Assessment and Accountability Comprehensive Center

AACC • A WestEd and CRESST partnership

Benchmark Assessment for Improved Learning

AN AACC POLICY BRIEF

Joan L. Herman, Ellen Osmundson, & Ronald Dietel

For a more detailed report, please refer to our Full Report available at aacompcenter.org

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 Policy Brief). 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

An AACC Policy Brief

Joan L. Herman, Ellen Osmundson, & Ronald Dietel

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 assessment[1], 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.

A BALANCED ASSESSMENT SYSTEM

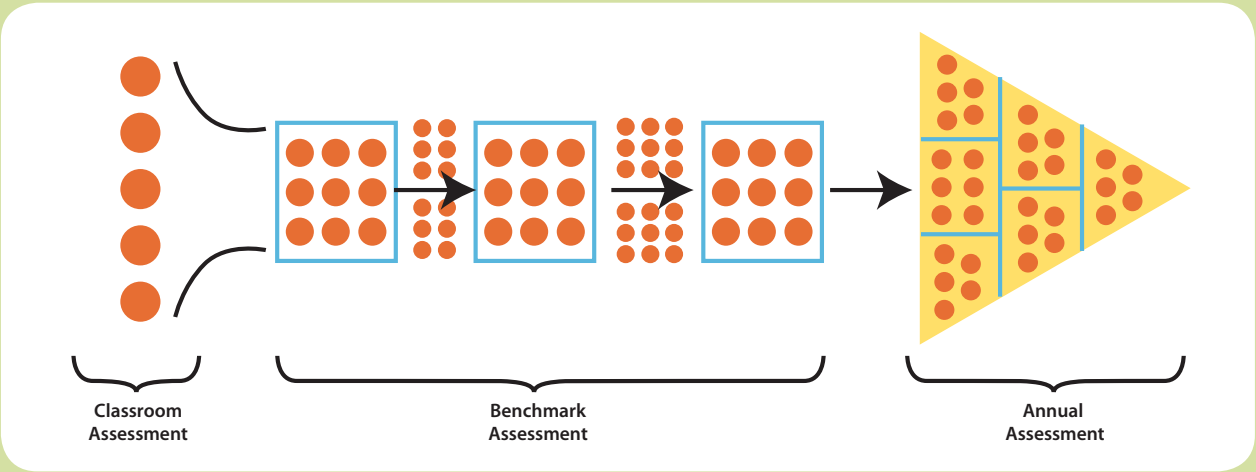
Benchmark assessment operates best when it is seen as one component of a balanced assessment system explicitly designed to provide the ongoing data needed to serve district, school, and classroom improvement needs. The National Research Council (NRC) defines a quality assessment system as one that is (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, (i.e. Important learning standards). A comprehensive system addresses the full range of knowledge and skills expected by standards while providing district, school and teachers with data to meet their decision-making needs. A system that is continuous provides ongoing data throughout the year.

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

FIGURE 1.

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



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—classroom, benchmark, and annual—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?

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. 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

Benchmark assessments communicate a strong message to students, teachers, and parents about what knowledge and which skills are important to learn. Teachers, who want their students to perform well on important assessments, tend to focus classroom curriculum and instruction on what will be assessed and to mimic assessment formats (see, for example, Herman, 2009). This last quality, how learning is measured, provides additional rationale for not limiting benchmark assessments to traditional multiple-choice formats, which too often emphasize low-level knowledge. Constructed response items (i.e., essays, extended multi-part questions, portfolios, or even experiments) as can innovative multiple-choice items, not only provide an important window into students' thinking and understanding, but also can communicate expectations that complex thinking and problem-solving should be a regular part of curriculum and instruction.

Plan Curriculum and Instruction

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

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 are helping students achieve learning goals. Benchmark assessments can help administrators or educators make mid-course corrections if data reveal patterns of insufficient performance, and may highlight areas where a curriculum should be refined or supplemented. Special care, however, must be taken in using benchmark assessment to monitor or evaluate student progress. Most benchmark assessments are designed to measure what students have learned during the previous period of instruction and do not provide an indicator of progress. Simply comparing students' scores from one time point to the other does not tell you whether student performance is improving, unless the tests are specially designed to do so.

Predict Future Performance

Benchmark assessment can provide data to predict whether students, classes, schools and districts are on course to meet specific year-end goals—or commonly, be classified as proficient on the end-of-year state test. Results that predict end-of-year performance can be disaggregated at the individual student, subgroup, classroom, and school levels to identify who needs help and to provide it.

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 National Research Council warns: "…the more purposes a single assessment aims to serve, the more each purpose is compromised. (NRC, p. 53, 2001).

CRITERIA FOR BENCHMARK ASSESSMENTS

A plethora of benchmark assessments are currently available to educators, ranging from glossy, high-tech versions designed by testing companies to locally developed, teacher-driven assessments. Below we describe important criteria and principles that schools, districts, and states should consider when selecting 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.

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).

Alignment

Alignment describes how well what is assessed—the content and the type of learning—matches both what schools are trying to teach and the assessment purposes. The over-riding questions are the extent to which the assessments reflect:

- What is most important for students to know and be able to do in a specific content area?
- What is the depth and breadth of district and school learning goals?
- What is the sequence of the local curriculum?

What to look for:

Many assessment companies report that their assessments are aligned with specific state standards. But educators should turn to the assessment framework or specifications and even the item pools provided by assessment providers to examine the match between the assessments and their curriculum and purposes.

Alignment to Curriculum.

An independent benchmark assessment analysis should compare school or district goal specifications with the:

1. Framework used to develop or classify available items: Is it consistent with local conception of curriculum and what's important for students to learn?
2. Distribution and range of assessment item content by grade level: do they match those of the local curriculum or domain specification?
3. Distribution and range of types of learning addressed by grade level: do they cover the full range including complex thinking and problem solving applications?
4. Specific distribution of items on each assessment by content and cognitive demand: do they reasonably represent the learning goals for the instructional period being assessed?

Alignment for Intended Purposes.

Alignment with learning goals is fundamental, but benchmark assessments must also be designed to serve their intended purpose(s). If benchmark assessments are intended to serve instructional planning purposes, they must provide diagnostic feedback on student strengths and weaknesses and help to identify the source of student difficulties. Good diagnosis involves mapping assessment items to a theory of how each learning toward particular objectives or standards develops and of the major obstacles or misconceptions that occur. Distractors for multiple-choice items, for example, can be designed to represent common errors. Look for the developers' theory of diagnosis and whether it is consistent with teacher and curriculum perspectives.

For benchmark assessment that are intended to serve predictive purposes, the benchmark assessment scores should be highly correlated with proficiency levels on the end-of-year test. 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). However, if teachers successfully use benchmark results to help low performing students improve, then the strong relationship between benchmark and end-of-year state test scores may decline.

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. 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 or occasions (i.e., whether a student took a test early in the day or later, now or next week). Measurement experts also look for reliability in scoring: results should be consistent regardless of who scores the test or when. 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.

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).

Fairness and Bias

Fairness and bias comprise the next 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 assessment itself impede 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.

What to look for:

Technical information regarding fairness and bias should be provided by benchmark assessment developers and should include demographics of the sample as well as scores and other technical evidence for various subgroups. Additionally, benchmark assessments should be examined prior to their use to ensure that the particular items will not be offensive to students in that organization. Guidelines for developing benchmark assessments that are free from bias reflect the same steps outlined above:

1. Developers should be sensitive to the demographic characteristics of the students;
2. Documentation describing the steps taken to minimize bias in the assessment items should be provided;
3. Organizations should examine how well an item functions for specific subgroups; and
4. If particular subgroups perform differently, validity of those items for the subgroups should be investigated.

High Utility

A final criterion important to consider when selecting or developing benchmark assessments is utility. The overarching question schools, districts and states should ask to determine a benchmark assessment's utility is: How useful will this assessment be in helping us to accomplish our intended purposes?

What to look for:

Specific utility questions to ask include what are the costs of purchasing or creating benchmark assessments? Who scores the assessments and what are the associated costs? What training may be necessary and how will it fit into existing schedules? Are there opportunity costs to be considered for students, teachers and administrators? How will the results be reported? Who analyzes and reports the scores and to which groups? How will the results fit with other assessments, both formative and end-of-year state tests? Who will use the results? How will we evaluate the quality of the assessments and their use?

Balance

No benchmark assessment program will perfectly match the preceding criteria nor meet all of a school or district's assessment needs. There will be tradeoffs, 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 be 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. However, the preceding criteria, supporting the validity of a quality benchmark system, will help you find the right balance for your district or school 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. Below we describe four conditions necessary to sustain effective use of benchmark assessments adapted from current research (Goertz, Olah, & Riggan, in press) and practice. 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, reporting, data use, professional development, resources, and evaluation.

2. Identify systems for analyzing and reporting data.

Whether an organization 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 to ensure users have access to the data they need without compromising confidentiality or the data system.

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 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 identified by assessment results. Time is also needed to enact instructional plans. There is little value in pinpointing gaps in student understanding if the district pacing guide mandates that teachers forge ahead to the next topic, regardless of student performance and needs

CONCLUSION

Good benchmark assessments can be an important addition to 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, sufficient resources, and professional development are three other key components.

For a more detailed report, please refer to our Full Report available at aacompcenter.org

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/PredictiveValidity.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).

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

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

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

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