment results can be aggregated at the classroom, grade, school, and district levels to school and district decision-makers, as well as to teachers. This interim indication of how well students are learning can fuel action, where needed, and accelerate progress toward annual goals.

Figure 1 highlights our conceptualization of the interrelationships between these three types of assessments—state, benchmark, and formative—in a balanced system. The learning targets assessed by frequent classroom-formative assessment contribute to the long-term targets addressed by periodic benchmark assessments. Benchmark data flows into the annual assessment, which in turn transfers into subsequent years of teaching, learning, and assessment.

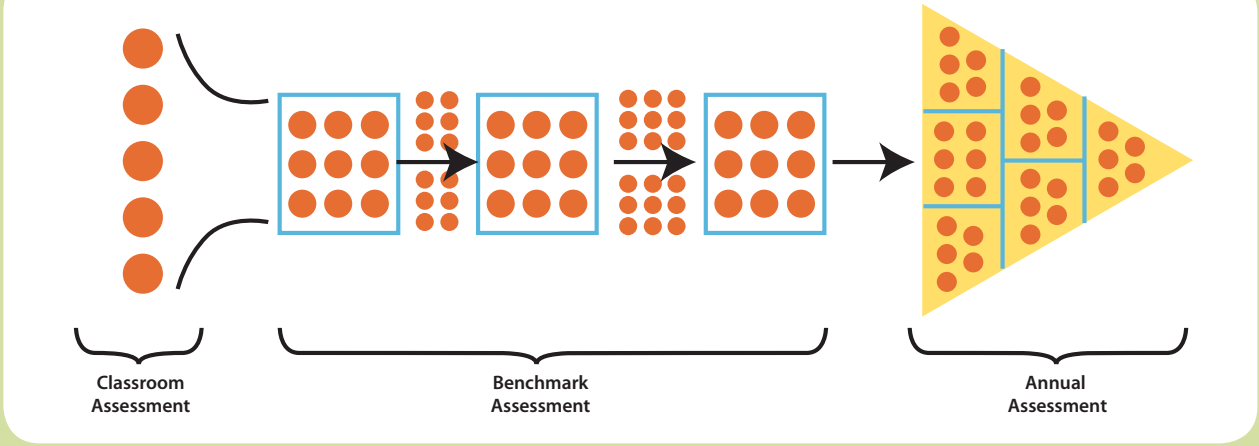
### Key Questions to Consider When Selecting Benchmark Assessments

As educational leaders consider the addition of benchmark assessments to an already assessment-heavy calendar, it is important to establish clear understandings of the nature and purpose of these assessments. We suggest that policymakers answer the following questions prior to adopting or developing benchmark assessments for their school or district:

1. What purposes do you expect benchmark assessments to serve?
2. What criteria should you use to select or create benchmark assessments?
3. What organizational capacity is needed to successfully support a benchmark assessment program?

**FIGURE 1.**

Quality Assessment System: Multiple formative classroom assessment feeding into each benchmark assessment and multiple benchmark assessment feeding into annual assessment.




## PURPOSES OF BENCHMARK ASSESSMENTS

Benchmark assessments often serve four interrelated but distinct purposes: (a) communicate expectations for learning, (b) plan curriculum and instruction, (c) monitor and evaluate instructional and/or program effectiveness, and (d) predict future performance. In the following sections, we briefly discuss and illustrate examples of each purpose, highlighting what, how, and by whom the results could be used. Note that the four purposes are not mutually exclusive—many benchmark assessments address more than one purpose.

### Communicate Expectations

How many of us remember hearing (or asking), "What's going to be on the test?" Benchmark assessments, as holds true for all assessments, communicate a strong message to students, teachers, and parents about *what knowledge* is important to learn, *what skills* are valued, and *how* learning will be measured. Similarly, teachers want their students to perform well on important assessments and thus tend to focus classroom curriculum and instruction on what will be assessed and to mimic assessment formats (e.g., Herman, 2009). This last quality, *how* learning is measured, provides additional rationale for not limiting benchmark assessments to traditional multiple-choice formats. Constructed response items, when appropriately designed, provide an important window into students' thinking and understanding. In particular, they communicate the expectation that complex thinking and problem solving should be a regular part of curriculum and instruction.

### Plan Curriculum and Instruction

Benchmark assessments can serve *instructional curriculum* and *planning purposes* by providing educators the information needed to develop and adjust curriculum and instruction to meet student-learning needs. To do so, benchmark assessments must be aligned with content standards and major learning goals for the period addressed by the assessment *and* provide feedback on students' strengths and weaknesses relative to those goals.

Consider, for example, a first quarter benchmark

assessment in one district where fourth graders are learning how whole numbers and decimals relate to fractions. Specifically, students are learning about tenths and hundredths in decimal and fraction notations, decimal and fractional equivalents (e.g., ½ = 0.5 or 0.50), and the conceptual models for how these representations relate to one another. A benchmark assessment that provides good information for planning instruction would provide data on how well students have learned these concepts. Ideally, the assessment could also diagnose challenges students encountered in each focus area (e.g., how well students can convert fractions to decimals and how well they can solve problems that require understanding of proportional reasoning). This benchmark assessment would not, however, include items on concepts not taught, such as negative numbers or the multiplication and division of fractions and decimals.

When commonly administered across classrooms, grade level, and/or content areas, benchmark assessment data provide teachers an opportunity for collaborative reflection, analysis, and action. School administrators and/or leadership teams may also use benchmark assessments for planning and targeting specific program interventions.

**Monitor and Evaluate Learning**

Benchmark assessments can also be used for monitoring and evaluation purposes by providing information to teachers, schools, or districts about how well programs, curriculum, or other resources currently being implemented are helping students achieve learning goals. The use of benchmark assessments can help administrators or educators make mid-course corrections if benchmark assessment data reveal patterns of insufficient performance and may highlight areas where a curriculum can be refined or supplemented to better support students and their learning. For example, benchmark assessment results can help a district or school experiment with two different approaches to reading or math instruction. Or they can serve as an early warning system for an instructional approach that is not meeting its goals.

Districts and schools also can use benchmark data to evaluate patterns and trends in school-by-school or teacher performance. Such data provide guidance for standardizing or adjusting curriculum and instruction across a district if there are substantial differences in performance between schools. Teachers or schools whose students outperform similar students may be asked to share their practices.

The use of benchmark assessments for progress monitoring merits a special note. Educators often report using benchmark assessments to evaluate or monitor student progress, but most benchmark assessments are designed to measure knowledge and skills learned during a specific time period. They are not cumulative progress indicators.

For example, imagine that a student gets 65% of the items correct on the first quarter benchmark assessment focusing on fractions and decimal equivalents. The same student gets 75% correct on the second benchmark, covering geometry and measurement concepts. Because the two benchmarks assess substantially different topics and one may contain more difficult items than the other, it is inaccurate to compare the two scores (65% vs. 75%) and say that the student is improving. Similarly, with English language arts assessments, if one test focused on students' ability to read

and analyze informational texts or skill in writing narrative stories, whereas the next test focused on analysis of literature and comparisons of plot structure and expository writing, results from the two would not be directly comparable.

### Predict Future Performance

As an alternative perspective on monitoring performance, district, and schools; teachers can use benchmark assessment data to predict whether students, classes, schools and districts are on course to meet specific year-end goals—or commonly, to be classified as proficient on the end-of-year state test. Benchmark results that predict end-of-year performance can be disaggregated at the individual student, sub-group, classroom, or school level to identify who needs help. Once those who are struggling are identified (e.g., students who are not on target for meeting specified year-long goals or schools or grade levels where too many students are not on track for meeting performance targets), steps can be taken to provide additional support and resources. Schools and districts may use benchmark results to allocate resources (i.e., time, staff, professional development, technical assistance, special interventions). Conversely, assessment results can identify students or groups who are excelling and may benefit from a more advanced instructional program.

### Addressing Multiple Purposes

Given the scarcity of time and resources in educational settings, it should come as no surprise that many organizations attempt to use one assessment for multiple purposes. However, the NRC warns: "…the more purposes a single assessment aims to serve, the more each purpose is compromised. … assessment designers and users [need to] recognize the compromises and trade-offs such use entails" (2001).

In line with NRC's recommendation, researchers suggest that benchmark assessments be selected or designed to fit the finest grain-size of information needed (Herman & Baker, 2005; Perie, Marion, & Gong, 2007). For example, a benchmark assessment created to serve instructional planning purposes by providing student learning feedback may also serve evaluative or predictive purposes, if appropriately designed. However, even though benchmark assessments developed especially for predictive purposes may be aligned with standards and long-term learning goals, their results are less likely to be fine-grained enough to support instructional planning. So advance planning and careful consideration of necessary details is essential in selecting or developing benchmark assessments to serve their intended purpose(s).

## CRITERIA FOR BENCHMARK ASSESSMENTS

A plethora of benchmark assessments is currently available to educators, ranging from glossy, high-tech versions designed by testing companies that feature built-in data analysis tools and reports, to locally developed, instructionally-driven assessments. Our purpose here is not to evaluate specific benchmark assessment delivery systems or individual assessments, but rather to describe important criteria and principles that schools, districts, and/or states should consider when selecting and/or developing benchmark assessments.

### Validity

Validity is the overarching concept that defines quality in educational measurement. Simply put, validity asks the extent to which an assessment

actually measures what it is intended to measure *and* provides sound information supporting the purpose(s) for which it is used. The dual definition means that benchmark assessments themselves are not valid or invalid, rather that validity resides in the *evidence* underlying an assessment's specific use. An assessment whose scores have a high degree of validity for one purpose may have little validity for another. For example, a benchmark reading assessment may be valid for identifying students likely to fall short of proficiency on a state test but may have little validity for diagnosing the specific causes of students' reading difficulties.

***What to look for:***

The evaluation of quality in any benchmark assessment system starts with a clear description of the purpose(s) an assessment is intended to serve and serious consideration of a range of interrelated issues bearing on how well a given assessment or assessment system serves that purpose(s). Consequently, benchmark assessments must

- Be **aligned** with district and school learning goals and intended purposes,
- Provide **reliable** information for intended score interpretations and uses,
- Be **fair**, **unbiased**, and **accessible**,
- Be **instructionally sensitive**,
- Have **high utility**, and
- Provide useful **reporting** for intended users and purposes.

## Alignment

Alignment describes how well what is assessed—the content and processes elicited by assessment items—matches both what schools are trying to teach and the assessment purposes. When evaluating benchmark assessment alignment, consider the following interrelated questions about alignment:

1. Do the assessments capture the *depth* and *breadth* of district and schools learning goals?
2. Do the assessments reflect what is most important for students to know and be able to do in a specific content area?
3. Is the framework on which the assessment is based consistent with that of the local curriculum?
4. Is the sequence of assessment content on successive assessments consistent with that of the local curriculum?

Questions 1 and 2 ask schools to look beyond simply mimicking the content and format of what is on the end-of-year state test. It causes them to consider what is most important for students to learn and how the students should apply that knowledge.

Questions 3 asks whether the curriculum and assessments reflect the same or a similar conceptual framework, particularly in terms of what types of knowledge and levels of cognitive demand are valued, how knowledge is expected to develop, and possible obstacles along the way.

Question 4 asks what content is to be addressed in each benchmark assessment over the course of a year, a decision that depends on the purpose of the assessment. For example, to provide diagnostic information for subsequent instructional interventions, a benchmark assessment should target the curriculum goals of the prior period.

In short, aligning benchmark assessments requires that districts and schools clearly lay out their learning goals for each instructional period.

**Specifying Expected Learning.**

Herman and Baker (2005) have suggested that districts and schools use the "big ideas" or key principles of each subject area to organize the specific learning expectations described by state standards. For example, major ideas such as rational number equivalence, properties of operations, functions, equations, etc., are foundational to understanding algebra. These ideas can provide a powerful framework for organizing the specific learning objectives inherent in state standards to make them more manageable and at the same time help both teachers and students to understand the meaning and see the relationships between these major ideas. A similar rationale underlies the common practice focusing on "power standards," the standards and learning objective that are most central to understanding a content area and that are most central in preparing students for success for the next level of learning. Current common core state standards, in fact, use such a strategy.

Beyond laying out the essential content, specifications also need to define what students are expected to *do* with the content and at what *level* they are expected to be able to apply or use it. The terms *cognitive demand*[2] or *cognitive complexity* may be used to classify the level of application. A commonly used scheme developed by Norman Webb (1997) identifies four general categories of cognitive demands: recall, conceptual understanding/application, problem solving/strategic thinking, and extended thinking.

2 For assessment purposes, **cognitive demand** refers to the type of thinking required by an assessment. The thinking may range from factual recall (memory) to applied knowledge while solving a difficult problem.

***What to look for:***

Many assessment companies report that their assessments are aligned with specific state standards and can serve many purposes. Savvy and prudent consumers of benchmark assessment systems, however, will want to conduct an independent analysis.

**Alignment with Learning Goals.**

With their learning goals by assessment period clearly in hand, educators should evaluate alignment by comparing their specifications with the following:

1. The framework used to develop the benchmark assessment or to classify assessment items or available item banks. Is it conceptually consistent with the local curriculum/domain[3]? For example, is there a similar conceptualization of the big ideas or major principles defining each subject area, a similar organizational framework laying out content and cognitive demands?
2. The distribution and range of items and assessment content by grade level. Do the big ideas and specific concepts addressed by grade level match those of the local curriculum or domain specification?
3. The distribution and range of cognitive demands addressed by grade level. Do assessment items cover the full range addressed by the local curriculum/domain specification? Are the types of applications elicited by assessment tasks consistent with local expectations? Balanced benchmark assessments should include all categories of cognitive demand.
4. The specific distribution of items on each assessment by content and cognitive demand. Do they represent a reasonable sample of the content and cognitive demands that are

3 **Domain** is the content of a particular field of knowledge. (Taken from: wordnetweb.princeton.edu/perl/webwn)

specified in the local curriculum? Do they emphasize targeted big ideas or power standards? Take the time to review individual items to understand what specific standards developers address and how. Examine whether the local review matches that which is provided by the publisher or developer.

5. The number and range of items. For assessments intended to guide instruction, are there a sufficient number and range of items to diagnose specific (and important) learning strengths and weaknesses? If not, the assessment will not serve the purpose.

**Examples of Alignment**

Table 1 (see page 9) displays one protocol[4] for conducting benchmark alignment studies. In this example, educators were interested in investigating the alignment of quarterly mathematics benchmark assessments with curriculum learning goals. They examined individual fourth grade items (representative ones, if the item pool is very large) and classified each by content and cognitive demand. In summary, they compared the content alignment of four different benchmark products and noticed

that for one vendor, "Data Analysis" got the most emphasis on the fourth grade benchmark math assessments (27/100, representing 27% of all items), with "Patterns and Algebraic Thinking," "Number Sense," and "Measurement" each representing 17%, 22%, and 21%, respectively. "Geometry" represented only 10% of all items on the assessment. Is this a reasonable distribution of content? Does the distribution reflect intended learning goals?

There is no right or wrong answer to the distribution; rather the analysis provides an opportunity for educators to ask themselves if the distribution aligns with what most matters conceptually in Grade 4 mathematics. If not, the school, district, or state will need to take steps to improve benchmark alignment with learning goals.

Next, the group might choose to review the distribution of cognitive demands, as displayed in Table 2. Notice that 52% of all benchmark items invoke recall and memorization, while problem-solving items represent 10% of the total tasks. Does this distribution of cognitive demand and application align with district or school goals and emphases? Again, there are no right or wrong distributions; the issue is how well the items match the kinds of learning expected in state and local standards and how well they represent the standards valued by local educators. If they don't match, then steps should be taken to better assess students' mathematical problem solving skills.

Table 2

| Cognitive demand | *N* = 33<br>Frequency (% of total) |
|---|---|
| a. Recall/memorization | 17 (52%) |
| b. Conceptual knowledge/ application | 8 (24%) |
| c. Problem solving | 3 (10%) |
| d. Extended thinking | 5 (15%) |

The alignment process can help educators think hard about what is taught and what is assessed. To make benchmark assessment use worthwhile, the benchmark assessment system should be aligned with instruction and vice versa.

4 In assessment, a protocol is usually an agreed upon procedural method or instrument development process used to measure knowledge or learning.

Table 1

Grade 4 Mathematics Benchmark Item Alignment and Analysis

| Math Concept | Benchmark/ Item from MCT Blueprint | Benchmark Assessment | | | | Total # items on benchmark assessments | Content: + = match - = not matched | Intellectual Demands a. Recall/ memorization b. Conceptual knowledge c. Problem solving d. Strategic thinking | Content: Focus: "Big Idea" |
|---|---|---|---|---|---|---|---|---|---|
| | | T1 | T2 | T3 | T4 | | | | |
| **Patterns & Algebraic Reasoning: 17 Items Total** | 1 | 1 | 1 | 1 | 1 | 4 | + | a | Variable as unknown quantity |
| | 2 | 2 | 1 | 2 | | 5 | + | b | Properties: commutative, associative, distributive |
| | 3 | 2 | 2 | | | 4 | - | a | Patterns & relationships: multiplication & division |
| | 23 | 2 | 1 | 1 | | 4 | + | d | Represent & analyze patterns, using words, tables, graphs |
| **Data Analysis: 27 Items Total** | 4 | 3 | 2 | 2 | 1 | 8 | + | d | Represent & analyze data using tables, line plots, bar graphs |
| | 5 | 1 | 1 | 1 | 1 | 4 | + | d | Data distribution |
| | 6 | 3 | 2 | 3 | 3 | 11 | - | a | Propose & justify predictions & conclusions from data |
| | 7 | 1 | 1 | 1 | 1 | 4 | + | a | Predict: likely/unlikely |
| **Measure-ment: 21 Items Total** | 2 | 1 | | 2 | | 3 | + | b | Estimation |
| | 8 | 1 | | 1 | 1 | 3 | + | a | Standard unit of measurement |
| | 9 | | | 1 | 2 | 3 | - | c | Standard unit of measurement |
| | 10 | | | 1 | 1 | 2 | + | b | Select and apply standard units |
| | 11 | | | 1 | 1 | 2 | - | b | Estimation |
| | 12 | | | 1 | 1 | 2 | + | a | Standard unit of measurement |
| | 13 | | | 1 | 1 | 2 | + | a | Estimation |
| | 14 | | 2 | 1 | 1 | 4 | + | a | Select & apply standard units & tools to measure length, area, volume, weight, time, temperature, angles |
| **Geometry: 10 Items Total** | 1 | | | 1 | 1 | 2 | + | d | Congruence & similarity |
| | 18 | | | 1 | 1 | 2 | + | c | Coordinate Grid |
| | 19 | | | 1 | 1 | 2 | - | a | Describe, analyze & classify 2- & 3-dimensional shapes |
| | 20 | | | 1 | 1 | 2 | + | b | Geometric models to solve problems |
| | 21 | | | 1 | 1 | 2 | + | b | Ordered pairs |
| **Number Sense: 25 Items Total** | 3 | 1 | 1 | | | 2 | + | a | Multiplication & division concepts |
| | 22 | 2 | 1 | | | 3 | + | a | Multiplication & division concepts |
| | 23 | 1 | 1 | | | 2 | - | a | Fluency & recall facts |
| | 24 | 1 | 1 | | | 2 | + | a | Estimations |
| | 25 | 1 | 1 | | | 2 | + | b | Multi-digit multiplication & division |
| | 26 | | 2 | | 1 | 3 | - | c | Place value: whole numbers and decimals |
| | 27 | | 2 | 1 | 1 | 3 | + | d | Estimations |
| | 28 | | | | 1 | 1 | + | a | Number sense |
| | 29 | | | | 1 | 1 | - | a | Number sense |
| | 30 | 1 | 1 | | | 2 | + | a | Place value: whole numbers and decimals |
| | 37 | | | | 1 | 1 | + | a | Multi-digit multiplication & division |
| | 38 | 1 | 1 | 1 | 1 | 4 | + | b | Multiplication & division concepts |
| **Total Number of Items** | | **25** | **25** | **25** | **25** | **100** | **34 + = 26 - = 8** | **34 a = 18, b = 8, c = 3, d = 5** | |

It can be difficult to know what is being assessed—in content or cognitive demand—simply by visually inspecting the items or tasks. Look for other evidence in addition to your or other experts' review of the content. For example, are there empirical analyses demonstrating that the assessment addresses different dimensions of content and cognitive demand (technically, factor analyses[5] ) or qualitative studies of the content and processes students use to respond to items (e.g., think aloud protocols[6] ).

### Psychometric Indicators Related to Alignment

In addition to reviewing alignment, psychometricians also often evaluate whether an assessment is measuring what it is intended to measure by examining its correspondence to other established measures of the same construct. If benchmark assessment results are highly correlated with the established measure, then this finding is taken as evidence that the new assessment also measures that construct. For example, one might look to see the relationship between benchmark results on student understanding of rational numbers and equations and teacher ratings or classroom assessments of the same concept.

### Alignment for Intended Purposes

Alignment with intended learning goals is fundamental to sound benchmark assessment, but the assessment also needs to provide suitable data for the purpose(s) of the assessment. If benchmark assessments are intended to serve instructional planning purposes, they must provide diagnostic feedback on student strengths and weaknesses and help identify the source of student difficulties. For example, score reports that simply tell a teacher that a student or group of students is performing poorly in mathematics provide little guidance for addressing student learning needs. More helpful are score reports that identify how students are performing on specific learning objectives or the depth of their learning (e.g., addition and subtraction of fractions, rational number equivalence, problem solving and application, etc., in mathematics; or identifying the main idea, using evidence, and analyzing in reading).

### Theory of Diagnosis

Benchmark assessments for instructional planning should also have an explicit design for providing diagnostic feedback. Good diagnosis is more than simply putting together a set of random items ostensibly related to the same objective and computing a correct percentage. Rather, quality diagnosis involves mapping assessment items of how each major knowledge and skill develops and of the major obstacles or misconceptions to student learning in that domain. Assessment items can then be designed to capture different kinds of misconceptions or obstacles. Distractors for multiple-choice items, for example, can be designed to represent common errors. Look for the developer's theory of diagnosis and determine whether it is consistent with teacher and curriculum perspectives.

The choice of multiple-choice versus complex-constructed response items represents an interesting tension in composing benchmark assessments. On the one hand, constructed response assessment items that ask students to: explain

5 **Factor analysis** is a statistical technique used to determine the number of components in a set of data. These components are then named according to their characteristics allowing a researcher to break down information into statistical groups. (Taken from: allpsych.com/dictionary/dictionary2.html)

6 **Think aloud protocols** involve participants thinking aloud as they are performing a set of specified tasks. Users are asked to say whatever they are looking at, thinking, doing, and feeling, as they go about their task. (Taken from: http://en.wikipedia.org/wiki/Think_aloud_protocol).

their thinking, create models, design and interpret, etc., provide both an important window into student understanding and rich data for planning and guiding subsequent instruction. Well-formulated constructed response tasks also are better at eliciting complex thinking and problem-solving (using high levels of cognitive demand), than are multiple-choice items. At the same time, however, constructed response items need special scoring, which requires time-consuming (and thus costly) training and, consequently, raises feasibility challenges. Teachers may well gain important insights for teaching and learning from scoring student work, but it may be hard to find the time to engage in this task. The costs of paying others to do the scoring may be cost-prohibitive and inhibit the timeliness of results. These competing tensions need to be balanced.

### Predictive Purposes

For benchmark assessments that serve predictive purposes, the relationship between benchmark results and end-of-year state assessment is of high interest. If the benchmark assessment scores are highly correlated with proficiency levels on the end-of-year test, benchmark scores can be used to identify students who are likely to achieve proficiency and those who are not. Known as predictive validity, correlations in the range of 0.7 and above between benchmark assessment results and the state assessments provide a reasonable amount of certainty that students who perform well on the benchmark assessment will also perform well on the state assessment (Williams, 2009). It should be noted, however, that one purpose of using benchmark assessments is to break the strong relationship between interim and end-of-year results. Ideally, teachers intervene with low-performing students and help them improve their end-of-year state test performance. In this case, the relationship between benchmark performance and state test performance would decline, but the benchmark assessment would have served an important purpose.

## RELIABILITY

Reliability is an indication of how consistently an assessment measures its intended target and the extent to which scores are relatively free of error. Think of reliability in terms of a 12-inch wooden ruler: whether you're using it today or tomorrow to measure the length of your desk, it should give the same result; the unit of measurement is fixed, thus providing a consistent measure. Consider, however, if the ruler is rubber, and stretched one day, but not the next. It could give different measures each time, not because what was being measured had changed (i.e., the length of your desk) but because of error associated with the measurement tool (in this case, the ruler). Further imagine that this rubber ruler is used by different people to measure your desk, and some have better eyesight than others and some are more committed to precision. Whoever does the measuring, in other words, may produce differences in results. Similarly, in educational testing, reliability indices tell you the extent to which scores reflect the stable item you're trying to measure, as opposed to errors introduced by the measurement process. Low reliability means that scores should not be trusted for decision-making.

Measurement experts have a variety of ways of looking at reliability. For example, they look for consistency of scores across different times (morn-

ing vs. afternoon) or occasions (taking the test twice), across different items that could comprise the test, and across different forms of the test. In terms of the ruler scenario, these are ways to test whether the ruler—the assessment—is wooden or rubber based on the different times and ways in which it can be used. Another reliability issue is having an adequate number of test questions to measure each component of a student's knowledge or skill. A test that has four or five multiplication problems, for example, will be a more reliable measure of a student's multiplication skills than a test that has only one question.

Measurement experts also look for reliability in scoring: results should be consistent regardless of who scores the test or when it is scored. In the ruler scenario, the human element (the fact that a person is doing the measuring—the level of experience using the ruler, the guidelines for rounding numbers, etc.), as well as the environmental factors in which the ruler is used (amount of light, texture of item to be measured, etc.), may influence the accuracy of the measurement, thereby introducing measurement error.

Consistency in machine-scored benchmark multiple-choice items is rarely problematic. However, reliability of scoring is an issue with complex, constructed response tasks, such as explanations and extended essays that need to be scored by humans. Here, we need to make sure that the score represents the quality of the response and not the idiosyncrasies of the individual doing the scoring. To combat the latter, raters are trained to consistently apply a scoring rubric for each benchmark assessment. Furthermore, previously scored examples of student work, called anchor papers, are often used to help raters accurately apply the rubric. Before starting official scoring, raters typically have to pass a consistency check, demonstrating that the scores they assign are consistent with scores given by experts. During long scoring sessions, occasional consistency checks may be required to further assure reliability, with additional training provided to anyone who needs it.

Reliability is a necessary but not sufficient criterion of test validity. For example, an assessment may be highly reliable but might not measure the "right" knowledge and skills. Alternately, an assessment may provide a highly reliable total score, but not provide reliable diagnostic information. Think again of the "foot-long" ruler but one that is a half-inch too short. The short ruler will provide a consistent measure each time it is used, but the measure will be inaccurate.

***What to look for:***

Test publishers typically provide reliability indices for their benchmark assessments with other technical information about item difficulty and discrimination. It is essential to review this technical information before purchasing or using benchmark assessments or item banks. For schools and districts developing their own benchmark assessments, specific statistical guidelines should be used to evaluate the reliability of assessment items prior to their widespread use (Brown & Coughlin, 2007).

Standard indices of reliability require a collection of items that are at an appropriate level of difficulty and differentiate high-achieving versus low-achieving students on a test. Item difficulty is measured in *p*-values, an indication of the proportion of students answering the item correctly. Appropriate difficultly is in the range of 0.2–0.8,

meaning that 20–80% of students answer an item correctly. Item discrimination indices essentially compare the *p*-values for high- and low-scoring students. If an item has a much higher *p*-value for students who have high overall scores than for those who have low scores on the test, this item is considered to have high discrimination power. On the other hand, if an item has a similar *p*-value for students who have high scores and those who have low scores on the test, this item is considered to have low discrimination power. A discrimination index is scaled from –1.0 to +1.0. A positive discrimination index of 0.3 and above indicates the item appropriately discriminates between high- and low-scoring students. A negative discrimination index means that low-scoring students are more likely to get it right, suggesting a faulty item that should not be on the test.

Reliability indices are derived by examining the consistency of student responses to individual items within the same test (internal consistency) or by examining the consistency of student test performance on multiple testing occasions (test–retest reliability) or across different parallel tests (parallel forms reliability). Reliability coefficients in the range of 0.8 and above (internal consistency based on Cronbach's alpha) represent reasonable assurance that benchmark assessments are providing reliable information.

Reliability indices may be calculated for the total score (all items on the test) and for each subscore. In social studies, for example, the total score may consist of five different subscores: behavioral science, economics, geography, history, and political science. Subscore reliability is very important for benchmark assessments used for instructional planning. Typically, at least five items (and often more) are needed for a reliable subscore—in other words, for an accurate diagnosis. A general rule of thumb is that the more items on an assessment, the higher the reliability. Practical considerations however, will most certainly limit the total number of items on a given benchmark assessment.

If the benchmark assessment contains open-ended items that are automatically scored, look for data on the reliability of scoring, typically the agreement between human and machine scores. Where complex responses need to be scored locally, look for data demonstrating the effectiveness of scorer training, including

1. Clear training procedures;
2. Rubrics that clearly describe the rules and criteria that should be used in assigning scores;
3. Anchor papers to be used in training and monitoring;
4. Procedures for monitoring the scoring process, recalibrating raters, where needed, and resolving discrepancies; and
5. Evidence that teachers or others can reliably apply the rubrics (e.g., interrater reliability statistics from prior use of the rubric).

Benchmark assessments that fail to provide information about their reliability will likely also fail to provide dependable information about student learning.

### Fairness, Bias, and Accessibility

Fairness and bias comprise a third, critical feature of quality benchmark assessments. A fair test is accessible and enables all students to show what they know; it does not advantage some students over others. Bias emerges when features of the assess-

ment itself impede some students' ability to demonstrate their knowledge or skill. Technically, bias is present when students from different subgroups (e.g., race, ethnicity, language, culture, gender, disability) with the same level of knowledge and skills perform differently on an assessment.

There are two primary forms of test bias: offensiveness and unfair penalization (Popham, 1995). Offensiveness becomes an issue when the content of an assessment offends, upsets, or distresses particular subgroups, thereby negatively impacting performance. Assessment items that present unfavorable stereotypes of different cultures, genders, or other subgroups could adversely affect these subgroups' performance.

Unfair penalization occurs when the content-irrelevant aspects of an assessment make the test more challenging for some students than for others because of differences in language, culture, locale, or socioeconomic status. For example, math or science tests that use complex language may reduce English learners' ability to show their mathematical knowledge because students struggle to understand the question, even if they know the content. Or a benchmark assessment that includes a reading passage and questions about a regatta might disadvantage students from an urban community and favor students from a suburban community because urban students may be less familiar with terms such as leeward or stern.

To avoid such problems, modern assessment development procedures emphasize the use of *Universal Design*. With Universal Design, student characteristics that may impede access are identified at the start of the assessment design process. Items are purposively designed or selected to be as free as possible from content-irrelevant features that could negatively impact performance of some students. For example, it is common practice to minimize the language load of questions intended to measure students' proficiency in subject matter other than reading and English language (e.g., mathematics and science).

Even with application of Universal Design principles, test items may not be accessible for all students. Some students with disabilities will still need special accommodations; for example, students with visual impairments may need magnified text or blind students may need Braille versions. Students with limited English proficiency may need access to bilingual glossaries or may need to have assessment items read aloud (if reading proficiency is not the target of the assessment). Fair benchmark assessments will provide students with the same accommodations they receive in classroom instruction.

***What to look for:***

Technical documentation regarding fairness and lack of bias should be provided by benchmark assessment developers and should address a range of issues, including:

- Steps taken to increase accessibility for all students (e.g., guidelines for reducing unnecessary linguistic complexity);
- Bias reviews to detect and remedy any identified item characteristics that may disadvantage or be offensive to some students;
- Technical analyses of differential item functioning, that is, analyses that identify items on which students of equal ability from different groups respond differently;
- Available accommodations; and

- Technical data describing the demographics of students with whom the tests have been developed or used, as well as reliability and other technical indicators for relevant subgroups (i.e., gender, ethnicity, language, disability, culture).

Districts and schools will want to ensure that there is evidence that the test they select is appropriate for their students. If, for example, a school or district has a high percentage of students who are English learners, then selecting a benchmark assessment that has been tested with a similar student population makes sense, increasing the likelihood that the assessment is appropriate for their students. Similarly, reliability indices for a benchmark assessment may be reasonably high when based on all students but low when computed for students from particular subgroups (e.g., English learners, students with disabilities) and thus not appropriate for use with those students.

### Instructional Sensitivity

The next criterion, instructional sensitivity, refers to the degree to which students' performance on an assessment accurately reflects the quality of instruction they have received (Popham, 2010) and their learning. If students have been well taught and have learned that which is assessed, they should perform well on the test. Pretty obvious—right?—and often assumed, but not always the case. Students may underperform on an assessment if some items are only tangentially related to what teachers have taught or if items are confusing, excessively difficult, or involve unfamiliar applications or item contexts.

When instructional sensitivity is lacking, assessment scores have little value for evaluating schools, teachers, or their instructional programs; furthermore, they provide faulty evidence for improving teaching and learning.

***What to look for:***

To avoid instructional sensitivity problems, teachers, schools and districts should conduct thorough reviews of the alignment between assessment items and curriculum. Furthermore, educators should insist that the assessment focuses on concepts that are central—and not tangential—to learning goals. Be on the lookout for item flaws that may confuse students or enable them to guess the right answer without having content knowledge. A list of potential flaws for multiple-choice questions can be found in Appendix A.

### High Utility

A fourth quality important to consider when selecting or developing benchmark assessments is utility. The overarching question that schools, districts, and states should ask to determine a benchmark assessment's utility is, "How usable will this assessment be in helping us to accomplish our intended purposes?" Does the assessment, for example, provide teachers with useful results to identify struggling students and guide instruction? To maximize utility, benchmark assessments must be user-friendly and feasible to administer, score, and interpret in a timely way.

***What to look for:***

Publishers typically showcase various ways that test data can be summarized and displayed. While these reporting features may be appealing on the surface, districts and schools should closely review the technical manuals that accompany each test for evidence supporting each intended use.

## Useful Reports

Benchmark data only becomes useful when reported in a timely and effective manner. We recommend six interrelated report features as follows:

***1. Different reporting levels.***

Most systems can automatically generate reports summarizing student performance for users at different levels of the system, for example: student, classroom, school, and district level summaries. Reports for classroom teachers may include both rosters with individual student performance and class summaries by content strand or standard. School-level reports should summarize overall and topic strand performance by class, grade, and subgroup. District reports should provide similar information plus performance between schools. Item-by-item performance detailing how many students got each item correct and the percentage of students choosing each distractor (or answer) should also be available for users at all levels.

***2. Reporting formats that are consistent with intended uses at each level.***

Having information at the right level of aggregation is one need, but reporting it in ways that easily fit with intended uses is another. When guiding instruction, for example, it may be helpful for teachers to receive not only class rosters that detail each student's score by content strand but also groupings of students who are performing at similar levels and who face similar learning challenges. School users may want to know how many students need special attention to reach proficiency by the end of the year. They also may want to know which teachers appear to be struggling and, concomitantly, teachers who may have promising strategies to share. Assuming the assessments remain the same or are comparable, districts and schools should have reports to compare this year's performance to the previous year(s).

***3. Scoring metrics and reporting categories that are consistent with local circumstances and proficiency levels.***

A scoring metric is simply a way of conveying a score. For example, state test reports typically include both scale scores and performance levels; teachers often score classroom tests based on percentage correct. These are different metrics for making meaning of the same type of responses on a test. These metrics may be summarized in a variety of ways: for example, in terms of mean (average) performance and the percentage of students scoring at different performance levels (e.g., below basic, basic, proficient, advanced). Many systems can and do report scores in a variety of ways (e.g., raw score, percent correct, scale or standardized score, percentile score) and use a variety of metrics to group students in different performance categories (e.g., students who score in the bottom 25%, students who got at least 75% of the items correct, or all students proficient and above). Some questions to ask include: What are the most meaningful metrics and groupings for the local district or school context? How do these metrics fit (or not) with the way your state, school or district defines proficiency or mastery? Are the grouping categories sensible given the range of students in your school or district? Will the metrics and categories be understandable to users?

***4. Reporting formats that use multiple representations.***

Data reports can be intimidating and difficult to

interpret. Just as it's a good idea to use multiple representations to help students understand and apply new concepts and principles, so can multiple representations (tables, graphs, text) help educators better "see" the data and draw inferences from it. User-friendly reports, aligned with intended use, help effectively convey benchmark data to diverse audiences.

***5. Flexibility.***

As districts and schools gain experience with benchmark assessments, they may discover new uses for the data and need reports that do not come standard with the system. Flexible systems provide the capacity to generate custom reports (e.g., a comparison of the performance of students in the new afterschool tutorial program to similar students who did not participate or a report on how students' performance on a unique cluster of items compare to another).

***6. Reliability of reported scores and inferences.***

Reports can look professional and appear very useful for decision-making. However, as noted earlier, if the reports do not provide reliable information, they should not be trusted for teacher and school decision-making. Look for data on score reliability, particularly reports about student performance on individual strands or topics. If reports include designations of mastery or proficiency, look for evidence to justify such classifications. Confirm interpretations with other available knowledge and data (e.g., teachers' observations and experience, evidence from class assignments, and tests).

A number of vendors now link diagnosis of student needs to targeted instructional activities or materials to help students improve their skills. However, just as wise benchmark test consumers will look for data to ensure that such diagnoses are reliable, so too will they ask for evidence that the targeted resources actually are effective in promoting student learning.

There will be tradeoffs in any system, with each benchmark assessment embodying different strengths and weaknesses. One benchmark assessment, for example, may be well aligned with district learning goals and state standards but is too difficult or costly to administer. Another benchmark assessment may be more feasible but fail to provide good diagnostic information or lack data on fairness or accommodations for special populations. The preceding criteria, which support the validity of a quality assessment system, can help you find the right balance for your district or school.

## BUILDING ORGANIZATIONAL CAPACITY FOR BENCHMARK ASSESSMENT

In the process of selecting or developing benchmark assessments, districts and schools need to carefully consider the infrastructure and systems needed for the benchmark assessment process to run smoothly and efficiently so that educators can make good use of assessment results. Decisions about how, when, and by whom the assessments will be administered, scored, analyzed, and used will influence the kinds of resources and support school personnel need. Efficient implementation and scoring procedures, management information systems that provide easy data access and analysis, organizational supports for appropriate interpretation and use, instructional supports for teachers and students, and time are all critical factors to be considered when implementing a benchmark assessment program.

Finally, we describe four conditions necessary to support effective use of benchmark assessments adapted from current research (Goertz, Olah, & Riggan, in press). Undergirding the plan is a culture conducive to data use, including high expectations, trust, and valuing data.

***1. Begin with a written plan.***

Investing in benchmark assessments is a costly, time-intensive undertaking. A written district or school accountability plan, including benchmark assessments, can help your school or district reach goals and save time. At minimum, we suggest that your plan includes the purposes of the benchmark assessments, individual responsibilities and timelines, reporting, data use, professional development, resources, and evaluation. A professional learning community, data champion, or accountability officer may be tasked with the effort and asked to involve a full range of stakeholders in its development.

***2. Identify systems for analyzing and reporting data.***

Whether a district or school chooses to purchase a software program to make benchmark assessment data available to teachers, principals, and districts or to develop their own tool, it is important that data are quickly and easily available to all stakeholders. Clear rules should be established and communicated widely regarding access to data (who, when, how, and for what purpose), and for protecting confidentiality, to ensure users have access to the data they need without compromising confidentiality of the subjects.

***3. Provide professional development.***

Schools and districts can do much to encourage the use of data from the benchmark assessments by providing high-quality, ongoing professional development. Teachers, schools, and district personnel require assistance in building their technical skills to access, organize, and interpret benchmark assessment data. Professional development should include content and pedagogical skills that help teachers differentiate instruction and revise instructional strategies and approaches based on data.

***4. Allocate time.***

Districts and schools should carefully build time into their calendars to make effective use of benchmark data. Data users, including assessment and content experts, need time to adequately analyze the data in ways that are both meaningful to their context and robust in terms of the analyses. Teachers need time for instructional planning to address weaknesses in student skills identified by the assessment. There is little value in pinpointing gaps in student understanding if the pace of the district curriculum mandates that teachers forge ahead to the next topic, regardless of student performance and needs.

## CONCLUSION

Good benchmark assessments should be part of a comprehensive assessment system. Benchmark assessments should be well aligned with curriculum and provide a continuous, comprehensive stream of information to plan and guide instruction. Validity, adequate resources, and ongoing professional development are three other key components. Benchmark assessments, if selected or developed carefully and used appropriately can serve as key tools in schools' continuous improvement efforts and provide important benefits for student learning and achievement.

## REFERENCES


Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice. 5*(1), 7–74.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region.* Retrieved from: http://www.mhkids.com/media/articles/pdfs/resources/Predictive-Validity.pdf

Goertz, M. E., Olah, L. N., & Riggan, M. (in press). *Using formative assessments: The role of policy supports.* Madison, WI: Consortium for Policy Research in Education.

Herman, J. L. (2009). *Moving to the next generation of standards for science: Building on recent practices* (CRESST Report 762). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership, 63*(3), 48–55.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy of Sciences.

Niemi, D., Vallone, J., Wang, J., & Griffin, N. (2007). *Recommendations for building a valid benchmark system: Interim report to the Jackson Public Schools* (CRESST Report 723). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

No Child Left Behind Act of 2001, Pub. L No. 107–110, 115 Stat. 1425 (2002).

Perie, M., Marion, S., & Gong, B. (2007). *The role of interim assessment in a comprehensive assessment system: A policy brief.* Aspen, CO: The Aspen Institute.

Popham, W. J. (2010). *Everything school leaders need to know about assessment.* Thousand Oaks, CA: Corwin Press.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know.* Boston, MA: Allyn and Bacon.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science* (Research Monograph No. 6). Madison, WI & Washington DC: National Institute for Science Education, University of Wisconsin, Wisconsin Center for Educational Research and Council of Chief State School Officers.

Williams, L. L. (2009). *Benchmark testing and success on the Texas Assessment of Knowledge and Skills: A correlational analysis* (Doctorate dissertation, Publication No. AAT 3353754). Phoenix, AZ: University of Phoenix. Retrieved from: http://gradworks.umi.com/33/53/3353754.html

## APPENDIX A

### Multiple Choice Item-writing Guidelines[1]

**David Niemi**



Multiple-choice items are items in which a question (stem) and a set of possible answers (responses) are presented and the student is expected to choose the correct one. Incorrect responses are called distractors.

Two principles of assessment item design:

1. Students who know the content should be able to complete the item successfully. The language and format of the question should not mislead them.

2. Students who don't know the content should not be able to complete the item successfully (except for a slight possibility of guessing correctly).

**Question wording**

The most important and most obvious consideration in writing an assessment item is to make sure that the item measures the student's content knowledge and not some irrelevant skill. Science and math items, for example, should not require exceptional reading ability.

Information should be presented as clearly and concisely as possible so that students are not confused by what is being asked. At the same time, it should not be possible for students who do not know the content to get the item correct based on language cues.

**Instructions should accompany each question or set of questions; for example:**
**Multiple Choice Item Instructions:**

*Circle the picture that shows six birds.*

*-OR-*

*Fill in the oval next to the correct answer.*

Instructions in an item should be placed before graphics or other text.

[1] This appendix was taken from CRESST Report 723: Appendix B originally published July 2007 (Niemi, Vallone, Wang, & Griffin, 2007)

## GUIDELINES FOR MULTIPLE CHOICE ITEMS

**GUIDELINE 1** Avoid unusual or difficult vocabulary (unless that vocabulary is the focus of the item). Eliminate unnecessary wordiness, and avoid overly complex sentence structures.

Problematic Item:

Which of the following materials is known for having the property of combustibility, and in addition to possessing that property is one that will ignite in the least laborious fashion?

a. coal
b. gasoline
c. rubber
d. steel

**GUIDELINE 2** In general, avoid fill-in-the blank and sentence completion formats. It is preferable to write questions.

Problematic Item:

Sacramento is the _____ city of California.

a. largest
b. westernmost
c. richest
d. capital

**GUIDELINE 3** Make sure there is only one correct or best answer. The correct answer should be one that subject-area experts would agree is clearly the best.

Problematic Item:

Who is the best writer?

a. J. K. Rowling
b. C. S. Lewis
c. E. B. White
d. F. L. Baum

**GUIDELINE 4** Avoid negative sentence structures, which tend to be confusing. Use "none of the above" and "all of the above" sparingly, especially in items for younger students.

Problematic Item:

Which statement about the set of whole numbers {1, 2, 3, . . . 20} is not true?

a. half are not even numbers.
b. more than 50% are not prime numbers.
c. 70% are not multiples of 3.
d. none of the above

**GUIDELINE 5** Use plausible distractors. Watch for unintentional clues that make it possible to identify the correct response or to exclude incorrect response options.

Problematic Item:

What is the largest mammal?

a. huge blue whale
b. ant
c. redwood tree
d. small rock

**GUIDELINE 6** Avoid absolute terms (e.g. always and never) and vague qualifiers (e.g. usually and generally).

Problematic Item:

Which of the following is true?

a. Cotton is never white.
b. Cotton is always white.
c. Cotton is usually white.
d. Cotton is generally used to make clothes.

**GUIDELINE 7** In multiple choice questions, use 3 response options for kindergarten and first grade students, 4 for students in Grades 2-12.

**GUIDELINE 8** Avoid repetitious language in response options.

Problematic Item:

How did Isaac Newton learn algebra?

a. Isaac Newton learned algebra by taking classes at Oxford University.
b. Isaac Newton learned algebra by studying algebra.
c. Isaac Newton learned algebra by taking classes at Cambridge University.
d. Isaac Newton learned algebra by inventing it.

Usually it is not necessary to use articles at the beginning of short responses; e.g., just use "boat" instead of "a boat" or "the boat" (where the question is, say, "Which of these is best for traveling across water?").

**GUIDELINE 9** Avoid intentionally deceptive or "trick" questions.

Problematic Item:

Where did George Bush send his diplomats to negotiate with Napoleon?

a. Paris
b. London
c. Berlin
d. Nowhere

**GUIDELINE 10** In general, there should only be one correct response to a multiple choice item.

**GUIDELINE 11** Whenever possible, put measurement units in the stem rather than repeating them in the responses, e.g., "What is the length of this line in centimeters?"

**GUIDELINE 12** Avoid using "you," e.g., "What should you use to measure volume?"

**GUIDELINE 13** Randomize the order of response options, except for numbers, which can be arranged in ascending order.

**GUIDELINE 14** These are guidelines, not laws. Item writers and reviewers will have to use judgment to determine (a) how to apply the guidelines in specific cases and (b) whether the two principles of assessment design have been observed or not.