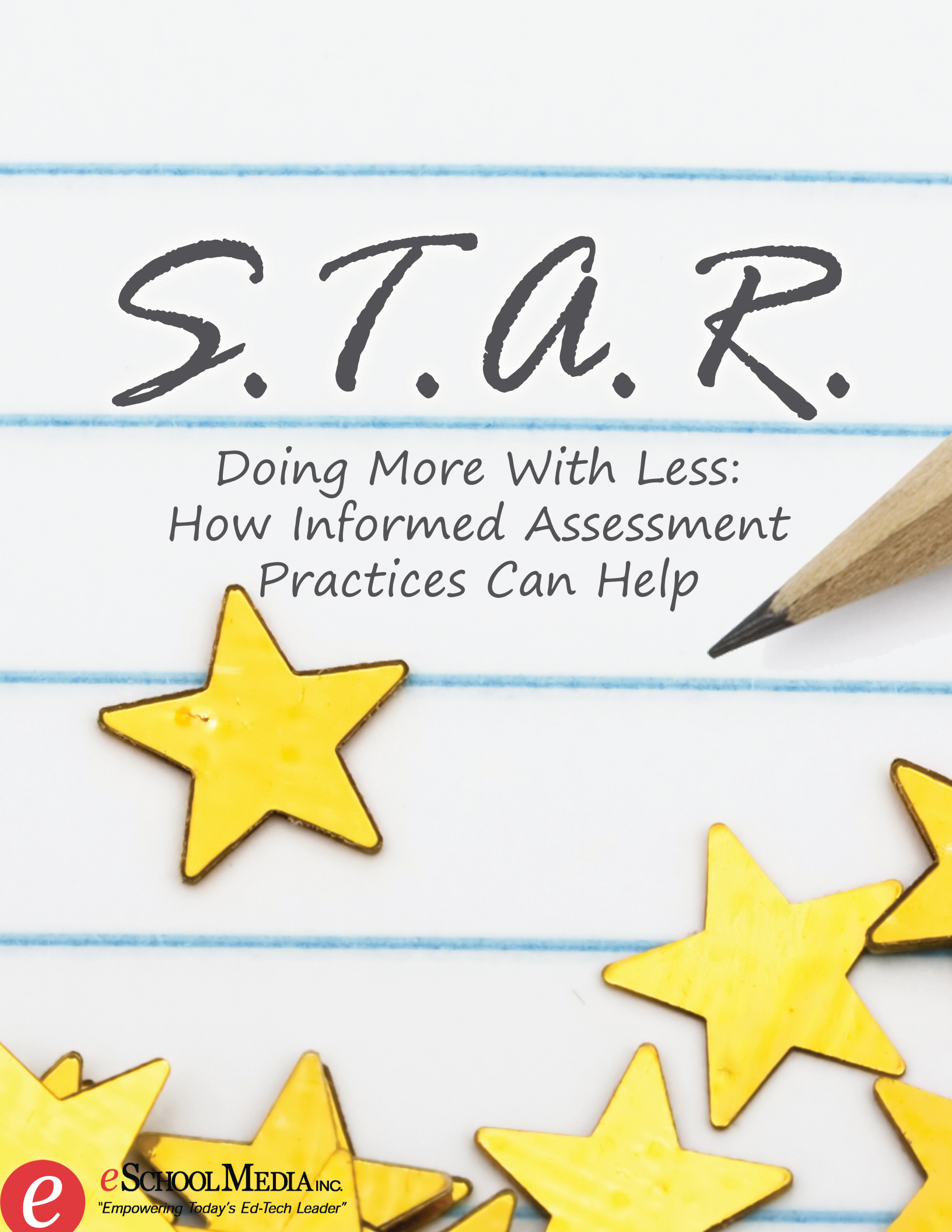
S.T.A.R.

Doing More With Less: How Informed Assessment Practices Can Help



eSCHOOL MEDIA INC.
"Empowering Today's Ed-Tech Leader"

Dear Reader,

With the current national emphasis on education reform and Common Core State Standards, too often assessments are highlighted without any real understanding of exactly how they can directly measure 21st-century skills and provide data to inform school and district practices.

Without a comprehensive understanding of what makes a good 21st-century assessment, the different types of assessments, and how these assessments contribute to short- and long-term goals and planning, schools, districts, and states can have the best-designed assessments in the world…but they won't make a difference if no one understands how they work or what they can provide.

In this latest School Technology Action Report, "Doing More with Less: How Informed Assessment Practices can help"—part of a series of **STAR** documents from *eSchool Media*: timely collections of news stories, case studies, white papers, and industry reports and surveys on pressing issues and relevant topics in education technology—you'll find a collection of reports and research detailing 'Assessment Literacy,' or what educators, administrators and policy-makers should know about formative assessment, performance assessments, the importance of assessment alignment, using and understanding assessment data, and accessible testing and universal design.

Not only can informed assessment practices ultimately provide students a better chance at proficiency and success, but can also help schools and districts spend less and save much-needed time.

Thank you for reading this latest report, and be sure to check back soon for another **STAR** on a new topic.

Sincerely,
The editors at ***eSchool Media***

Dennis Pierce,	*Laura Devaney,*	*Meris Stansbury,*	*Dennis Carter,*
Editor	Managing Editor	Associate Editor	Assistant Editor




measured progress

It's all about student learning. Period.

50 Education Way, Dover, NH 03820 | P.O. Box 1217, Dover, NH 03821-1217 | Web: measuredprogress.org | Office: 603.749.9102 | Fax: 603.749.6398

February 2012

Dear Fellow Educator –

NCLB drove a dramatic increase in testing at all levels. To some, we've gone too far. Although we're a testing company, we think educators can save time—and money—by more effectively using assessment to improve student learning. We're sponsoring this School Technology Action Report because we want to help educators do just that—do more with less. *Informed assessment practices* can efficiently shed light on student learning; without them, educators cannot determine the most appropriate next steps to strengthen instruction or promote growth.

Informed Assessment Practices (outcomes of high levels of "assessment literacy") include a coherent approach to (1) selecting or creating the right items, test, or activity for a purpose, and (2) appropriately understanding and using the evidence generated to fulfill that purpose. Most educators enter the profession without the training and skills in basic measurement principles—such as validity, reliability, and comparability—and other aspects of assessment. It, therefore, falls to professional learning and coaching to build their capacity in this body of knowledge and skills increasingly called for in professional and accreditation standards.

Measured Progress is a not-for-profit company. Founded in 1983, we've worked in 46 states, primarily focusing on customized, standards-based, large-scale assessment. We've also delivered assessment-related professional development through state and district programs. In 2010 we acquired DATAWISE, a California company with a student assessment and data management platform used by districts that serve more than 500,000 students. The platform, with an embedded item bank and professional development, supports effective student assessment throughout the year.

Over the years we've collaborated with our clients and others to develop and promote innovations in assessment that helped improve teaching and learning—

- being among the first to include performance measures—from tasks and events to portfolios—in large-scale testing programs in order to provide educators with richer insights into student learning;
- pioneering ways to accurately assess students with severe and complex disabilities; and
- promoting formative assessment as a set of classroom *practices*.

We continue to support improved student outcomes by developing new online tools, building new assessment content for the Common Core State Standards—including performance-based and technology-enhanced items—and expanding the impact of

measured progress

It's all about student learning. **Period.**

50 Education Way, Dover, NH 03820 | P.O. Box 1217, Dover, NH 03821-1217 | Web: measuredprogress.org | Office: 603.749.9102 | Fax: 603.749.6398

our groundbreaking assistive technologies (developed by Nimble Assessment Systems—which we also acquired in 2010).

Recent contract awards for Race to the Top initiatives and foundation-funded studies are spurring further innovations to help improve K-12 education. For example, we're

- helping to lead the way in developing test content, technology infrastructure, and accessibility supports for the "next-generation" Common Core assessments;
- creating new performance assessments to promote and track learning of higher-order skills, helping to build educator capacity to create and score them, and researching their use as an input to evaluating teacher effectiveness; and
- building interim Common Core assessment content and a flexible online item banking and test delivery system for districts and schools as part of a statewide program.

This School Technology Action Report includes some handy definitions, examples of common mistakes in using assessment, and articles addressing a range of relevant topics. We hope you find the report stimulating and useful. It might even prompt you to build greater capacity in *Informed Assessment Practices* in your district.

Sincerely,

Stuart R. Kahl

Stuart R. Kahl, Ph.D.
Founding Principal

Table of Contents **Page**

Assessment Literacy

Ensures the efficient and accurate use of assessment to boost student growth

Informed assessment practices, or "assessment literacy," encompass the knowledge and skills educators need to:

1. Identify, select, or create assessments optimally designed for various purposes, such as:

a) Accountability

b) Instructional program evaluation

c) Monitoring and/or promoting student growth

d) Diagnosing specific student needs (learning gaps)

2. Analyze, evaluate, and use the quantitative and qualitative evidence generated by external summative and interim assessments, classroom summative assessments, and instructionally embedded formative assessment practices to make appropriate decisions to improve programs and specific instruction to advance student learning.

Educators need not be experts in psychometrics to apply informed assessment practices. However, they do need to understand three key measurement principles:

1. Consistency of measurement—would you obtain the same results if you tested students on the same knowledge and skills again (reliability)?

2. The extent to which a test measures what is needed to provide the right information for the intended purpose (validity)

3. The legitimacy of test score comparisons (comparability—probably the most misunderstood concept in test results interpretation)

In general, teachers and administrators need similar levels of expertise in using assessment. While teachers must practice the knowledge and skills daily in their classrooms, administrators must (1) provide the appropriate opportunities for professional development and ongoing

collaboration to sustain this competency, (2) practice it at the school or district level, and (3) evaluate teacher practice for both formative and summative purposes.

The following examples illustrate common mistakes—opportunities where assessment literacy can help educators do more with less:

Vignette: A district curriculum specialist makes significant remediation decisions based on individual student subtest scores from commercial tests.

Vignette: A school administrator applauds the diagnostic value of a multiple-choice adaptive test with secure items.

Vignette: A teacher using a new online grading tool enters all the quiz and test scores given during the year and, looking at the score trends, determines whether students have demonstrated adequate growth during the year.

What teachers as assessors must know and be able to do

A variety of factors and initiatives seem to be converging to create a "perfect storm" of reform in education. Dissatisfaction with American students' performance on international assessments, concerns about U.S. global competitiveness, the state of the economy, and evidence that many high school graduates are not "college or career ready" are all taking their toll. And as a result, Race to the Top initiatives, the Department of Education's ESEA reauthorization blueprint, and plans of many states and state consortia all call for significant changes in curricula, instructional delivery, and the ways we monitor student achievement.

Education reform movements are nothing new. And there are some who would contend that past efforts have had little effect. Twenty-first century skill advocates point out that our current system of education was created to address the workplace needs of an emerging industrial nation—to turn out people who were armed with some basic, low-level skills and ready to take their place on an assembly line prepared to arrive on time, respect authority, and conform to established rules. They would assert that little about this system has changed since that time, despite the radical shift in the demands of the workplace.

New curricular emphases, programs, and instructional techniques have come and gone. But rather than seeing gargantuan improvements in student achievement, the small gains in large-scale assessment results have been far from adequate. If one accepts that interactions between students and teachers are the key to significant improvements in student achievement, then it becomes obvious where we should focus attention—on teaching and testing practices that have been shown to lead to real improvement.

Without shortchanging content, teachers in the future will be expected to better address a broad range of student skills, some cognitive (problem solving, critical thinking, communication) and some not (collaboration, self-direction). They will be expected to place greater emphasis on project-based learning and assessment leading to multiple, scorable student products and performances. Done well, these activities can lead to greater depth of knowledge of content.

The learning environment will not be limited to a school building or classroom, but instead will make greater use of out-of-school resources. Computers and other technology tools will be relied on extensively in all aspects of teaching and learning. Thus, teachers will have to be comfortable with changing learning environments and proficient with new, high-tech tools and systems. (Partnership for 21st Century Skills, 2009).

Assessment literacy

With all these changes, however, there is still something critical that teachers will need; something they have been lacking for some time – a far higher level of assessment literacy. They need a great deal more grounding in the use of assessment than the limited exposure to testing concepts they receive in pre-service training. I'm not talking about more definitions of such things as stanines or percentile ranks, but rather a far deeper understanding of the roles many kinds of assessment play in the processes of teaching and learning. Consider this response, which a teacher today might offer to answer a question about testing and grading practices.

> *I do formative assessment. I give quizzes or tests almost every day. I create the tests online from an item bank, and my students take the tests online, too. That way I get results back immediately and can use them to adjust my instruction. The scores are automatically recorded in my electronic grade book, and I can see right away how well each student did, as well as how the class did as a whole. My district gives two interim assessments each year. These are developed by our curriculum coordinator working with teachers, also using the item bank. These are general assessments we use to monitor growth and to identify students who are likely to have trouble passing the state test at the end of the year. We also use the diagnostic information they give us.*

On the surface, these comments may seem quite reasonable. But they may well depict poor practice. For example, there is a significant disconnect between the teacher's concept of formative assessment and the dramatically effective process of formative assessment supported by research. The latter is an ongoing process that occurs during instruction and involves (1) letting students know the learning targets and criteria for success, (2) gathering rich evidence of student learning by a variety of means (e.g., observation, questioning, quizzes), (3) providing descriptive feedback on gaps in student learning, (4) the teacher and student using the feedback to adjust instruction and learning activities, (5) student self-assessment, and (6) activating other students as resources (Wiliam, 2007).

Back to our teacher's response, timing (immediacy of results) is only one attribute of effective formative assessment. A score on a multiple-choice quiz hardly constitutes rich evidence or descriptive feedback leading to appropriate changes in instruction. In fact, the assignment of scores to many kinds of student work before the completion of an instructional unit is one of many grading practices that destroy students' motivation to learn and thus inhibit learning (Schafer, 1993).

The district testing the teacher describes also seems reasonable on the surface. Early warning and growth monitoring are legitimate uses of interim testing. With respect to the latter, I wonder if

the test items were selected for the two tests in such a way that comparisons of performance on the two measures are appropriate. It is doubtful that the two tests were statistically equated. Were raw mean scores compared? If percentage of proficient students was reported, were the cut scores for proficiency arbitrarily set at 70 percent on both tests? In either case, what if the second test was just easier than the first—would a higher score on the second one really be an accurate reflection of growth?

A school administrator once mentioned to me that the district was looking forward to implementing a data management system, so that results from multiple tests could be aggregated to help the teachers better understand their students' capabilities. With respect to total tests or subtests, how would the content covered by different measures compare? Are the results reported on the same scales? If not, how can they be aggregated? Does it make sense to aggregate data gathered months apart? For monitoring growth with respect to a general area or specific standard, are the measures comparable, based on content and difficulty?

Whether using self-developed tests or off-the-shelf instruments from the publishers, these are the kinds of questions to which district educators need to know the answers. Those answers, known in fact by too few, determine the tests' legitimate uses, as well as what legitimate conclusions can be drawn from the results.

Assessment types and uses

There are several categories of assessments that are used in schools today, and several approaches that might be used within each. Very different from the process of formative assessment described earlier are summative assessments, which could include teacher-made classroom tests, interim assessments like the district tests the teacher described above, and high-stakes external tests, such as state accountability assessments. Summative assessments are "those assessments that are generally carried out at the end of an instructional unit or course of study for the purposes of giving grades or otherwise certifying student proficiency" (Shepard et al, 2005).

Some of these tests might be general achievement tests covering the whole domain of mathematics at a grade, for example, or benchmark tests, perhaps covering material taught within the last two or three months.

There are tests made up of multiple-choice questions, tests made up of constructed-response questions, and tests made up of combinations of item types. (Generally, extended constructed-response questions are better for testing higher-order thinking skills or greater depth of knowledge.) There are fixed tests (the same tests taken by all students in a group) and computer-adaptive tests, which are tailored to each student's ability level. General achievement measures,

whether fixed or adaptive, are not designed to provide rich, diagnostic information. They can be used to monitor growth. Also, they are quite useful as a source of information to guide program improvements that will benefit the next group of students to pass through a tested grade–such as general areas of weakness within a discipline or identification of low-performing subgroups of students.

Conclusion

It is true that teachers of the future will need to deal with many changes in education–including new environments and new tools. But these changes won't lessen the need for a much greater level of assessment literacy, here defined as the knowledge and skills teachers need to:

- Identify, select, or create assessments optimally designed for various purposes, such as: grading or certifying proficiency, diagnosing specific student needs (gaps in learning), and assessing higher order thinking; and
- Analyze, evaluate, and use the quantitative and qualitative evidence generated by external summative and interim assessments, classroom summative assessments, and instructionally embedded formative assessment practices to make appropriate decisions to improve programs and specific instruction to advance student learning.

Better equipped with assessment literacy, teachers will be in a much better position to weather the "perfect storm of reform" and maximize student learning.

References:

Partnership for 21st Century Skills. (2009) 21st century learning environments. White paper from series on support systems, http://www.21stcenturyskills.org/route21/.

Schafer, W. (1993) Assessment literacy for teachers. *Theory into Practice*, 32(2), College of Education, The Ohio State University.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F. (2005) Assessment. In Darling-Hammond, L. and Bransford, J. (Eds.), *Preparing Teachers for a Changing World*, 275-326, San Francisco: Jossey-Bass.

Wiliam, D. (2007) Keeping learning on track: Classroom assessment and the regulation of learning. In Lester F. (Ed.), *Second Handbook of Mathematics Teaching and Learning*, 1053-1098, Greenwich, CT: Information Age Publishing.

Innovative assessment: Effective and ineffective assessment reform

Disappointing performance of U.S. students on international tests, concerns about the nation's global competitiveness, questionable readiness of our students to enter college and the workforce, and growing dissatisfaction with the assessment and accountability requirements of NCLB have all led to another wave of efforts to significantly reform American education. Unfortunately, policy makers peck away at things they can address easily – common state standards, increased NCLB flexibility with respect to growth models, monies for data management systems, etc. – but all these fail to address the problem at its source, the classroom.

The approach to schooling that dominates precollege programs has existed for well over a century. This approach served us well when the nation was at the height of industrialization. However, in this very different age, our efforts to assure that no child gets left behind have done little to eradicate practices that seem to be more consistent with an orientation toward no child getting ahead. The federal government's obsession with grade-level curricula and accountability requirements has seriously stifled efforts to prepare our students for a 21st century workplace—and have pushed assessment back 25 years to an era of minimal competency and basic skills testing. Most high school students are bored in school, not motivated to learn, and not encouraged to think critically or creatively and work collaboratively. Dropout rates are all too high for this diverse, new breed of digital natives.

For good or bad, assessment practices (both local and large-scale) have a significant impact upon the attitudes, behaviors, and practices of students and teachers. And in a time of technology-driven change, it is especially important to ensure that the educational reforms we undertake do indeed advance the cause of improved educational practice and outcomes, rather than solidify poor practice and inhibit progress toward higher levels of achievement.

Background–balanced assessment systems

There are many types of educational assessments, which have different characteristics and address different purposes. Unfortunately, the state of assessment literacy among policy makers and decision makers at various levels of the educational enterprise is not as it should be. Many tests are used for purposes they were not designed to address. This leads to unjustified conclusions from results and subsequently to ineffective, even damaging, follow-up action. While this paper does not purport to describe all types of tests available to educators, the description below of the components of a district's balanced assessment system should help frame the discussion of effective and ineffective assessment reforms. In their own ways, assessments in each of the four categories can improve teaching and learning

Classroom formative assessment

Formative assessment is a process by which teachers and students assess learning during instruction in a topic or skill so that, if necessary, they can make instructional adjustments to improve learning. The process includes letting students know the learning targets and criteria for success; gathering good evidence for diagnosing students' misconceptions; providing descriptive feedback; and making appropriate adjustments to instruction. Peer and self-evaluation play important roles, too.

A large body of research indicates that effective formative assessment practices have dramatic effects on student achievement. Formative assessment is a process, not a test. Evidence gathering approaches can include a wide variety of techniques, such as questioning, observations of class work, quizzes, projects, etc. The best evidence is actual student work, which reveals student thinking. Formative work is typically not graded, because it is gathered before instruction on the topic is completed and thus, before the student has reached the level of mastery he/she will achieve by the end of instruction.

Classroom summative assessments

These classroom assessments determine and document students' level of achievement upon completion of a unit, a marking period, semester, etc. They include tests, papers, and projects that typically are counted toward grades. Although more can be done with the results of classroom summative assessments, seldom is that the case. Their results could also be used in much the same way as those from external assessments described below.

External interim assessments

Many districts use tests, either commercially produced or ones they construct from item banks, at some midpoint during the year. Their purpose might be as an early warning, to predict performance on subsequent high-stakes summative tests and identify students at risk of failing the latter. They might also be used as benchmark tests, given a few times during the year to measure performance with respect to the material covered by instruction since the previous interim test.

External summative assessments

High-stakes statewide accountability tests, as well as national and international tests, fall into this category. In addition to their use in NCLB-required accountability reporting , the results from state tests (and interim tests, too -- see above), particularly subtest and subgroup scores, can also

help identify areas in which schools' instructional programs can be improved for the next group of students to pass through the tested grades.

Shortcomings of current assessment systems

Obsession with scoring/grading at the local level

It is common practice these days for teachers to grade *everything*. This includes work that is done early in the learning process, before the students have reached the level of competence they will ultimately attain after instruction on a topic is completed. Non-cognitive behaviors are graded, as well.

Technology is an accomplice in this matter. Electronic grade books allow teachers to enter scores of grades on every student every week; and with weightings of different types of entries predetermined, teachers can compute overall grades at any time at the push of a button. It is not unusual for a parent to see a full-page printout of scores a child earned in a single week and the course grade the student would receive if the marking period were at an end.

Some teachers claim that if they didn't grade everything, students would never do what's expected of them. They also use grades as a primary means of maintaining discipline. These notions are contrary to the principles of formative assessment, which stress motivating students to learn and take responsibility for their own learning. While switching to not grading formative work may seem problematic at first, once students understand that their grasp of all the material addressed via formative assessments is crucial to their performance on subsequent measures that do count, they ultimately fall in line with the new practice. Learning will be the motivator, rather than the grade and a mindset that says that 80 percent is good enough and there's no need to understand the other 20 percent of the content tested.

Research bears out the fact that many common grading practices actually inhibit learning. In addition to grading formative work, for example, such practices include assigning zeroes to incomplete work (a grade average killer) and assigning high scores (albeit giving them little weight) for completed homework without regard for the quality of the work and absent any formative use of it at all.

Mismatch between components of an assessment system and good instruction

There's a lot of talk about the different components of an assessment system not being aligned. Of course, all of the components should be aligned with the content standards in a subject, and it is almost certain that this alignment is lacking in a lot of places. It is also apparent that the

components are becoming more and more aligned with one another in terms of the nature of the assessments. While such alignment is desirable, currently it too is a problem. NCLB requirements for statewide testing at so many grades and the quick turnaround times for results demanded because of the parental choice option, have led many states to discontinue the use of non-multiple-choice formats in their assessment programs. Research in the early 1990s documented the negative impacts on instruction of the use of all-multiple-choice tests for high-stakes, accountability testing. A lot of higher-order content standards are not covered by these tests. More important, however, is the impact on all other components of a district's assessment system.

Interim assessments from commercial sources tend to be all multiple-choice, due in part to the use of online test delivery systems. While online testing can include constructed-response items, the demand for immediate results often leads to all-multiple-choice tests. Item banks made available to school districts for interim and classroom summative tests also tend to be exclusively multiple-choice. Several vendors even go so far as to claim that their tests and item banks can "satisfy all your formative assessment needs." This leads to over-interpretation of item-level results for individual students and misguided remedial efforts. Clearly, multiple-choice items do not yield the rich information on students' thinking that actual student work can offer. In addition, student work provides the basis for the descriptive feedback and appropriate instructional adjustments that are central to effective formative assessment.

Teachers, understandably, try to make their tests emulate high-stakes tests. This practice would be fine if the high-stakes tests modeled good classroom testing. Unfortunately, that is generally not the case. Furthermore, not many people can write good multiple-choice items, so typical classroom tests expose students to items that are ambiguous or which focus on low-level cognitive skills. If teachers would simply drop the options from their test items, they and their students would be better served. Short-answer questions can be quickly scored, and offer teachers the opportunity to see revealing student work. Of course, more substantial, higher-order, constructed-response tests would be desirable, too.

Classroom assessment, both formative and summative, plays a critical role in the achievement of students. Formative assessment, as described earlier in this paper, is simply good instruction. Clearly, the components of an effective balanced assessment system should be consistent with one another—and in ways that encourage good instruction.

Mismatch between learning environments and good assessment

Two current hot topics in education are formative assessment and the teaching of 21st century skills. If the education community could make big strides in these areas, achievement levels of

our students could increase significantly and our students would be better prepared for college and the workplace. Unfortunately, many factors prevent our schools from making the desired progress. The nature and impact of high-stakes testing have already been mentioned, but there are many more inhibitors.

When teachers are first introduced to all the characteristics of effective formative assessment, many of them are overwhelmed and claim they just don't have that much time to give to assessment. Advocates of true formative assessment recognize that what is required in many schools is a whole new mindset, a culture of learning that is manifested in different roles and activities for teachers and students. Of course, as described earlier, formative assessment is just good instruction – gifted teachers have been doing it well for years. However, the individualization that is required does mean that for most teachers to become proficient formative assessors, they will have to use their time differently.

Twenty-first century skills include higher-order cognitive capabilities, such as critical thinking, problem solving, and communication, as well as some workplace competencies, like collaboration, teamwork, innovation, and leadership. (They also include literacy in areas such as entrepreneurship, finance, global awareness, and information technology.) To assess many of these skills effectively requires performance assessment. In measurement parlance, performance assessment refers to almost any assessment approach that is not fixed-response. Short-answer or constructed-response questions are examples of performance assessment. However, when people speak of performance assessment today in the context of 21st century skills, they are referring to more substantial activities.

The most common example of such performance assessment in education is direct writing assessment–the administration of writing "prompts," which elicit essays or other forms of extended student writing. Other performances include oral presentations, research projects yielding scorable products, and the like. Unfortunately, administering and scoring such measures can take a great deal of time and effort. Neither is accommodated in the typical school environment.

Lots of things have to change if sufficient progress is to be made with respect to formative and performance assessment—grading practices, time allocations for both class work and out-of-class time for teachers, and mindsets. An individual teacher cannot make these changes alone; they require whole-school involvement and the active and total support of school leadership. Continuous professional development of teachers and administrators is critical; large-scale assessment and accountability requirements that encourage good practice are, too. A later section of this paper presents a vision of how all this might come together.

The promise and pitfalls of technology-based assessment

It is obvious that technology will play a significant role in the schools of the future. But just as nuclear energy could be used for good and for bad, so too can technology be used productively and unproductively in education.

Multiple-choice emphasis

The major use of computers in educational assessment seems to be to deliver multiple-choice tests, which, as mentioned earlier, tend not to cover many higher-order content standards. Many state assessment programs have moved partially to computer-based testing. In addition, commercial interim tests and tests school personnel create from item banks are delivered by computer. The quick turnaround of results associated with computer-delivered tests may be very appealing, but the downside is that some vendors overemphasize the "formative" value of these tests. In truth, many of these instruments don't address in detail the topic being taught at the time of testing or provide the type of rich, diagnostic information that leads to effective feedback.

For statewide tests, administered in a narrow window of time and with stringent security requirements, significant challenges arise due to the uniqueness of every school's software and hardware configuration, as well as to the ever-changing third party software the test delivery systems employ. For smaller special populations, computer-delivered tests, which offer accommodations customizable to the individual student (large-print, signing, translation, text-to-speech conversion, for example), have proven very effective.

At the local level, "clickers" and Fly Pens (FLY Fusion™ Pentop Computers) are being used more and more. Questions are projected on a screen in front of a group of students who use clickers to submit their answers. Immediate results are provided after each item is administered. Needless to say, while this technology has some utility, it is not being used to address higher-order skills. The Fly Pen is a computerized pen that, when used with special pad paper, "knows" exactly where it is on a page, so that it can digitize images it leaves on the page. Students can take multiple-choice tests with the pens (downloaded to print by the teacher) by entering a test identification number, then simply writing the letters of the multiple-choice options they choose. The test data from an individual student are captured, results can be provided immediately by the pen to the student, and the data from all the students' pens can be downloaded and aggregated by the teacher. (Since the Fly Pen captures any images the students produce on the special paper, it would be interesting to see what applications of this technology could be made to do constructed-response testing on a larger scale.)

The quick turnaround of results is primarily associated with multiple-choice testing, largely because of the additional time required for the scoring of students' responses for other item types. Artificial intelligence (AI) scoring is being used in some testing programs, particularly for the scoring of writing samples. For high-stakes testing, it is more likely to be used for "second reads" with discrepancies between human and AI scores resolved by human third readers. The jury is still out on the use of AI for the scoring of constructed responses in the content areas.

New item types

In statewide testing, it is rare that all students can take a test by computer; therefore, both computer-delivered and paper tests are used. Because of comparability concerns, computer-delivered tests are designed to mimic the paper tests. Thus, states have not been able to benefit from the capability of computers to administer items making use of techniques such as drag-and-drop, hot spots, and animation. One must question, however, whether items using these approaches really address knowledge or skills that cannot be assessed readily by traditional multiple-choice items. This is not to say that the new approaches shouldn't be considered, but they may not be worth the expense of developing if they don't accomplish more than traditional formats.

Simulations are being used for some testing, particularly in science. For example, in assessing skills with respect to scientific investigation, simulations can eliminate the need for real materials and the wait for reactions to take place or plants to grow. However, the costs associated with taking simulation testing to scale may be prohibitive for some testing programs.

An alternative use of computers in testing

Clearly, computers will become the primary tool for test delivery in the future—both for large-scale and local testing programs. In the meantime, there are other computer applications that can and should play a role in the assessment of students. It is important that we look carefully at the knowledge and skills we value, those embodied by the content standards we are using or developing, and decide how best to measure each. If we are to make strides toward the effective assessment of many 21st century skills, then we have to turn again to performance testing.

During the authentic assessment era of the 1990s, there were states that successfully implemented performance assessments on a large scale. However, an unready public (including some in the measurement community), politics, and some less-than-sterling examples of good practice led to the decline of interest in authentic performance assessment. NCLB was the nail in the coffin. Now, despite the lack of resources due to a suffering economy, performance

assessment's time has come again. A lot of lessons were learned during the '90s. We know how to do it.

Some of the past efforts at performance assessment that were unsuccessful failed to exhibit two essential attributes. First, the activities, whether on-demand performance tasks or longer projects, were often not closely tied to content standards. Loose connections to general categories of standards would not help advance learning or documentation of learning of the content and skills reflected in a state's or district's standards. Second, some failed attempts did not include rigorous, high quality measurement of performance. These two attributes – close alignment to standards and good measurement – are absolutely necessary in renewed efforts to bring performance assessment to the forefront of assessment and accountability.

Performance assessment, particularly for large-scale assessment programs, can be time consuming and expensive. For this reason, there are many who support the notion of locally administered and scored, curriculum-embedded performance assessments contributing to state accountability results. This, along with a scoring audit process, is probably the most feasible approach to bringing performance testing to scale and at the same time having a positive impact of large-scale testing on curriculum and instruction, rather than the opposite.

So what is the role of computers in this scenario? There are multiple roles. First, if performance assessments are curriculum-embedded, they are probably associated with longer-term activities or projects, which could well involve students in online work – either research or finding resources required for particular tasks within the project. Research efforts associated with projects could include information gathering from out-of-school resources, such as museums and other local attractions or businesses. Designing projects around such resources, projects that may require considerable pre-field-trip and post-field-trip activity, could make the excursions much more than fun and games—more valuable as learning activities.

Critical to the success of project-based learning are scorable products students generate, whether they are working alone or on teams. These products can take the form of written work (using word processing software), oral presentations (using Power Point or the computer's graphic capabilities), and even images of other kinds of products. Finally, electronic portfolios can be the repository of written documents, images, video clips, etc. This technology would facilitate the submission of student work for centralized auditing of scores. Thus, for assessing many 21st century skills, it may be desirable to turn Internet access and other computer applications on, rather than off, as is typically done during traditional testing.

A vision of reform for schooling and assessment

In the authentic assessment era, there was a great deal of conversation around "removing the barrier" between instruction and assessment. Yet it could be argued that as performance assessments were conducted in that era, those barriers were never really removed. As suggested earlier in this document, effective formative assessment and performance testing have the potential to break down those barriers and significantly raise performance levels. However, they require changes in our schools that go beyond merely convincing individual teachers to try something new.

It's not hard to find futuristic descriptions of 21st century schools that challenge the very image of schools as we know them today. With the ever-increasing development and use of new technologies, their time will come. However, technology is not a panacea—it enables both good and bad practices. Also, change in American schools comes very slowly. There is much we can do now that may not be so revolutionary, but which makes use of a vast body of educational research that has been conducted over many decades. In fact, it is astounding in this era when everything is supposed to be "research-based," how much educational practice ignores findings of very conclusive research.

We know what it takes to bring about significant change in schools, and the role school leaders should play in that process. We know the value of sustained, on-the-job professional development for teachers; yet one-shot professional development presentations during twice-yearly in-service days still predominate. We know which grading and feedback practices enhance learning and which ones inhibit learning. We know about the impacts of high-stakes testing on instruction. We know how to conduct meaningful performance assessments and reliably score the results.

The bullets below describe what some aspects of effective schooling might look like today—not in the distant future. These research-based practices aren't new, but they are not nearly as widely implemented as they will have to be if achievement levels of American students are to be raised significantly.

What it might look like:

- Teachers are proficient formative assessors. This means their students clearly understand the learning targets and criteria for success; the teachers use a variety of evidence-gathering techniques that produce rich information about students' capabilities relative to the targets; the teachers provide descriptive feedback on the students' work and make appropriate decisions about instructional adjustments or moving on to new learning targets.

- Teachers' grading practices reflect student achievement and are not designed primarily to keep students in line or make them do their work. Instead of inhibiting learning, grading practices motivate students and make them responsible for their own learning. Formative work is not graded, incomplete work is completed, and tests that count are not surprises, nor do they contain surprises.
- Classroom assessment, formative and summative, makes minimal use of multiple-choice items.
- The teacher's role is not to serve as the source of all information for student learning, but rather as a facilitator and monitor of learning, a process in which the students play active roles. Self and peer assessment play significant roles in the classroom.
- Core knowledge and skills are not ignored, but rather applied through the use of short- and long-term projects and performance assessments. Block scheduling can be used to allow the introduction and review of core knowledge and skills, at the same time providing time for students to work alone or in teams on projects that are closely tied to standards and that lead to scorable products or performances. Through these projects, students make use of a variety of 21st century skills, such as problem solving, critical thinking, communication skills, teamwork, collaboration, and media skills.
- Several times a week teachers spend time with colleagues reviewing and discussing actual student work and subsequent next steps for instruction. Time for such "communities of learning" enables newer teachers to benefit from the content knowledge and pedagogical skills of more experienced teachers.
- Principals are instructional leaders and leaders of reform efforts, where needed. Teachers cannot independently change grading practices, time allocations, etc. Principals take part in communities of learning, and monitor and guide teachers' practices relative to formative assessment, testing, and instruction generally.
- Principals make good decisions regarding choice of interim assessments and use of data from interim assessments and accountability testing. They exhibit a high level of assessment literacy in making these decisions, and they and their teachers avoid succumbing to "quick fixes," such as using testing products for purposes for which they are not designed or able to address.
- When significant changes are necessary to implement the practices described above, principals get help, perhaps in the form of coaches or consultants who can help them lead their reform efforts; who know how to "phase in" changes over time, so that they are not overwhelming to their teachers; and who know how to create a culture of learning in a school. Educator development is continuous, relying heavily on in-school communities of learning and coaching provided by internal and/or external sources.
- State assessments include multiple-choice and constructed-response components, as well as curriculum-embedded performance assessments with local scoring and centralized audit procedures. They model good assessment practice. Federal laws and their implementation are changed to enable and encourage good practice at the state and local levels.

Equity

Years ago, there were concerns that certain assessment formats were inherently biased against minority populations. These concerns were largely the result of differing performance of subgroups of students on tests. However, improved understanding of the nature of test bias has quieted that concern somewhat. Test bias may exist when identifiable subgroups, equated on ability, perform differently on a test. Even with this clarification, differing performance can relate to many factors, including motivational ones.

Nevertheless, significant performance differences persist among various subgroups of students, and federal mandates intended to reduce or eliminate them have been far from successful. However, characteristics of assessments, as described in this paper, have the potential to improve the situation. For example, effective formative assessment is individualized in that by this instructional process, students' individual misconceptions are identified and appropriate instructional adjustments are made. Students proceed to summative assessments relative to a learning target when they are ready. This is vastly different from the prevalent practice of moving whole classes on to the next lesson or unit whether everyone is ready or not. Student motivation is a critical factor in formative assessment, also known as assessment for learning. Of course, as explained previously, formative assessment (and the individualized instruction it embodies) requires teachers and students to play very different roles than they have in traditional instruction.

Also critical to student motivation are the learning activities in which students are engaged in and out of school. Project-based learning has the potential to motivate students to learn far better than many traditional instructional approaches. Again, however, close ties to content standards and good measurement are important features of projects. From an equity perspective, project-based learning can offer choices of projects, roles within projects, and even modes of demonstrating learning (performance assessment). Such choices allow individuals to play to their strengths and interests and motivate them to learn and take responsibility for their own learning. Student motivation, aspirations, and expectations are all keys to addressing equity concerns in education.

Raising the bar

Policy makers can mandate improvement, require more testing, encourage common state standards, fund data management systems, allow growth models, and on and on; but improved performance will be the result of what happens in our classrooms. The authentic assessment era's hope of breaking down the barrier between assessment and instruction must become a reality, and formative assessment and performance assessment may be the keys to this happening. Thus, innovations in assessment and related used of new technologies must be directed toward these activities and to the learning environments and teacher support that can enable them.

A vision built on educational research and successful practices

An analysis of common elements of effective assessment systems in the United States and abroad reveals several key themes:

1).The student assessment process is guided by common standards and grounded in a thoughtful, standards-based curriculum. It is managed as part of a tightly integrated system of standards, curriculum, assessment, instruction, and teacher development.

Large nations like Australia, Canada, and China manage curriculum and assessment at the state or provincial level, while small nations like England and Singapore—which have school populations about the size of California and Kentucky, respectively—have national systems managed by a ministry of education. Each of these jurisdictions has undertaken a careful process of developing standards (generally described as curriculum expectations) and curriculum guidance, often in the form of syllabi, to guide teachers' instruction in the classroom, as well as professional development that is organized around the curriculum.

- Curriculum guidance is lean but clear and focused on what students should know and be able to do as a result of their learning experiences. Assessment expectations are described in the curriculum.
- Curriculum and assessment are organized around a well-defined set of learning progressions along multiple dimensions within subject areas. These guide teaching decisions, classroom-based assessment, and external assessment.
- Teachers and other curriculum experts are involved in an extensively vetted curriculum development process and in the process of developing assessment measures grounded in the curriculum standards. These guide professional learning about curriculum, teaching, and assessment. Thus, everything that comes to schools is well aligned and pulling in the same direction.

2).A balance of assessment measures that includes evidence of actual student performance on challenging tasks that evaluate applications of knowledge and skills.

The curriculum and student assessment process seek to teach and evaluate knowledge and skills in authentic ways that examine a broad array of skills and competencies and generalize to higher education and multiple work domains. They emphasize deep knowledge of core concepts within and across the disciplines, problem solving, collaboration, analysis, synthesis, and critical thinking. As a large and increasing part of their examination systems, high-achieving nations use open-ended performance tasks and school-based, curriculum-embedded assessment to give students opportunities to develop and demonstrate higher order thinking skills such as the

abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, and apply learning to new situations. The curriculum and assessment systems evaluate students' abilities in a variety of tasks such as projects, group work, open-ended tasks, and oral presentations. The system would also employ summative measures such as examinations that include essays and open-ended tasks and problems, along with tests using selected-response (multiple-choice) items, usually given at the end of a course or year.

3).Teachers are integrally involved in the development of curriculum and the development and scoring of assessment measures for both the on-demand portion of state or national examinations and local tasks that feed into examination scores and course grades.

Most successful systems in the U.S. and other high-achieving nations invest in extensive moderation to ensure an accurate, reliable, and consistent scoring process and enable teachers to deeply understand the standards and develop stronger curriculum and instruction. The moderated scoring process is a strong professional learning experience, and as teachers become more skilled at using new assessment practices and developing curriculum, they become more effective at teaching the standards. The assessment systems are designed to increase the capacity of teachers to prepare students for the demands of college and careers in this new century and global society.

4).Assessment measures are structured to continuously improve teaching and learning.

Assessment ***as***, ***of***, and ***for*** learning is enabled by several features of successful assessment systems:

The use of school-based, curriculum-embedded assessment (more complex assessment exercises that all students are expected to complete over an extended timeframe) provides teachers with models of good curriculum and assessment practice, enhances curriculum equity within and across schools, and allows teachers to see and evaluate student learning in ways that can feed back into instructional and curriculum decisions.

Close examination of student work and moderated teacher scoring of both school-based components and externally developed open-ended portions of examinations are sources of ongoing professional development that improve teaching.

Developing both school-based and external assessment measures around learning progressions allows teachers to see where students are on multiple dimensions of learning and to strategically support their progress.

School-based, curriculum-embedded assessment engages students in their own learning process and builds their capacity to assess their own learning.

5).Assessment and accountability systems are designed to improve the quality of learning and schooling.

The student assessment process produces evidence of learning. That evidence is critical information for informing the learning process for both the student and teacher and for informing decision makers about the quality of the educational program and the accountability of the personnel who are responsible. But there must be a balance in the system between these two uses of the evidence. The need for accountability using large-scale, high-stakes, summative assessments should not overshadow assessment's primary purpose of providing timely feedback to the teachers and learners engaged in the instructional process. The interval of time between when the evidence is produced and when it is used to alter the course of instruction is crucial to improving the quality of the learning. A shorter time interval increases the value of the information used to modify the learning process.

High-achieving states and nations invest most of their resources in high-quality assessments that aim to drive the learning of ambitious intellectual skills in the classroom. In order to maintain investments in well vetted expert processes of development and scoring, most countries implement external tests for students only once or twice prior to high school (generally around grades three and six), with continuous school-based assessment throughout these years.

High school examinations in high-achieving nations are generally selected from an array of subjects by students to demonstrate their areas of competence for colleges and employers. These assessments also inform course grades, support individual student learning, and shape curriculum improvement. The tests are typically not used to determine student graduation from high school; they set a higher standard linked to college and career expectations.

High-achieving states and nations implement accountability systems that publicly report outcomes and take these into account, along with other indicators of school performance, in a well-designed system focused on continual improvement for schools. Many nations combine assessment data with information from school inspections to design intensive professional development supports and interventions that improve school performance. Many of these inspectorate systems use experts to examine teaching, learning, and school operations up-close in order to diagnose school needs and guide more targeted improvement efforts.

6).Assessment and accountability systems use multiple measures to evaluate students and schools.

High-achieving countries use multiple measures (multiple sources of evidence of varying types) to evaluate skills and knowledge needed for the demands of this dynamic, technological era. Students engage in a variety of tasks and tests that are both curriculum embedded and on demand, providing many ways to demonstrate and evaluate their learning. These are combined in reporting systems at the school and beyond the school level. School reporting and accountability are also based on multiple measures, including student achievement as one indicator among many. Other indicators often include student participation in challenging curricula, progress through school, graduation rates, college attendance, citizenship, a safe and caring climate, and school success and improvement.

7).New technologies enable greater assessment quality and information systems that support accountability.

New technologies enhance and transform the way the assessment process is developed, delivered, and used, providing adaptive tools and access to information resources for students to demonstrate their learning, and appropriate, immediate feedback by supporting both teacher scoring and computer-based scoring.

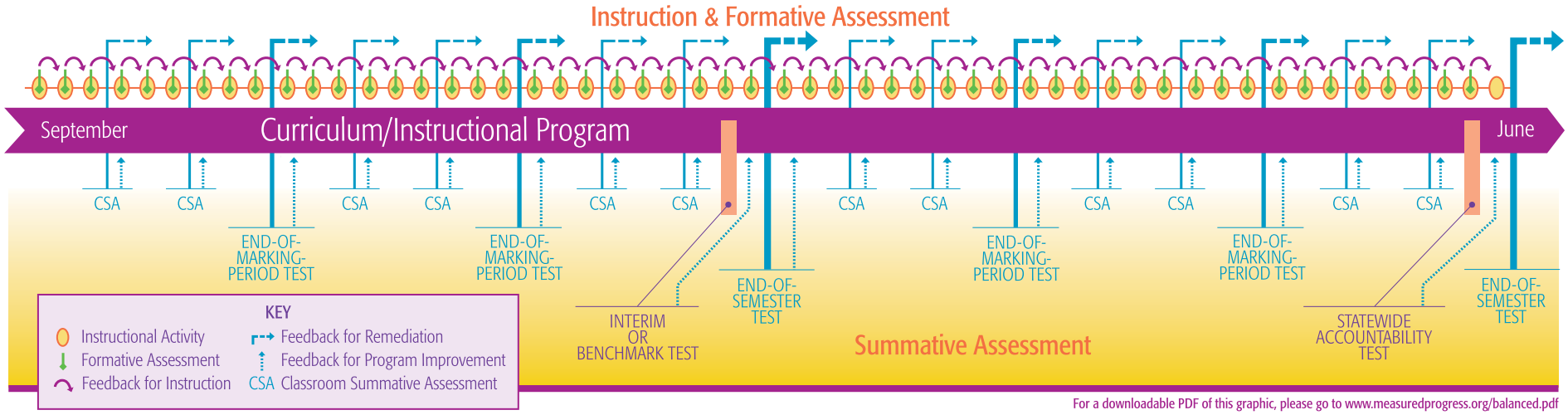
Technology also organizes data about student learning, enhancing system accountability for instruction and reporting by providing more efficient, accurate, and timely information to teachers, parents, administrators, and policymakers. In the current U.S. context, technology can help to integrate information at all levels of the system as part of a longitudinal state data system, contributing to a rich profile of accomplishment for every student.

By applying these lessons as well as new knowledge from the leading edge of assessment development, we can imagine a systemic approach to transforming assessment of learning in the United States.

Stuart Kahl, Ph.D.
President & CEO
Measured Progress

A Balanced Assessment System: A Different Perspective

Traditionally, instruction, assessment, and remediation are represented graphically as a circular process. Through all my musings about the correct uses of different types of assessment, I came to question that model. The school year is linear, so it's more appropriate and realistic to view assessment as part of the curricular/instructional time line. This graphic depicts where different types of assessment, both summative and formative, might fit in a balanced assessment system. Both forms of assessment are aligned with curriculum, which is derived from an established set of content standards. Formative assessment is embedded in and informs instruction, while various summative assessments inform remediation and program changes. Used correctly, both kinds of assessments improve teaching and learning, which should be the ultimate purpose of testing.



For a downloadable PDF of this graphic, please go to www.measuredprogress.org/balanced.pdf

Classroom Formative Assessment
- Includes instructionally embedded activities, e.g., observation, homework, quizzes, questioning
- Usually constructed locally
- Yields rich diagnostic information
- Happens while material is being taught
- Leads to instructional decisions
- Isn't used for grades

Classroom Summative Assessment
- Occurs after material is taught
- Includes unit tests or other graded performances
- Can be constructed locally or provided with curriculum materials
- Counts toward grades
- Isn't diagnostic

Interim and Benchmark Summative Tests
- Can be used as an early warning of performance on later high-stakes test
- Often constructed by external sources
- Can cover some or all of school year curriculum
- Provide broad domain or sub-domain coverage (not diagnostic)
- Subtest and subgroup results raise programmatic questions that require further investigation

Statewide High-Stakes Accountability Tests
- Provide broad domain or sub-domain coverage (not diagnostic)
- Constructed by external sources
- Subtest and subgroup results raise programmatic questions that require further investigation

Educational Measurement: Issues and Practice
Fall 2009, Vol. 28, No. 3, pp. 5–13

Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments

Marianne Perie, Scott Marion, and Brian Gong, *National Center for the Improvement of Educational Assessment, Inc.*

Local assessment systems are being marketed as formative, benchmark, predictive, and a host of other terms. Many so-called formative assessments are not at all similar to the types of assessments and strategies studied by Black and Wiliam (1998) but instead are interim assessments. In this article, we clarify the definition and uses of interim assessments and argue that they can be an important piece of a comprehensive assessment system that includes formative, interim, and summative assessments. Interim assessments are given on a larger scale than formative assessments, have less flexibility, and are aggregated to the school or district level to help inform policy. Interim assessments are driven by their purpose, which fall into the categories of instructional, evaluative, or predictive. Our intent is to provide a specific definition for these "interim assessments" and to develop a framework that district and state leaders can use to evaluate these systems for purchase or development. The discussion lays out some concerns with the current state of these assessments as well as hopes for future directions and suggestions for further research.

Keywords: comprehensive assessment system, formative assessment, interim assessment

First encoded in federal law as a result of the Improving America's Schools Act of 1994 (IASA), the standards-based reform movement has resulted in the widespread use of summative assessments designed to measure students' performance at specific points in time. Under IASA, testing was required at three grades: once each at the elementary, middle, and high school levels. The enactment of the No Child Left Behind Act (NCLB) of 2001 required increasing the number of these large-scale summative tests to every grade 3–8 and at least once in high school. Policymakers' goal for these assessments generally has been to measure students' attainment of the state content knowledge and skills against some defined level of performance, such as attaining the level of *Proficient* or *Distinguished* or simply *meeting the standard*. While many had hoped that these once-a-year tests would provide instructionally useful information, educators and others know this is not occurring. This is not because there is something "wrong" with these summative accountability tests; rather it is that they were not designed to meet instructional purposes. For example, these tests—by design—usually are administered as late in the year as possible and the results are returned after the students are home for the summer. In addition, the reports are designed to provide reliable total score and performance level information for each student across a wide range of content within a minimum of testing time, at low cost, under standardized conditions common to the whole state. This design precludes these general survey tests from providing useful diagnostic information for individual students. Therefore, educators and policymakers have realized that other forms of assessments are necessary to inform instruction during the school year.

This need for measuring student performance throughout the year has resulted in a rapid influx of products.

Marianne Perie is a Senior Associate at the National Center for the Improvement of Educational Assessment, P.O. Box 351, Dover, NH 03821; mperie@nciea.org. Scott Marion is the Associate Director, and Brian Gong is the Executive Director, of the National Center for the Improvement of Educational Assessment.

Many vendors are marketing assessments to states and districts that they call "benchmark," "diagnostic," "formative," and/or "predictive" with promises of improving student performance and helping schools and districts meet the federal NCLB requirements or increasing pass rates on high school exit exams. All of these terms fit under the umbrella term "interim assessment." A good interim assessment can be an integral part of a state's or district's comprehensive assessment system used in conjunction with classroom formative assessments and summative end-of-year assessments. Unfortunately, there is little research indicating that many of these commercially available interim assessments positively affect student achievement. Furthermore, vendors for many of these products cite research on classroom formative assessment (e.g., Black & Wiliam, 1998) implying that their assessments will improve student learning even though few, if any, of these commercial products are the types of assessments or activities described in the Black and Wiliam (1998) meta-analysis. There is a growing concern among researchers and educators that states and districts are buying assessment systems that promise to provide information to improve learning without fully examining the validity of these claims.

The focus of this article is two-fold. First, we define interim assessments, distinguish them from formative assessments, and focus on their uses. Second, we provide a framework for evaluating these interim assessments to help state and district leaders thoughtfully examine the commercially available products, develop strong specifications for a customized system, or develop their own interim assessments. A final purpose of this article is to promote interest in further research in this area, and to that end, we conclude with a section describing our vision for this research.

Throughout this article, our discussion will focus on how interim assessments fit into the comprehensive system and what unique value, if any, interim assessments serve. We attempt to describe the characteristics of effective interim assessments, discuss the different purposes these assessments may serve, provide information on how to choose the best type of assessment for a given situation, and then offer guidance on evaluating the products that already exist in the marketplace. Although we believe that there are some organizations trying to sell item banks and reporting systems as interim assessments without thoughtfully integrating them into a state assessment system, our goal is not to condemn all currently available products, but rather to provide a framework for the consumer to use, in evaluating them.

Distinguishing Among Assessment Types

Before we can begin a thoughtful discussion on interim assessments, we need to agree on definitions. We prefer the schema that places assessments into three categories—summative, interim, and formative—and distinguishes among the three types based on the intended purposes, audience, and use of the information. Summative assessments are given one time at the end of the semester or school year to evaluate students' performance against a defined set of content standards. These assessments are typically given statewide (but can be national or district) and are usually used as part of an accountability program or to otherwise inform policy. They could also be teacher-administered end-of-unit or end-of-semester tests that are used solely for grading purposes. They are the least flexible of the assessments.

Skipping to the narrowest type, formative assessment is used by classroom teachers to diagnose where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve student learning. The assessment is embedded within the learning activity and linked directly to the current unit of instruction. It can be a five-second assessment and is often called "minute-by-minute" assessment or formative instruction. Furthermore, the tasks presented may vary from one student to another depending on the teacher's judgment about the need for specific information about a student at a given point in time. Black and Wiliam (1998) defined formative assessment as just one part of formative instruction. In their seminal piece, *Inside the Black Box,* they argue that formative assessment cannot stand alone but must be a part of a whole system that uses the information from the assessment to adapt teaching to meet the learner's needs. Providing corrective feedback, modifying instruction to improve the student's understanding, or indicating areas of further instruction are essential aspects of a classroom formative assessment. There is little interest or sense in trying to aggregate formative assessment information beyond the specific classroom.

Finally, interim assessments are considered medium-scale, medium-cycle assessments, falling between summative and formative assessments and usually administered at the school or district level. Typically, interim assessments are given several times a year, although a test that was administered once at some midpoint during the year could also be considered interim. While the results may be used at the teacher or student level, the information is designed to be aggregated beyond the classroom level, such as the school or district level. That is, they may be given at the classroom level to provide information for the teacher, but a crucial distinction is that these results can be meaningfully aggregated and reported at a broader level. As such, the timing of the administration is likely to be controlled by the school or district rather than by the teacher, another critical feature separating these tests from formative assessments.

Although many others have used the term "interim assessment" to describe benchmark, diagnostic, predictive, and even some formative assessments, we offer the following definition:

> Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts.

By this definition, end-of-chapter tests available in most textbooks could be considered interim, if they are designed to be used to inform instructional decisions and reported in the aggregate. Teacher-created tests given at the end of a unit could be interim, formative, or summative, again depending on their purpose and design. The key components of the definition are that interim assessments (1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed to inform decisions

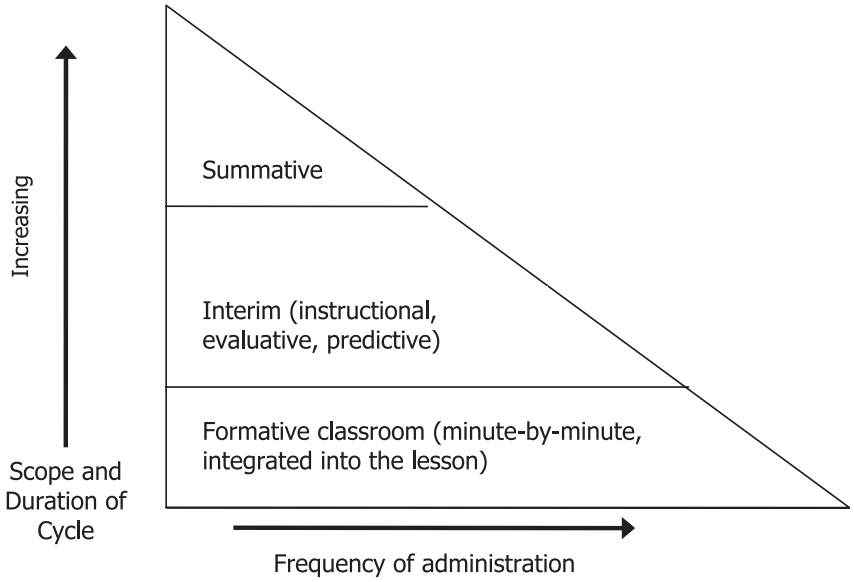


FIGURE 1. Tiers of assessment.

both at the classroom and beyond the classroom level, such as the school or district level. If the test is used simply for grading purposes, it is summative, while if it is used solely for the purpose of informing the teacher of a student's progress, it is most likely formative. These assessments may serve a variety of purposes, including predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program or pedagogy, or diagnosing gaps in a student's learning. It is these purposes that determine the necessary features of the assessments.

These three tiers of assessment—summative, interim, and formative—are shown in Figure 1. The triangle illustrates that formative assessments are used most frequently and have the smallest scope (i.e., the narrowest curricular focus) and the shortest cycle (i.e., the shortest time frame, typically defined as 5 seconds to 1 day), while summative assessments are administered least frequently and have the largest scope and cycle. Interim assessments fall between these other two types on all dimensions.

Overview of Interim Assessments

We encourage the reader to think broadly about the possible forms of interim assessments, from commercially purchased, computer-based sets of multiple-choice items to more locally created sets of extended performance tasks administered commonly throughout a school, district, or state. We do not intend to tout one type of interim assessment as being the best—although we argue that some are clearly superior for improving learning than others—but to encourage users to be explicit about the desired purpose of the assessment and then find the assessment that best fits that purpose. For example, an interim assessment may be given in order to

(1) Evaluate how well the student has learned the material taught to date.
(2) Predict students' performance on a summative assessment.
(3) Determine whether one pedagogical approach is more effective in teaching the material than another.
(4) Provide aggregate information on student achievement at a district level.
(5) Provide specific feedback on where the gaps in a particular student's knowledge are at the classroom level.
(6) Determine whether students are on track to succeed on the summative assessment.
(7) Diagnose and provide corrective feedback to help a group of students get on track to succeed on the summative assessment.
(8) Motivate and provide feedback to students about their learning.
(9) Provide information to help the instructor better teach the next group of students by evaluating the instruction, curriculum, and pedagogy.
(10) Ensure that teachers are staying on track in terms of teaching the curriculum in a timely manner (i.e., pacing).
(11) Provide a more thorough analysis of the depth of students' understanding.
(12) Determine whether students are prepared to move on to the next instructional unit.

Summarizing this large list brought us to three general classes of purposes for interim assessments: instructional, evaluative, and predictive. Although this categorization is not perfect, it seems to capture the essence of most of the goals of using an interim assessment system. We recognize that many assessments are not designed to serve only a single purpose, but we argue that few assessments or assessment systems can serve more than two or three purposes well and they tend to work best when the various purposes have been prioritized explicitly. Thus, an important additional step is to check not only whether the assessment is being used for its intended purposes, but to check the quality with which it meets those purposes.

Instructional Purposes

The primary goal of an interim assessment designed to serve instructional purposes is to adapt instruction and curriculum to better meet student needs. Of the three purposes, this one aligns most closely with the previous definitions of formative assessment. That is, the results of these assessments are used to adjust instruction with the intent of helping the students assessed meet the learning goals. However, the testing and reporting time frame of these interim assessments is typically medium cycle, whereas classroom formative assessments tend to operate on shorter cycles.

Subsumed under this purpose are other types of assessment that certainly would not meet the definition of *formative* presented earlier, but are instructional nonetheless. Consider, for example, features included in many commercially available systems. A typical system contains a bank of items nominally aligned with the state curriculum that teachers can use to create a test to evaluate student learning on the concepts taught to date. Results

may be reported immediately, and data are disaggregated by content standard allowing teachers to identify strengths and weaknesses in the students' learning. This type of interim assessment might be labeled formative, but we would argue that to be truly formative it must be timed appropriately for adjustments to instruction to occur, be aligned with specific local curriculum, provide more in-depth analyses of student misconceptions or lack of understanding, lead to strategies for improving instruction, and lead the teacher to modify instruction. Nevertheless, this type of assessment falls under the instructional category.

To serve instructional purposes, an assessment system must go beyond simply providing data. Educators must have strategies for interpreting and using the data to effectively modify classroom instruction. It is worth noting a tension between the need for professional development to accompany these assessment systems and the ownership of that responsibility. It is the contention of many assessment developers that tools and strategies for improving instruction are the teacher's responsibility, not the instrument provider's. Whether that professional development support is or should be included in the instructional interim assessment package will be debated among policy makers, developers, and educators. We feel strongly that no matter what the source of professional development, an assessment system purchased for instructional purposes will be effective only when used by educators who have the knowledge and tools to use the assessments and the results appropriately. Ideally, we believe that promoting informed use would be supported by development and training by both the developer and the user.

Evaluative Purposes

Another type of purpose an interim assessment might serve is to provide evaluative information about the curriculum or instruction. Think of this as a programmatic assessment designed to change instruction not necessarily in mid term but over the years. The students benefiting from the information gleaned from these assessments would not necessarily be the students assessed, but the students receiving the instruction in the future. Many had hoped that summative end-of-year assessments would fulfill this purpose, and in many cases these end-of-year tests have provided useful evaluative data, but most are too short and designed to cover too much content to provide the depth of information required for most evaluative purposes.

District-level policymakers are often interested in interim assessment systems for reasons other than to inform modifications to instruction. For instance, their goals may be to enforce some minimal quality through standardization of curriculum and pacing guides, to centralize coordination for highly mobile urban student populations and high teacher turnover, or as a lever to overcome differences in learning expectations and grading standards. These types of purposes are evaluative in nature.

Assessments used for evaluative purposes could be given district wide to compare the effectiveness of various instructional programs for improving student learning. Consider, for example, a district that is experimenting with more than one reform program or pedagogical strategy across different schools. The use of interim assessments in this context could be an effective way of monitoring the relative efficacy of each program. Similarly, assessments could be given at various points throughout the year to measure growth—not with the intention of intervening but for evaluating the effectiveness of a program, strategy, or teacher.

The assessments could also be used on a smaller scale, providing information on which concepts the students understood well and which were less clear. Teachers within one or more schools could use this information with the goal of helping them modify the curriculum and instructional strategies for future years. Other purposes could be to provide a more in-depth understanding at the school level on how the test items link to the content standards and how instruction can be better aligned with improved performance on the test. Of course, teachers can and should always learn from their experience. Any instructional interventions that could improve instruction in a current year should be implemented.

In our definition, an *evaluative* assessment would be designed explicitly to provide information to help the teacher, school administrator, curriculum supervisor, or district policymaker learn about curricular or instructional choices and take specific action to improve the program, affecting subsequent teaching and thereby, presumably, improving the learning. Assessment systems designed to serve evaluative purposes must provide detailed information about relatively fine-grained curricular units. However, not every student needs to be assessed in order for the teacher or administrator to receive high-quality information from the assessment. A matrix sample could be used to maximize the information while minimizing the time spent on assessments in the classroom.

Predictive Purposes

Predictive assessments are designed to determine each student's likelihood of meeting some criterion score on the end-of-year tests. Predictive purposes of interim assessments are important to many users and this interest could increase as the annual NCLB targets continue to rise. In addition, assessments in this category could be used to predict performance on a high school exit exam or success with postsecondary curriculum. Although predictive purposes are important in high-stakes testing situations, we suspect that there are few assessment systems where the sole purpose for the system is prediction. Rather, most users want additional information to help them improve the performance of students for whom failure is predicted. This additional information might come from the assessment itself or from further probes to determine areas of weakness in those not on track to succeed. This scenario could be an example of how interim and formative assessments work together to help improve student performance on a summative assessment. It also highlights the importance of aligning all components of a comprehensive assessment system.

A confounding variable on any predictive test is that if it provides good feedback on how to improve a student's learning, then its predictive ability is likely to decrease. That is, if the test predicts that a student is on track to perform at the basic level, and then appropriate interventions are used to bring the student to proficient, the statistical analysis of the test's predictive validity should underpredict student performance over time. However, it is important to track the performance of students predicted to succeed on the summative test, and questions should be raised if too many students predicted to pass the summative test actually fail it.

Identifying the Goal

As policymakers decide to bring an interim assessment system to their state/district/school we encourage them to have a theory of action for how the particular assessment system will work in the teaching-learning cycle. Policymakers and educators using assessments need to understand the limitations of any assessment for fulfilling particular purposes. As a start, we think it will be helpful for educational leaders to address the following questions:

(1) What do I want to learn from this assessment?
(2) Who will use the information gathered from this assessment?
(3) What action steps will be taken as a result of this assessment?
(4) What professional development or support structures should be in place to ensure the action steps are taken appropriately?
(5) How will student learning improve as a result of using this interim assessment system and will it improve more than if the assessment system were not used?

The answers to these questions will dictate the type of assessment needed and will drive many of the design decisions including the types of items used, the mechanism for implementing it, the frequency with which it should be administered, and the types of reports that will need to be developed from the data.[1] Importantly, these questions and the associated answers serve as the beginning of a validity argument to support (or refute) the particular assessment system.

Answering these questions also may suggest that it might be appropriate to consider primary and secondary purposes in designing or choosing an interim assessment system. For instance, while the primary purpose of giving an interim assessment may be evaluative, we would hope that given the results for a specific set of current students, teachers and school leaders would attempt to provide remediation programs for those students not understanding key concepts. Similarly, even when the primary purpose of an interim assessment is to predict success on the end-of-year assessment, a policymaker may also want the predictive assessment to provide some diagnostic information so that educators can intervene with students predicted to score below a critical level. Of course, the assessment may only fulfill secondary purposes if certain factors associated with a primary purpose—such as having a very short test—do not overly constrain other uses.

Finally, the answers to the above questions should help policymakers to determine whether the best approach is to adopt a currently existing system or to build their own. There are many vendors currently selling interim assessments under various labels. These assessments are marketed to serve a plethora of purposes, including serving as a diagnostic tool, providing information that can be used to guide instruction, determining student placement, measuring growth or progress over time, and predicting success on a future assessment. Typically these systems consist of item banks, test assembly supports, administration tools, and customized reports. These systems often are computer- and even web-based, allowing students to take the test whenever they wish (or their teacher wishes) and wherever a computer with an Internet connection is available. Others also have the option of creating pencil-and-paper tests. Teachers can construct the tests, the tests can be fixed by an administrator, or the tests can be adaptive. The items are "linked" to content standards,[2] and results typically are reported in terms of number correct or as scale score developed by the publisher. The "diagnostic" portion tends to be a summary of results by content standard. Often, these systems provide a variety of options for reports, with different levels of aggregation.

Other states and districts have experimented with developing in-house local assessments. These tend to be computer-based systems that include teacher-developed items linked directly to instructional units. They give quick feedback to teachers and produce in-depth reports at the student and classroom levels. It seems that most of these systems have been developed for instructional purposes rather than as predictive or evaluative.

There is no one-size-fits-all assessment, only a best design for a desired use and existing constraints and resources. We believe that many educational leaders consider a cost-benefit relationship before investing in such a system, but we fear that the equation often tips in favor of low costs and short testing time. For instance, it is cheaper to score multiple-choice items than constructed-response items or performance tasks, and it often costs less to buy a computer-based testing system than to invest in professional development for all teachers. We recognize the reality of constrained budgets, but argue that saving a few dollars on an assessment system might actually "cost" more in terms of opportunities for learning that may be lost as a result of cutting up-front purchase costs.

Characteristics of an Effective Interim Assessment System

This section of the article is intended to help educational leaders either choose or develop a strong interim assessment system for their schools. We provide evaluative criteria to help policymakers critically appraise their local assessments and also provide suggestions for the type of validity evidence to collect over time. We recognize that some districts or states will be looking to purchase an already available assessment system, while others will be looking to create a system customized to their needs. The considerations described below are appropriate for both needs.

Again, we emphasize that the purpose must be clearly stated before one can truly determine or evaluate the necessary characteristics of the assessment. Consideration should be given to all parts of the interim assessment, including item quality, administration requirements, and reporting elements. This last piece is important because the report is the mechanism for translating the assessment data into decisions, which then translate into action and should be one of the first considerations in designing a new assessment. It serves to transform raw data into results that can be interpreted meaningfully and acted upon appropriately. Time should be spent discussing the question: what do we want the tests to tell us? Assessments serving an instructional purpose will have different features in their reports than those serving predictive or evaluative purposes.

Evaluative Criteria

To help guide the evaluation of commercially available interim tests and the development of custom interim assessment systems, we have provided the following criteria for states and/or districts to consider prior to purchasing or developing an interim assessment system. We find that most, if not all, of these criteria fit under Standard 15.8 of the *Standards for Educational and Psychological Testing*

(American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999):

> When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Following our argument that the interim assessment design must be linked to the purposes and intended uses, we present evaluation criteria for the three major purposes articulated earlier: instructional, evaluative, and predictive. To avoid redundancy, we first present several general criteria that cut across all three purposes.

General

(1) A test can be no better than the quality of the items it contains. Therefore, the quality of the items needs to be evaluated against professional standards and expert opinion. The types of items/tasks may vary depending on the specific purposes and intended uses, but all should be of high quality as documented through traditional reviews for content and bias and sensitivity as well as pilot testing and data reviews.
(2) Alignment evidence should be provided to document the relationship of the items and sets of items in a test "form" to the knowledge and skills (including depth of knowledge) called for in the target content standards.
(3) The inferences resulting from the test scores should be validated for the intended uses and purposes.
(4) The test publisher must include clear guidelines regarding the appropriate uses of the assessment results, as well as indicating either potentially inappropriate uses of the results or limitations of the validity evidence.
(5) Tasks should be applicable to the target student populations; in most schools and districts these may include English language learners and students with disabilities.
(6) There should be evidence that the professional development associated with the assessment system facilitates educators' appropriate interpretation and use of the assessment results for the specified purposes. Clearly, assessments serving instructional purposes will require different professional development than is required for evaluative and predictive purposes, and the audiences (teachers, building administrators, district leaders) may differ for each.
(7) For interim assessment systems that require a "break" from instruction in order to test, educational leaders should consider the time required for assessment, which should be as short as possible to provide the desired information. For performance tasks embedded in instruction, the issue of "testing time" is less critical.

Instructional

(1) To the extent possible, interim assessments for instructional purposes should fit as seamlessly with instruction as possible and represent an opportunity for student learning during the assessment experience.
(2) Ideally, the system should provide evidence, based on scientifically rigorous studies, demonstrating that the assessment system has contributed to improved student learning in settings similar to those in which it will be used.
(3) There should be evidence that the results of the assessment and the associated score reports have been designed to facilitate meaningful and useful instructional interpretations.
(4) Clear guidelines should be provided explaining how the results of the assessment, including the results of particular tasks/items or sets of items, should be used to help inform instructional decisions.
(5) Each particular assessment in the system must link closely to the curricular goals taught prior to the assessment administration, preferably quite proximal to the assessment event. The assessment should include only content and skills for which the students have had a legitimate opportunity to learn, unless the purpose of the assessment is as a pretest to determine readiness for some learning in the near future or as a placement test.
(6) To best serve instructional purposes, each interim assessment should assess only a limited number of important curricular goals to make it more likely that instructional adjustments can be timely and targeted appropriately.
(7) In general, to serve instructional purposes interim assessments intended to support diagnosis of students' understanding and misconceptions should include high-quality open-ended tasks. All items, whether open ended or multiple choice, should be developed so that useful information about students' understanding and cognition can be gleaned from specific incorrect answers.
(8) Instructional interim assessments should measure instructional and curricular goals, provide information not easily gleaned from the state's large scale assessment such as more in-depth understanding demonstrated through extended tasks or synthesis works.

Evaluative

(1) The collection of tasks administered through the year should represent a technically sound range of difficulty and appropriate breadth, dependent on the focus of the evaluation.
(2) The assessments should comprise items and tasks with a mix of formats to provide users a deep understanding of the relative effectiveness of educational programs.
(3) The assessment must be targeted to the content standards that are the focus of the educational program(s) being evaluated or studied and/or to the expected domain of transfer.
(4) The reports must be designed to facilitate the intended evaluation and accurately portray the error associated with the scores and subscores.

Predictive

(1) The assessment should be highly correlated with the criterion measure (e.g., the end-of-year state assessment). The technical documentation should include evidence of the predictive link between the interim assessment and the criterion measure. However, in order to justify the additional testing and cost, the predictive assessment should be significantly more related to the criterion measure than other measures (e.g., teachers' grades) that could be used.
(2) The predictive assessment should comprise items with a similar mix of item types as the criterion measure.
(3) The predictive assessment should be designed from the same or similar blueprint as the criterion measure.

(4) The reports should be designed to facilitate the intended predictions, including an honest and accurate characterization of the error associated with the prediction, both at the total score and subscore levels.
(5) If the purpose of the assessment goes beyond solely predicting performance to identifying areas of weakness, the assessment should contain enough diagnostic information so that remediation can be targeted for students predicted to score below the cut on the criterion measure.

We are not suggesting that interim assessment systems must meet all the criteria listed above before being purchased for a district or state, but we recommend that educational leaders consider the criteria when evaluating which, if any, system to purchase or when evaluating a proposal to create a customized system.

Validity Evidence

Approaching this from a validity perspective, we argue that the interim assessment system should be validated for the specific purposes and uses. Validity evidence would include:

(1) A clearly articulated goal or target. An interim assessment serving an instructional purpose, for example, must include a rich representation of the content standards students are expected to master.
(2) High-quality items that elicit and assess what is intended. Items should be directly linked to the content standards and specific teaching units.
(3) Useful and clear interpretations to support the intended uses.
(4) Operational feasibility and low negative unintended consequences. A predictive interim assessment should minimize the loss of instructional time.

Additionally, any provider should be required to provide evidence of the validity of the system for the intended purposes. Once the system has been implemented, the sponsor—whether districts and/or states—should periodically evaluate the system to ensure that it is meeting intended purposes and uses. While any evaluation will have to be tailored to the specific purposes and uses, we offer the following general suggestions for exploring the validity of an interim assessment system:

(1) If the test is used for instructional purposes, follow up with teachers to determine how the data were used, if they provided useful information, and whether there was evidence of improved student learning, including evidence of generalizability and transfer, for current students.
(2) If the test is used for evaluative purposes, gather data from other sources to triangulate results of interim assessment and follow up to monitor if evaluation decisions, such as changes to curriculum and/or instruction, are supported.
(3) If the assessments are used for either instructional or evaluative purposes, look for evidence of increases in teacher knowledge of content, pedagogy, and student learning.
(4) If the test is used for predictive purposes, do a follow-up study to determine that the predictive link is reasonably accurate, provides more predictive power than information such as grades and teacher judgments, and that the use of the test contributes to improving criterion (e.g., end-of-year scores).
(5) Regardless of the purpose of the assessments, the manageability, including the quality of implementation, should be monitored.
(6) Finally, any unintended negative consequences should be monitored for all interim assessments including any adverse effects on student motivation as a result of engaging with the tasks, a narrowing of the curriculum, or a decreased focus on formative assessment.

Matching the Purpose with the Assessment

The main impetus for this article was to provide advice on how to evaluate the suitability of commercially available or locally created products for states and districts considering implementing some sort of interim assessment system. We have continued to emphasize the need to articulate the purpose(s) of such a system.

We recognize that in most instantiations of interim assessment educational leaders are trying to squeeze as many purposes as possible out of a single system. Unfortunately, one of the truisms in educational measurement is that when an assessment system purports to fulfill too many purposes—especially disparate purposes—it rarely fulfills any purpose well.[3] This does not mean that certain interim assessment systems cannot fulfill more than one purpose, depending on the level addressed by the primary purpose. If the system is intended to provide rich information about individual students' strengths and weaknesses tied to a particular set of curricular goals, then these results likely cannot be aggregated to the subgroup, school, and/or district level to provide evaluative information. On the other hand, if the primary goal is to gather predictive or early warning information, it is unlikely that the assessment will contain rich enough information to serve instructional or even evaluative purposes. Therefore, users should design a system that will adequately fulfill the more important and finest grain purpose first and then consider whether additional purposes can be fulfilled well within the same assessment, or whether it would be more appropriate to use multiple assessments—including formative assessment—within a comprehensive system.

We recommend that educational leaders considering purchasing a commercially available system follow the advice offered in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), specifically in Standard 11.1:

> Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are those that summarize the test's purposes, specify the procedures for test administration, define the intended populations of test takers, and discuss the score interpretations for which validity or reliability data are available.

Future Areas of Research Needed

Clearly, this field is rich for further research. New studies funded by the U.S. Department of Education's Institute of Education Sciences are exploring areas that may inform the field of formative uses of assessment. Many of these studies focus on interim assessments, sometimes as part of a tutoring session or computer-based learning. In general, they examine how testing a particular unit of instruction relates to retention of information after an extended period of time. One common finding across studies was that student performance

on the "repeated testing" was not nearly as important as the corrective feedback they received as a result. That is, a student who guessed incorrectly on an item on a unit test, but who received good corrective feedback, was just as likely to answer a similar item correctly on a future test as a student who had answered it correctly the first time. Another common finding we found interesting was that the repeated testing, in and of itself, contributed to retention. And this was particularly true when the short tests required students to generate their own responses on short-answer items (Viadero, 2006). We look forward to seeing the results of these studies when published.

We feel it is important to continue to examine how the use of interim assessments can help further student learning. Education leaders can find themselves in a difficult position if they do not want to adopt a test without validity evidence, while there is little validity evidence available. So, the first area we see the need for strong research efforts is in validating the use of these types of assessment. In general, we see the need for research in the following areas:

(1) Score-based inferences from interim assessments need to be validated for the use of improving performance on summative assessment and gather evidence to evaluate this argument. Choose several types of interim assessments and validate their uses.
(a) Are predictive assessments truly predicting student performance on end-of-year assessments more so than other readily available data? Of course, the results of this question could be confounded by the use of appropriate interventions, but those interventions may provide evidence of the validity of the consequences.
(b) Are instructional assessments actually improving instruction? Are there any unintended consequences?
(c) Are evaluative interim assessments effectively identifying differences in various pedagogies or instructional approaches? What characteristics make them more useful?
(2) Studies are needed to examine differential effects of interim assessments on students' intrinsic motivation to learn. Consider the concern that frequent assessments may diminish intrinsic motivation by shifting the effort and purpose from learning "to know" to learning "to display one's knowledge" (Lave & Wenger, 1991). How can we use the interim assessments constructively to further students' desire to learn rather than to further their desire for a high score?
(3) Kluger and DeNisi (1996) and others found that normative types of feedback or feedback that focuses on the person rather than on the task can actually have a negative effect on student performance. Their research showed that the most effective types of feedback were ones in which students were told not only what they needed to learn but how to get there. How does this research apply to the interpretation of results from interim assessment?
(4) It has been argued that evidence collected for summative purposes can rarely be disaggregated to support learning, but evidence collected for formative purposes can be aggregated to support summative inferences (Wiliam, 2006). However, we need to learn more about how to aggregate results of formative assessments before pursuing this path. What are the requirements for building a system that provides teachers the information they need but can still be scaled to compare results across students, teachers, and/or schools?
(5) What are the effective strategies for implementing interim assessments and presenting results so that teachers use the data appropriately for making effective educational decisions?
(6) What types of professional development are necessary to influence effective use of interim assessments and what factors (e.g., teacher qualifications) interact with various professional development models? What approaches are most effective for providing this type of professional development on a large scale?

There are a host of other lines of inquiry areas that one might pursue to build the research base on interim assessment, but we think that the ones listed above are an important starting point.

Discussion

We first approached this article from the perspective of investigating "formative" and "benchmark" assessments being used at the district and state levels. Because many assessments now in the field are marketed under the appropriated term "formative assessment," we realized that there needed to be a discussion regarding the current types of assessments being sold for formative purposes. Then, we turned to developing a framework to better understand interim assessments: how they are used and why they are proliferating at such a rapid rate. Furthermore, we were interested in the role that state and district leaders play in selecting/developing these assessments and how we might be able to help these leaders with this task. When asked why we chose to focus on interim assessments rather than the purer and research-based formative assessments, our answer was simple: states and districts are spending considerable resources to implement such systems.

We recognize the difficulty of developing, at a state level, strong formative assessment strategies as advocated by Black, Wiliam, Shepard, and others. Components such as weaving the assessment seamlessly into the curriculum and providing useful feedback that leads to appropriate modifications in instruction is difficult when the agent (state department of education personnel) is several steps removed from the classroom. While states can support professional development programs that help educators develop and use such tools, they could also help by purchasing a preexisting system, if such a system supports formative and professional learning needs. In addition, states may have other requirements for an assessment program, such as developing an early warning system to identify students who are not on track to succeed in order to help with additional supports. Or, the states may wish to use these interim assessments as evaluation tools for different schools, instructional programs, or pedagogies. That is why we chose to define interim assessments, focused on specific purposes and uses, as tools to evaluate students' knowledge and skills that are designed to inform decisions at the classroom level and above.

That said we are concerned that many of the commercially available systems are quite different from what the research currently supports, and those selling such systems promise far more than they can deliver. For example, these systems often lay claims

to the research documenting the powerful effects of formative assessment on student learning when it is clear that Black and Wiliam's (1998) meta-analysis evaluated studies with formative assessments of very different character than essentially all currently available commercial interim assessment programs.

We believe it is not worth spending scarce resources on interim assessments that simply administer a series of minisummative assessments. The assessments should be linked to specific instructional units to provide teachers with useful information. While pre- and posttest designs may be useful for some purposes, testing students on material they have not yet learned rarely provides teachers with helpful information. We have seen systems where shorter versions of the end-of-year assessment are given periodically throughout the school year. The items on these assessments are placed on the same scale as the items on the end-of-year assessment, so the results can be used to show progress toward the goal. A cursory examination of several of these systems revealed that they do not meet the criteria discussed in this article and suffer from such technical and content shortcomings that we believe they are a poor use of money and instructional time.

A good interim assessment can be an integral part of a state's or district's comprehensive assessment system, used in conjunction with classroom formative assessments and summative end-of-year assessments. As such, we believe that there are valid purposes for giving interim assessments beyond informing instruction at that point in time. However, the policymakers and educators using the assessment need to understand the purpose of the assessment and what it can and cannot do. If policymakers want an assessment to help educators improve instruction, they should look for one that ties directly to the classroom instruction and provides in-depth examination of not just which items students miss but why they miss them. Actually, if this is the sole goal of the assessment, we argue that resources would be better spent helping teachers learn formative assessment techniques, including using the information to intervene with students who do not yet understand key concepts.[4] If policymakers want an assessment to tell them how students are likely to perform on an end-of-year assessment, they need to examine the reliability of the predictions and the information describing what to do next.

At a minimum, we argue that any expenditure of resources (teacher time, money, etc.) for an interim assessment system must provide experiences and information that are not available on the state large-scale assessment or in the classroom through daily instructional activities, including formative assessment. Finally, any of these assessment types need to provide evidence of their validity. Are they demonstrating their intended positive consequences and are there any unintended negative consequences of their use? For instance, do additional assessments solidify a student's understanding of a concept or inure him to tests in general? Such validity evidence should be examined prior to adoption of the assessment program and should also be generated for the specific populations and context of the state's or district's program. These interim assessments can be an integral part of any comprehensive assessment system and should be considered as a piece of a whole and evaluated as such.

Notes

[1]For more information about these types of design decisions within the context of interim assessments, please see Perie, Marion, and Gong, 2007.

[2]Unfortunately, the strength of the alignment between such commercial tests and the state content standards is rarely evaluated by independent analysts, so the "link" between the two is often based on the publishers' claims.

[3]This should also be a red flag to any educational leader considering purchasing a system that promises to fulfill many purposes or to solve all educational problems.

[4]However, this assumes that the curriculum is sound; our experience has been that often considerable attention needs to be paid to the curriculum before fine tuning any instruction through formative assessment.

Acknowledgment

The authors wish to acknowledge the work of the Council of Chief State School Officers' formative assessment working group (FAST SCASS) that helped us to clarify the distinctions between formative and interim assessment. In particular, we received influential feedback from Jim Popham, Margaret Heritage, and Fritz Mosher. We also would like to thank others that reviewed earlier drafts including Bob Linn, Sue Brookhart, and three anonymous reviewers. Finally, we wish to acknowledge our colleagues at the Center who focused our thinking on the definition and use of interim assessments, specifically Rich Hill, Charlie DePascale, Karin Hess, and Jennifer Dunn.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice*, *5*(1), 7–74. Also summarized in an article entitled, Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

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

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. National Center for the Improvement of Educational Assessment. Dover, NH: NCIEA. Available at www.nciea.org.

Viadero, Debra. (2006). Cognition studies offer insights on academic tactics: U.S.-funded projects eye ways of helping students remember more material. *Education Week, 26*(1), 12–13.

Wiliam, D. (2006). Assessment for learning: Why, what and how. *Orbit: OISE/UT's Magazine for Schools*, *36*(2), 2–6.

The Assessment-Literate School Administrator

Stuart Kahl, Ph.D.
Founding Principal

We hear a lot about the need for higher levels of assessment literacy. I believe that the need is urgent, and it can only be satisfied by moving beyond simple fact sheets and glossaries of measurement terminology. Just as our students need to be able to apply their knowledge in situations requiring higher-order thought, school administrators should apply their knowledge of assessment to inform many of the decisions they're called upon to make. Here are some characteristics of assessment-literate administrators.

- The assessment-literate school administrator understands that there are different purposes for interim assessments and knows why some types are better suited to a particular purpose than others.
- The assessment-literate school administrator is equipped to evaluate testing program options armed with the knowledge that alignment of a test or set of test items is more than the assignment of items to broad categories in a state's content standards.
- The assessment-literate school administrator knows that the use of general achievement measures several times during the school year better demonstrates measurement error than student growth.
- The assessment-literate school administrator knows that subtest scores from general achievement measures are unreliable because subtests are essentially very short tests; therefore, they should be used to identify patterns of deficiencies over multiple administrations or to raise questions that could lead to further investigation.
- The assessment-literate school administrator recognizes the types of assessment the formative assessment research was all about. To bolster formative assessment practices, he/she invests in professional development for teachers instead of many of the "formative" tools available from external sources.

Each year, districts and states invest significantly in testing. It's critical that such spending decisions be informed by a deep level of assessment literacy. But the onus is not solely on administrators. Assessment literacy is important for policy makers, teachers, students, and parents, as well. Assessment plays too many important roles in the education of young people for us to allow any of these stakeholders to be less than proficient when it comes to assessment literacy.

What do you think?
Let us know at twocents@measuredprogress.org

You Can't Squeeze Blood out of a Turnip: What Diagnostic Testing Is —and Isn't

Stuart Kahl, Ph.D.
Founding Principal

With local educators feeling greater pressure than ever to raise student achievement levels, there is a parallel increase in demand for diagnostic information from testing. Unfortunately, many are looking for such information from tests that were not designed to be diagnostic. "Diagnosis" is defined as the identification of the nature and cause of something. A physician diagnoses illness, while an information technologist diagnoses problems with software code or some other component of a computer system.

In education, clinical testing can diagnose cognitive or learning disabilities for individual students. However, non-clinical, diagnostic testing sheds light on specific concepts or skills that a student is having difficulty learning and determines "why"—possibly pinpointing prerequisite concepts and skills that may need to be re-taught. Two issues of concern with this latter type have to do with the specificity of the "diagnostic" information and the validity of inferences that can be drawn from it.

Growing emphasis on statewide summative tests (as well as interim tests that look, smell, and feel like the state tests), coupled with concerns about over-testing and the loss of instructional time, have led teachers to look to such tests for diagnostic information. While the tests have several important uses, diagnosis as described above is generally not one of them. Most of the tests were designed to produce total test scores, along with a few student subtest scores.

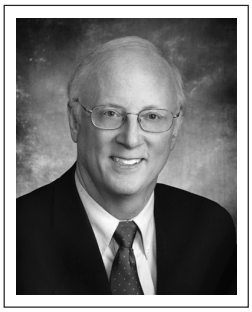
A subtest might represent only a quarter or fifth of the larger test, with items representing only a sparse sampling of a very broad subdomain (e.g., geometry/measurement in mathematics). Thus, subtest scores are not particularly reliable and no more diagnostic at the student level than a thermometer is for a medical diagnosis—it can indicate something's wrong, but it doesn't tell you what.

Summative tests should be used to raise programmatic questions that require further investigation: Why is one subgroup of students performing lower than another? Why are students performing more poorly than expected in a particular subdomain? Further investigation could involve discussions with teachers and students, review of curricula, and more testing. These tests can raise similar questions about individual students, but because inferences regarding individual students are even less reliable, deeper probing is absolutely necessary in order to acquire useful diagnostic information.

Educators may try to squeeze diagnostic information out of general achievement measures by over-interpreting single item results and speculating, based on probabilities, on hypothetical performance by students on test items unrelated to ones they actually took. While these approaches can be useful in a discussion of the general strengths and weaknesses of a large group of students, they would be no more accurate than a coin toss for individual student diagnoses.

In the next few years, I expect we're going to see more and more testing and greater and greater emphasis on test results. It is critical that educators and non-educators, parents and policy makers alike, become more assessment literate and therefore more likely to use the right assessments for the right reasons, avoid misdiagnosis, and assure that testing will be as effective as it can be in support of teaching and learning.

What do you think?
Let us know at twocents@measuredprogress.org



Helping Teachers Make the Connection between Assessment and Instruction

Stuart Kahl, Ph.D.
Founding Principal
Measured Progress

Most discussions about interim and formative assessments end in agreement that many teachers would benefit from professional development on a variety of assessment topics. The trick is to identify the right topics.

First, though, let's rid ourselves of the notion that teachers don't know anything about testing or that their testing practices should be presumed to be bad. Remember, high school grade-point average is a better predictor of college success than are college entrance examination scores. At all grade levels, teachers usually have a pretty good idea of what their students' relative performance will be on standardized tests. On the rare occasions when their predictions are wrong, the smart money is on the teachers—not the results of that single test.

The reason for this high level of awareness is simple: the grades teachers give students are based on multiple and varied measures administered during a marking period, semester, or year. Those cumulative measurements are generally more reliable than a single standardized test—even one of the highest quality.

It's also true, however, that a single standardized test may be more reliable than a single, teacher-made measure. Plus, with teacher-assigned grades, variation in standards across classes and schools (e.g., the quality of performance associated with specific letter grades) is an issue. (This is why there are standardized tests.) But this situation should be expected, just as it should not be surprising that standards for proficiency vary across states using different tests and establishing those standards independently.

So let's start with the premise that teachers already know a lot about gathering evidence of student learning. That being said, what are some areas of teacher training or experience that would provide fertile ground for continuing education? Here are some possibilities:

- Creating good tests—and test items
- Developing and using rubrics for scoring open-response items and performance-type assessments
- Evaluating student work—gathering information beyond a number
- Knowing what different types of assessments can and cannot do
- Using information from different types of assessments to modify instruction

In terms of assessment mastery, let's give teachers what they really need—professional development to ensure that assessments of all kinds lead to better teaching and learning.

ISSUE PAPER

TOPICS:

Standards-Based Tests

Norm-Referenced Tests

Vertical Scaling

Value-Added Model

Computer-Adaptive Tests

Diagnostic Tests

Customized vs. Off-the-Shelf Tests

Large-Scale Assessment: Choices and Challenges

Stuart R. Kahl, Ph.D.
Kevin P. Sweeney, Ph.D.

measured progress

Number 3

March 2004

States struggling to comply with the No Child Left Behind Act of 2001 (NCLB) are in a race against time to satisfy the assessment and accountability requirements of the law. At the district level, because of NCLB-required change data, which classifies large numbers of schools as failing to attain initial improvement targets, there is a scramble to find ways to raise student achievement levels.

A "trickle-down" effect has reached the testing industry, in the form of client demands for "faster, better, cheaper" tests. Achieving all three of these characteristics is often very difficult. Faster and cheaper tests are easy to produce. Adding "better" to the mix is relatively easy, too, if "better" refers only to the psychometric properties of tests. Testing companies, state departments of education, and local school systems can produce technically sound, reliable tests. However, to evaluate if a test is better, the question becomes, "better for what purpose?" Once the purpose is articulated, a particular test must fulfill the dual requirements of covering the "right stuff" and yielding appropriate information. If the test doesn't meet both of these criteria, then faster, cheaper, and even technically better testing may, in the long run, be of little use.

. . . to evaluate if a test is better, the question becomes, "better for what purpose?"

Unfortunately, the pressures of NCLB, coupled with an incomplete understanding of assessment issues, are pushing many policy makers and educators toward cheaper and faster tests that may not cover the "right stuff" or produce the best information—either for accountability purposes or school improvement. This issue paper describes the different testing options available to schools, districts, and states, and points out the pros and cons of each.

Standards-Based Tests

A standards-based test addresses standards in two ways. First, the test questions pertain to a particular set of content standards—statements of objectives defining the domain of knowledge and skills to be learned and assessed at a particular grade and in a particular subject. Second, results are reported in the context of performance standards—relative to various threshold scores, which create test-score ranges that correspond to different categories or levels of performance.

Content Standards and Alignment

The coverage of appropriate content is central to the development of a standards-based test. Off-the-shelf commercial products marketed to a wide audience address more generic, "common denominator" content. NCLB and the previous Title I reauthorization require a state's accountability tests to be aligned closely with the state's own content standards. Two-way alignment is required. Two-way alignment ensures that all test questions address the standards and that all measurable standards are addressed by questions in the test. In the current climate, alignment means much more than mere categorical alignment (whether a test question belongs to one of the content categories). The test questions must also reflect the depth of knowledge (cognitive level and complexity) and breadth of knowledge communicated by the content standards. Attention must also be paid to the balance of items across the standards. A true, standards-based test meets these alignment requirements.

Until a few years ago, many states' content standards were written for grade spans, leaving it to local school systems to determine at which grades within a span to teach particular expectations within standards. In the near future, NCLB requires that grades three through eight and one high school grade be tested in reading and mathematics. While states with grade-span standards will not have to redo their standards, they will have to develop grade-level expectations (grade-specific knowledge and skills to be taught and assessed). Thus, two states with similarities in their grade-span standards may assign the same specific content and skills to different grades.

Performance Standards and Achievement Levels

Student performance relative to established thresholds or "cut-scores" is the focus of the reporting of results in a standards-based program. The score range into which a student's numerical score falls is a critical piece of information. Thus, a student's performance level is reported along with the numerical test score. The critical statistic for schools, districts, and states is the percentage of students at or above the "proficient" level.

NCLB proficiency. NCLB requires that states use at least three performance categories or achievement levels, the critical one being the "proficient" level. By 2014, NCLB expects 100 percent of the students in every school and in each of several required subgroups within a school to achieve the "proficient" level. While the reasonableness of the interim targets and the ultimate expectation are the subject of much debate, these are not a topic of this paper. However, it should be noted that because each state independently determines the quality of work earning the "proficiency" designation, the percentages of students performing at "proficient" or above vary from state to state. Thus, the reasonableness of the proficiency expectation is even more questionable for schools in states with very high standards.

Consistency of performance standards across grades. Many states already had standards-based assessment programs in place before NCLB—some in just a few scattered grades, some in many grades. Those states that are now expanding their programs to additional grade levels to meet NCLB requirements face the added challenge of creating a coherent set of cut-scores across grades. In the past, a state that tested at a few grade levels, such as grades four, eight, and eleven, would set cut-scores independently for each grade. There would not be great concern if the percentages of students scoring at a particular level varied across grades. With the adjacent-grade testing mandated by NCLB, it would be inappropriate for a student passing through the grades to jump back and forth between "proficient" and "not proficient" just because of the way the cut-scores had been set. Thus, the task of standard setting must result in a coherent set of cut-scores across grades.

Decisions about individual students. Many states use the results of standards-based tests for decisions about graduation, advancement, or summer school. While legitimate, this use of testing results accentuates a sad truth about cut-scores and tests; a student scoring just above a cut-score and a student scoring just below the same cut-score are virtually indistinguishable from one another in terms of their proficiency based on the evidence from the test. In fact, there is close to a fifty-fifty chance that their true proficiency levels are reversed. This is the case, even for a test of the highest quality. This anomaly is the result of a psychometric concept called measurement error.

Measurement error does not refer to a mistake that has been made; it indicates that any test is subject to some degree of imprecision (e.g., while a student may attain a slightly different score when he or she retakes a test, that student's true, underlying ability has not changed). This is why accepted practice for using test scores to inform high-stakes decisions is to offer students multiple testing opportunities. A truly proficient student is not likely to fail due simply to measurement error when given several retest opportunities. The fact that different students can demonstrate their proficiency better by different methods is a reason it is also important that districts and states offer alternative measures and take into account additional, relevant information in making decisions about individual students. These matters are discussed in greater depth in "Measurement Error, Human Error, and Decisions Based on a Test," the second paper in this series.

A truly proficient student is not likely to fail due simply to measurement error when given several retest opportunities.

Norm-Referenced Tests

The assessments most familiar to several generations of Americans are off-the-shelf, norm-referenced tests (NRTs). "Norm-referenced" means that at some time the test was administered to national samples of students participating in norming studies. The data gathered from such studies allow the developers of these tests to produce national norms, which in turn, allow the reporting of students' scores in terms of national percentile ranks.

A student's national percentile rank on an NRT shows how he or she performed on the test compared to other students nationwide. (A percentile rank is the percentage of students a particular student's score equaled or exceeded). Unfortunately, the accuracy of norms is suspect, because it is becoming increasingly difficult to get truly representative samples of well-motivated students to participate in norming studies. Nonetheless, norm-referenced testing programs do a good

job of letting students and schools know where they stand in a general sense because all students' scores are placed on a common metric. Local school and district tests can differ significantly in difficulty and in the quality of work that is required for acceptable performance. District and state percentile ranks can be quite different from national percentile ranks depending on how the district or state performs relative to the nation. Thus, nationally normed NRTs provide valuable information on relative performance.

The purpose and content of NRTs. NRTs are designed to place students on a general achievement continuum—to rank order students and to produce reliable total test scores for students representing a great range of abilities. NRT items address "common denominator" concepts and skills; they are not designed to comprehensively cover a set of content standards that might be used by a state or local school district. In other words, NRTs do not measure how well schools are teaching or students are learning the material defined by the relevant content standards.

> . . . NRTs do not measure how well schools are teaching or students are learning the material defined by the relevant content standards.

Thus, for purposes other than ranking students on general achievement in a subject, traditional NRTs have significant alignment issues. And, with grade-level expectations and alignment studies focusing on far more than simple categorical alignment, those issues are of even greater concern. Students' range of ability in a particular grade spans many grades in terms of grade-level equivalents. Because NRTs must discriminate effectively across that full range, many test items in a test form for a particular grade are, in fact, more appropriate for students in grades above and below the target grade. These out-of-grade items are also useful in the vertical scaling (equating) of tests across grades.

NRTs and NCLB. Because NRT items do not address the breadth of many state content standards, and because of the grade inappropriateness of some items, the U.S. Department of Education only allows the use of off-the-shelf NRTs for NCLB assessment and accountability purposes if the tests are augmented by a substantial number of additional items, to ensure adequate coverage of state content standards.

Security is another potential problem related to the use of off-the-shelf NRTs for NCLB purposes. In the late 1980s, when new state laws called for accountability tests, the only tests readily available were the publishers' off-the-shelf products. Many states opted for these tests, although numerous school districts had been using the same instruments at the local level. This lack of security and the limited numbers of alternative forms available from the publishers may have contributed significantly to score inflation and, ultimately, a scandal that occurred in the testing industry. Because the stakes are much higher today, the potential increases for a repeat of the problem if state testing programs use the same off-the-shelf forms in multiple years.

National norms for customized tests. Some states are asking contractors to create national norms for the customized tests they are using for NCLB accountability. Several different possible approaches can be taken to meet this demand. One approach would be to administer the state's test to a national norming sample—a very costly and difficult endeavor. Another option would be to conduct a linking study, which entails administering an NRT and the state's customized test to the same sample of students, so that the two tests can be linked or equated. Once linked, students' results on the state test could be reported in terms of national percentile ranks derived from the NRT. A variant of this approach would be to link the customized state test results to the state's performance on the National Assessment of Education Progress (NAEP) and construct national percentile ranks through the relationship between state and national NAEP performance data. This approach, though relatively straightforward, has some technical problems that are beyond the scope of this paper. Assuming those technical issues could be addressed satisfactorily, there is a significant issue of equating two tests that do not measure the same content. This difference in content coverage between the tests is one of the primary reasons why use of an unaugmented NRT for NCLB is not acceptable. Additionally, by establishing a linkage between the NRT and a state's standards-based test, comparisons are being

made between two groups—students within the state and students across the nation—with different experiences with respect to the content covered by the standards-based state test. This is a questionable practice.

It is important to note that different norming approaches address different norming questions. For example, comparing a student's performance to that of a national sample on a state's own mathematics test addresses the question of how the student compares if a nationally representative sample took this particular state test, as compared to linking existing norms to the state test, which addresses the question of how the student compares to a national sample on mathematics in general.

It is important to note that different norming approaches address different norming questions.

Vertical Scaling

Scaled Scores

Before dealing with the issues associated with vertical scaling, it is important to understand the concept of scaled scores. Just as temperature can be reported either on the Celsius or Fahrenheit scale or distance can be measured either in miles or kilometers, test results can be reported on a variety of different metrics. Very often, it is necessary to make comparisons of performance on different tests. These might be different tests measuring the same standards administered at different times, or they might be tests on entirely different subjects. Suppose a student earns 28 points on a reading test and 35 points on a math test. Based on this "raw score" information only, one might conclude that the student did better on the math test. However, tests can differ in length and difficulty, among many other things.

A distribution of test scores can be described, in part, by an average or mean test score and a standard deviation, which is a measure of the spread of a distribution. Suppose that, when administered to the same group of students, the reading test had a mean of 24 points and a standard deviation of 4, and the math test had a mean of 40 points and a standard deviation of 10. That means the student scored one standard deviation above the mean (+1.0) on the reading test and one-half standard deviation below the mean (–0.5) on the math test. (The two raw scores have been converted to a common scale—the standard deviation scale.) The scores, +1.0 and –0.5, are called z-scores, which are one type of scaled score. Looking at these two scores, one would more correctly conclude that the student performed better on the reading test. Obviously the math test was longer and/or easier.

Scales other than the z-score scale might be used. For instance, using a scale with a mean of 250 and a standard deviation of 50, the reading score, one standard deviation above the mean, would be 300. The math score, one-half standard deviation below the mean, would be 225. Many state testing programs use these kinds of scaled scores. The only difference is that for those programs, the first transformation of raw scores is not done by computing z-scores. Instead a far more sophisticated technique based on something called Item Response Theory is used.

Vertical Scales

Transforming scores to different common scales does not change the scores' relative position within their distributions. Nevertheless, it is useful to convert scores to a common scale, to compare scores across time or subjects. Many states scale their scores separately for each grade. Thus, in the first year of a program, the mean score might be set at 250 and the standard deviation 50 at each grade level. Assuming the tests used at a grade level in different years are equated, an average score higher than 250 in subsequent years at a particular grade might show that program improvements in a school are working and performance in that grade is improving. If there were no change in performance, then the mean for the grade would remain at 250.

One problem with scaling scores separately at each grade is that the resulting test scores mask the gains students really are making as they progress through the grades. For example, a student scoring at the state average two years in a row would get scores of 250 each year if the state used the approach described above. Yet, the student has almost certainly learned a great deal over that one-year period. Because of interest in a student's progress in a year, "vertical scales" are sometimes used.

One way to obtain the data to place tests at different grade levels on one unified scale is to have students answer test questions intended for grade levels above or below the students' current grade. Assume that scale ranged from 300 to 900, for example. Fourth graders' scores might be clustered in the 400s, although the full range of fourth graders' scores would go well below 400 and above 500. Seventh graders' scores might be clustered in the 700s but would range well outside of the 700s. Using a vertical scale such as this, a particular student's "growth" would be seen in his or her test scores in successive years.

Concerns about Vertical Scales

As suggested earlier, students in a particular grade have ability levels that vary significantly. Thus, on a vertically scaled test, it would not be unusual for a high-achieving fourth grader to receive a score in the area of the vertical scale where the sixth or seventh graders' scores are clustered. If grade-level equivalents were reported (these are not used as much as they used to be), this would suggest that the student is operating at the sixth- or seventh-grade level. Don't believe it! The process of creating a vertical scale, the administration of mixed-grade items, and psychometric scaling itself spreads students' scores out considerably. However, the fact is that a fourth grader scoring similarly to a large number of sixth or seventh graders has neither been exposed to nor tested on the bulk of sixth- or seventh-grade content.

. . . a fourth grader scoring similarly to a large number of sixth or seventh graders has neither been exposed to nor tested on the bulk of sixth- or seventh-grade content.

The process of vertically scaling different grades' tests is actually test equating. Equating of tests of the same subject intended for use in different years at the same grade is clearly appropriate and advisable. Equating adjusts for small differences in difficulty, so that a particular scaled score means the same in one year as it does in another in terms of what can be concluded about the student's proficiency. Equating is clearly appropriate for tests measuring the same content. However, many measurement experts would argue that a fourth-grade math test and an eighth-grade math test hardly measure the same content, and that vertically equating them is as meaningful as equating a reading test and a math test.*

It is desirable to track student progress in a subject across years. However, results from vertically scaled tests can be misleading. There is a variety of effective alternatives to vertical scaling. One involves examining student performance on a test relative to predicted performance. With good longitudinal test data, analyses can be done that allow the computation of a predicted score for a student, based on his or her previous-year performance. A score that is higher or lower than the predicted score indicates whether the student is falling short of, achieving, or exceeding typical, expected growth. Another alternative to vertical scaling involves longitudinally assessing how far above or below a performance standard a student scores in successive years. For example, if a student achieves a score that is just above the "proficient" cut-score in grade three, then achieves a score substantially above "proficient" in grade four, it would be fair to conclude that student made substantial gains in grade four. To properly implement this alternative, it is important to pay special attention to the process of determining cut-scores to ensure consistency in the meanings of "proficient" across grades and that this consistency is reflected in cut-scores.

Value-Added Model

Standards-based tests compare a student's performance to a pre-established performance standard. Norm-referenced test results compare a student's performance to that of his or her peers. Test results based on the value-added model compare a student's performance to his or her previous performance. The simplest manifestation of the value-added model is pre-/post-testing (two administrations of the same or equivalent test forms) to determine the gains a student makes over the course of a school year.

* Independently transforming reading and math scores to a common scale as described at the beginning of this section is not the same as equating. That is just making it easier to compare where in the two tests' distributions the two scores fell. Equating is a far more sophisticated merging of item results, ultimately treating the two tests as if they were really one big test measuring the same construct.

Some programs call for summative testing of students in every grade. The value-added model can work with one administration per year. For example, suppose every grade is tested in the late spring. Then the scores for the students in a particular grade can serve as post-test scores when compared to the scores of those same students from the previous year at the preceding grade. Those same scores can serve as pre-test scores a year later, when there are scores for the same students the next spring. Of course, different grade-level tests are not equivalent forms. Therefore, the tests have to be either vertically equated, or some alternative must be implemented, as explained in the previous section.

More sophisticated applications of the value-added model bring other factors into the picture. For example, factors other than pre-test scores may be used in the computation of predicted scores against which post-test scores are compared. This would be a way of controlling for variables, such as socioeconomic factors, over which teachers and schools have no control. Some value-added systems actually pinpoint the contributions to gains of different factors, such as program characteristics, teachers, etc. In fact, one somewhat controversial system uses this practice as a teacher evaluation tool.

The value-added model does not produce results that satisfy NCLB requirements. Information concerning the magnitude of gains does not address where students fall on a proficiency continuum—an NCLB requirement. Furthermore, a system that adjusts for factors such as socioeconomic status, ethnicity, etc., would be inconsistent with a critical requirement of NCLB, which calls for the same standards for all students. However, assessment systems can be designed to report both value-added information, as well as the standards-based information mandated by NCLB.

Computer-Adaptive Tests

Computer-adaptive testing can be very efficient if the sole purpose of a testing program is to place individual students on a proficiency continuum with respect to a single, underlying trait (e.g., math proficiency) as do total scores on a typical test. Test items of different difficulty discriminate at different places on such a continuum. For example, a difficult item may differentiate an "A" student from a "B" student with respect to the subject of the test because the "A" student would answer the item correctly and the "B" student (or "C" or "D" student) would answer the item incorrectly. An easy item may discriminate between a "C" student and a "D" student because the "C" student would answer the item correctly while the "D" student would answer the item incorrectly. Traditionally, all students would respond to all items on a test to attain total test scores. That means that for any particular student, given the limited purpose of placing the student on the performance continuum, most of the items do not provide useful information.

. . . computer-adaptive tests were not designed to determine how well students are achieving the breadth and depth of knowledge embodied by the content standards.

Techniques associated with Item Response Theory allow estimates of proficiency to be made on the basis of whatever items a student answers. The more items a student answers that discriminate near where the student actually is on the proficiency continuum, the better the estimate of proficiency. Using computer-adaptive testing, it is possible to derive a proficiency score for a student based on relatively few items. This is accomplished by the computer picking a question for the student to answer based on the student's success (or lack thereof) on the previous question. In this way, the test can zero in on the student's proficiency without requiring him or her to answer many questions that are either far too easy or far too difficult for that particular student.

Computer-adaptive tests do not satisfy NCLB requirements, first, because the students do not all take the "same test" as required by the law. More importantly, the tests do not adequately cover state content standards, a key NCLB requirement. In short, computer-adaptive tests were not designed to determine how well students are achieving the breadth and depth of knowledge embodied by the content standards.

Computer-adaptive testing programs may produce subtest scores, but for such scores to be considered as useful diagnostic information, the test would have to be much longer. The

efficiencies gained by adaptively targeting a test to a student's ability level would be lost, for all intents and purposes. It may not even be known which skills were measured by the items answered by a particular student. Furthermore, if computers use an item selection process geared to measuring proficiency in the total test subject area, then it is very questionable what subtest scores even estimate. Finally, computer-adaptive tests generally are all multiple-choice. If the skills assessed by constructed-response questions are of value to a state (and part of the state's content standards), the alignment problem is even more significant. Nevertheless, computer-adaptive testing is a legitimate, effective approach for the particular purpose it was designed to achieve.

Diagnostic Tests

The topic here is not diagnostic tests per se, but rather the diagnostic value of large-scale assessments. A diagnostic test provides detailed information about a student's strengths and weaknesses with respect to specific skills and concepts. Generally, large-scale assessments are not designed to be diagnostic. They usually feature breadth of coverage of a domain and therefore have very few items addressing one particular skill or specific concept. It is interesting that NCLB, with all its other requirements for state assessments, also expects them to be diagnostic.

Generally, large-scale assessments are not designed to be diagnostic.

Sometimes a testing program points to reported subtest scores as diagnostic information. The reality is that a few subtest scores hardly constitute valid diagnostic information. They provide some insight into relative strengths and weaknesses of students, but only in terms of very general capabilities in the subdomains they address. Again, items in a subtest generally are selected to provide breadth of coverage of a subdomain. Subscores can point to areas where more assessment is needed, much in the way that a blood pressure reading from a drug store machine points to the need for further medical examination. Additionally, one must keep in mind that the internal consistency (a form of reliability) of a test is affected significantly by test length (the number of items). If a total test is constructed with 50 to 70 items, so that the reliability is at an acceptable level, then a 10- or 20-item subtest will have a considerably lower reliability. In other words, subtest scores are seldom reliable enough to make firm diagnostic evaluations. One commercial testing program reported a spelling score (a sub-subtest score) based on three items. For an area such as spelling, how much faith can one have in the results of such a subtest?

While the scores of a short subtest may yield very limited information, more valuable information can be gleaned from a single test item if the item itself is available with its results. Seeing that a student or a large percentage of a group of students selects "2/6" as the answer to "1/2 + 1/4 = ?" tells teachers a lot. They don't need to see the results from many more items to determine that there is a problem and just what that problem is. Thus, the way a large-scale assessment can probably be most diagnostic is if the test items are released along with the results.

Customized vs. Off-the-Shelf Tests

Even though NCLB allows the use of augmented NRTs, one might conclude that the law encourages statewide assessment programs that are customized. Obviously, item sets expressly developed to address a state's content standards and more specific grade-level expectations provide the best alignment possible. In addition, customized programs permit the mix of multiple-choice and constructed-response items most consistent with the values of the individual states' education leaders.

Many states have chosen customized designs for their assessment programs, in many cases using different tests each year. This practice eliminates security problems associated with repeated high-stakes use of the same off-the-shelf products. Sometimes, many years' tests are developed early in the program and after they are used up, another mega-round of development is required. Sometimes, a design is used that involves common items, which are answered by all students and used for the student and school accountability results, and matrix-sampled items, which are spread out across different forms of the test. This embedding of additional items with the common items allows the matrix-sampled items to be field-tested as part of operational testing. These matrix-sampled items provide a pool from which the next year's common items can be drawn. This approach requires ongoing development work during all years

the program is in place—a positive practice if teachers and other educators within the state are involved. This affords them a sense of ownership in the assessment and enables them to become the program's best ambassadors.

Some states release all or some of their used common items to the field, along with test results. As suggested above, the release of items with results provides educators with extremely useful information for improving the instruction of individuals and groups of students. If some common items are not released, they are often returned to a pool so that they may be selected for use in later years' tests.

Customized programs involve more work and greater expense than programs using off-the-shelf products. However, for NCLB testing, the differences between the two approaches lessen somewhat because of the requirement that off-the-shelf products be augmented. Some concerns exist that the construction of annual tests in relatively short time frames may compromise the quality of customized tests. This generally is not the case. As suggested earlier, test reliability is primarily a function of test design and length. Customized statewide assessments typically have the same level of reliability as the off-the-shelf products. In terms of content validity, the customized tests have the edge. One might think that with annual customized tests, the potential for errors is greater in the tests themselves and their results. Interestingly, programs using off-the-shelf tests have had their share of problems as a result of analysis and reporting errors. Of course, with the requirement of item augmentation for alignment, additional psychometric work is required to scale the additional items in with the existing test's items, so that in many ways, augmented, off-the-shelf tests are really semi-customized.

Conclusion

Many years ago, before the demand for large-scale assessments associated with reform acts and accountability laws, test selection was a local school or district decision. Choices were generally based on cost and simple alignment of test content to the local systems' instructional objectives. Tests, not program designs, were being selected. Nowadays, for large-scale assessments, there are many options and approaches that can be used. Thus, decision makers have a great deal more to consider when planning a program. Cost obviously remains an important concern, but given the variety of approaches available, it is more important than ever to carefully evaluate the various designs in light of program purposes. Knowledge of the options for assessment programs—what they are designed to do, what they are not—is essential for making good decisions. Hopefully, this overview will be of use to the decision makers or to others with input to or influence over those decisions.

Stuart Kahl is president and chief executive officer and Kevin Sweeney is director of measurement, design, and analysis for Measured Progress. Founded twenty years ago as Advanced Systems in Measurement and Evaluation, Measured Progress provides highly customized, standards-based assessments, and professional development tools and services to clients across the nation.

It's all about student learning. Period.

171 Watson Road, Dover, NH 03820
800.431.8901
www.measuredprogress.org

Formative Assessment

Powerful, heavily researched, and affordable

In October 1998 Phi Delta Kappa International published an article by Paul Black and Dylan Wiliam [included in this STAR report: "Inside the Black Box…"] that summarized the results of a meta-analysis of international research concerning classroom assessment practices. Shortly thereafter, perhaps in response to the solid evidence (indeed, more than for any other form of assessment) demonstrating that formative assessment practices promote learning—among all, but particularly of low-performing, students, confusion started to reign (and continues to this day) about what formative assessment is.

The confusion isn't merely a matter of terminology. If educators think that using an off-the-shelf test in and of itself is practicing formative assessment, they will miss out on the power of formative assessment practices that the Black and Wiliam analysis highlighted. In its attempt to avoid such an outcome, the Council of Chief State School Officers (CCSSO) convened a steering committee of national experts and in 2007 published the following definition of formative assessment:

> *"Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes."*

CCSSO and others have subsequently expanded upon this definition, typically citing the key elements in the process:

1. Teachers ensuring students understand the learning targets and the criteria for success,

2. Teachers gathering rich evidence of student learning by a variety of means (e.g., observation, questioning, quizzes),

3. Teachers providing descriptive feedback on gaps in student learning (for teachers to provide the most effective feedback and make the most appropriate adjustments in instruction, they need to be knowledgeable about the learning progressions associated with the learning targets),

4. Teachers and students using the feedback to adjust instruction and learning activities,

5. Students engaging in self-assessment and meta-cognitive reflection, and

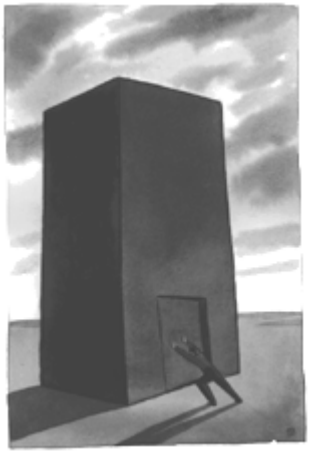
6. Teachers activating other students as resources.

From this simple statement and the key strategies/steps involved in the process, it's easy to see that formative assessment is essentially good teaching. So, with nothing more than an investment in professional learning and coaching, teachers can embed it in their practice.

The following examples illustrate common mistakes—opportunities where informed assessment practices can help educators do more with less:

Vignette: A teacher believes a test composed of randomly selected multiple-choice items from a large item bank is formative because the results are reported immediately upon completion of the test.

Vignette: A legislator believes formative assessments should count toward NCLB adequate yearly progress.



Inside the Black Box: Raising Standards Through Classroom Assessment

By Paul Black and Dylan Wiliam

Firm evidence shows that formative assessment is an essential component of classroom work and that its development can raise standards of achievement, Mr. Black and Mr. Wiliam point out. Indeed, they know of no other way of raising standards for which such a strong prima facie case can be made.

RAISING the standards of learning that are achieved through schooling is an important national priority. In recent years, governments throughout the world have been more and more vigorous in making changes in pursuit of this aim. National, state, and district standards; target setting; enhanced programs for the external testing of students' performance; surveys such as NAEP (National Assessment of Educational Progress) and TIMSS (Third International Mathematics and Science Study); initiatives to improve school planning and management; and more frequent and thorough inspection are all means toward the same end. But the sum of all these reforms has not added up to an effective policy because something is missing.

Learning is driven by what teachers and pupils do in classrooms. Teachers have to manage complicated and demanding situations, channeling the personal, emotional, and social pressures of a group of 30 or more youngsters in order to help them learn immediately and become better learners in the future. Standards can be raised only if teachers can tackle this task more effectively. What is missing from the efforts alluded to above is any direct help with this task. This fact was recognized in the TIMSS video study: "A focus on standards and accountability that ignores the processes of teaching and learning in classrooms will not provide the direction that teachers need in their quest to improve."1

In terms of systems engineering, present policies in the U.S. and in many other countries seem to treat the classroom as a black box. Certain inputs from the outside -- pupils, teachers, other resources, management rules and requirements, parental anxieties, standards, tests with high stakes, and so on -- are fed into the box. Some outputs are supposed to follow: pupils who are more knowledgeable and competent, better test results, teachers who are reasonably satisfied, and so on. But what is happening inside the box? How can anyone be sure that a particular set of new inputs will produce better outputs if we don't at least study what happens inside? And why is it that most of the reform initiatives mentioned in the first paragraph are not aimed

at giving direct help and support to the work of teachers in classrooms?

The answer usually given is that it is up to teachers: they have to make the inside work better. This answer is not good enough, for two reasons. First, it is at least possible that some changes in the inputs may be counterproductive and make it harder for teachers to raise standards. Second, it seems strange, even unfair, to leave the most difficult piece of the standards-raising puzzle entirely to teachers. If there are ways in which policy makers and others can give direct help and support to the everyday classroom task of achieving better learning, then surely these ways ought to be pursued vigorously.

This article is about the inside of the black box. We focus on one aspect of teaching: formative assessment. But we will show that this feature is at the heart of effective teaching.

The Argument

We start from the self-evident proposition that teaching and learning must be interactive. Teachers need to know about their pupils' progress and difficulties with learning so that they can adapt their own work to meet pupils' needs -- needs that are often unpredictable and that vary from one pupil to another. Teachers can find out what they need to know in a variety of ways, including observation and discussion in the classroom and the reading of pupils' written work.

We use the general term assessment to refer to all those activities undertaken by teachers -- and by their students in assessing themselves -- that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs.2

There is nothing new about any of this. All teachers make assessments in every class they teach. But there are three important questions about this process that we seek to answer:

- Is there evidence that improving formative assessment raises standards?
- Is there evidence that there is room for improvement?
- Is there evidence about how to improve formative assessment?

In setting out to answer these questions, we have conducted an extensive survey of the research literature. We have checked through many books and through the past nine years' worth of issues of more than 160 journals, and we have studied earlier reviews of research. This process yielded about 580 articles or chapters to study. We prepared a lengthy review, using material from 250 of these sources, that has been published in a special issue of the journal *Assessment in Education*, together with comments on our work by leading educational experts from Australia, Switzerland, Hong Kong, Lesotho, and the U.S.[3]

The conclusion we have reached from our research review is that the answer to each of the three questions above is clearly yes. In the three main sections below, we outline the nature and force of the evidence that justifies this conclusion. However, because we are presenting a summary here, our text will appear strong on assertions and weak on the details of their justification. We maintain that these assertions are backed by evidence and that this backing is set out in full detail in the lengthy review on which this article is founded.

We believe that the three sections below establish a strong case that governments, their agencies, school authorities, and the teaching profession should study very carefully whether they are seriously interested in raising standards in education. However, we also acknowledge widespread evidence that fundamental change in education can be achieved only slowly -- through programs of professional development that build on existing good practice. Thus we do not conclude that formative assessment is yet another "magic bullet" for education. The issues

involved are too complex and too closely linked to both the difficulties of classroom practice and the beliefs that drive public policy. In a final section, we confront this complexity and try to sketch out a strategy for acting on our evidence.

Does Improving Formative Assessment Raise Standards?

A research review published in 1986, concentrating primarily on classroom assessment work for children with mild handicaps, surveyed a large number of innovations, from which 23 were selected.[4] Those chosen satisfied the condition that quantitative evidence of learning gains was obtained, both for those involved in the innovation and for a similar group not so involved. Since then, many more papers have been published describing similarly careful quantitative experiments. Our own review has selected at least 20 more studies. (The number depends on how rigorous a set of selection criteria are applied.) All these studies show that innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains. These studies range over age groups from 5-year-olds to university undergraduates, across several school subjects, and over several countries.

For research purposes, learning gains of this type are measured by comparing the average improvements in the test scores of pupils involved in an innovation with the range of scores that are found for typical groups of pupils on these same tests. The ratio of the former divided by the latter is known as the *effect size*. Typical effect sizes of the formative assessment experiments were between 0.4 and 0.7. These effect sizes are larger than most of those found for educational interventions. The following examples illustrate some practical consequences of such large gains.

- An effect size of 0.4 would mean that the average pupil involved in an innovation would record the same achievement as a pupil in the top 35% of those not so involved.
- An effect size gain of 0.7 in the recent international comparative studies in mathematics[5] would have raised the score of a nation in the middle of the pack of 41 countries (e.g., the U.S.) to one of the top five.

Many of these studies arrive at another important conclusion: that improved formative assessment helps low achievers more than other students and so reduces the range of achievement while raising achievement overall. A notable recent example is a study devoted entirely to low-achieving students and students with learning disabilities, which shows that frequent assessment feedback helps both groups enhance their learning.[6] Any gains for such pupils could be particularly important. Furthermore, pupils who come to see themselves as unable to learn usually cease to take school seriously. Many become disruptive; others resort to truancy. Such young people are likely to be alienated from society and to become the sources and the victims of serious social problems.

Thus it seems clear that very significant learning gains lie within our grasp. The fact that such gains have been achieved by a variety of methods that have, as a common feature, enhanced formative assessment suggests that this feature accounts, at least in part, for the successes. However, it does not follow that it would be an easy matter to achieve such gains on a wide scale in normal classrooms. Many of the reports we have studied raise a number of other issues.

- All such work involves new ways to enhance feedback between those taught and the teacher, ways that will require significant changes in classroom practice.
- Underlying the various approaches are assumptions about what makes for

effective learning -- in particular the assumption that students have to be actively involved.
- For assessment to function formatively, the results have to be used to adjust teaching and learning; thus a significant aspect of any program will be the ways in which teachers make these adjustments.
- The ways in which assessment can affect the motivation and self-esteem of pupils and the benefits of engaging pupils in self-assessment deserve careful attention.

Is There Room for Improvement?

A poverty of practice. There is a wealth of research evidence that the everyday practice of assessment in classrooms is beset with problems and shortcomings, as the following selected quotations indicate.

- "Marking is usually conscientious but often fails to offer guidance on how work can be improved. In a significant minority of cases, marking reinforces underachievement and underexpectation by being too generous or unfocused. Information about pupil performance received by the teacher is insufficiently used to inform subsequent work," according to a United Kingdom inspection report on secondary schools.[7]
- "Why is the extent and nature of formative assessment in science so impoverished?" asked a research study on secondary science teachers in the United Kingdom.[8]
- "Indeed they pay lip service to [formative assessment] but consider that its practice is unrealistic in the present educational context," reported a study of Canadian secondary teachers.[9]
- "The assessment practices outlined above are not common, even though these kinds of approaches are now widely promoted in the professional literature," according to a review of assessment practices in U.S. schools.[10]
- The most important difficulties with assessment revolve around three issues. The first issue is *effective learning*.
- The tests used by teachers encourage rote and superficial learning even when teachers say they want to develop understanding; many teachers seem unaware of the inconsistency.
- The questions and other methods teachers use are not shared with other teachers in the same school, and they are not critically reviewed in relation to what they actually assess.
- For primary teachers particularly, there is a tendency to emphasize quantity and presentation of work and to neglect its quality in relation to learning.

The second issue is negative impact.

- The giving of marks and the grading function are overemphasized, while the giving of useful advice and the learning function are underemphasized.
- Approaches are used in which pupils are compared with one another, the prime purpose of which seems to them to be competition rather than personal improvement; in consequence, assessment feedback teaches low-achieving pupils that they lack "ability," causing them to come to believe that they are not able to learn.

The third issue is the *managerial role* of assessments.

- Teachers' feedback to pupils seems to serve social and managerial functions, often at the expense of the learning function.
- Teachers are often able to predict pupils' results on external tests because their own tests imitate them, but at the same

time teachers know too little about their pupils' learning needs.

- The collection of marks to fill in records is given higher priority than the analysis of pupils' work to discern learning needs; furthermore, some teachers pay no attention to the assessment records of their pupils' previous teachers.

Of course, not all these descriptions apply to all classrooms. Indeed, there are many schools and classrooms to which they do not apply at all. Nevertheless, these general conclusions have been drawn by researchers who have collected evidence -- through observation, interviews, and questionnaires -- from schools in several countries, including the U.S.

An empty commitment. The development of national assessment policy in England and Wales over the last decade illustrates the obstacles that stand in the way of developing policy support for formative assessment. The recommendations of a government task force in 198811 and all subsequent statements of government policy have emphasized the importance of formative assessment by teachers. However, the body charged with carrying out government policy on assessment had no strategy either to study or to develop the formative assessment of teachers and did no more than devote a tiny fraction of its resources to such work.12 Most of the available resources and most of the public and political attention were focused on national external tests. While teachers' contributions to these "summative assessments" have been given some formal status, hardly any attention has been paid to their contributions through formative assessment. Moreover, the problems of the relationship between teachers' formative and summative roles have received no attention.

It is possible that many of the commitments were stated in the belief that formative assessment was not problematic, that it already happened all the time and needed no more than formal acknowledgment of its existence. However, it is also clear that the political commitment to external testing in order to promote competition had a central priority, while the commitment to formative assessment was marginal. As researchers the world over have found, high-stakes external tests always dominate teaching and assessment. However, they give teachers poor models for formative assessment because of their limited function of providing overall summaries of achievement rather than helpful diagnosis. Given this fact, it is hardly surprising that numerous research studies of the implementation of the education reforms in the United Kingdom have found that formative assessment is "seriously in need of development."13 With hindsight, we can see that the failure to perceive the need for substantial support for formative assessment and to take responsibility for developing such support was a serious error.

In the U.S. similar pressures have been felt from political movements characterized by a distrust of teachers and a belief that external testing will, on its own, improve learning. Such fractured relationships between policy makers and the teaching profession are not inevitable -- indeed, many countries with enviable educational achievements seem to manage well with policies that show greater respect and support for teachers. While the situation in the U.S. is far more diverse than that in England and Wales, the effects of high-stakes state-mandated testing are very similar to those of the external tests in the United Kingdom. Moreover, the traditional reliance on multiple-choice testing in the U.S. -- not shared in the United Kingdom -- has exacerbated the negative effects of such policies on the quality of classroom learning.

How Can We Improve Formative Assessment?

The self-esteem of pupils. A report of schools in Switzerland states that "a number of pupils . . . are content to 'get by.' . . . Every teacher who wants to practice formative assessment must

reconstruct the teaching contracts so as to counteract the habits acquired by his pupils."14

The ultimate user of assessment information that is elicited in order to improve learning is the pupil. There are negative and positive aspects of this fact. The negative aspect is illustrated by the preceding quotation. When the classroom culture focuses on rewards, "gold stars," grades, or class ranking, then pupils look for ways to obtain the best marks rather than to improve their learning. One reported consequence is that, when they have any choice, pupils avoid difficult tasks. They also spend time and energy looking for clues to the "right answer." Indeed, many become reluctant to ask questions out of a fear of failure. Pupils who encounter difficulties are led to believe that they lack ability, and this belief leads them to attribute their difficulties to a defect in themselves about which they cannot do a great deal. Thus they avoid investing effort in learning that can lead only to disappointment, and they try to build up their self-esteem in other ways.

The positive aspect of students' being the primary users of the information gleaned from formative assessments is that negative outcomes -- such as an obsessive focus on competition and the attendant fear of failure on the part of low achievers -- are not inevitable. What is needed is a culture of success, backed by a belief that all pupils can achieve. In this regard, formative assessment can be a powerful weapon if it is communicated in the right way. While formative assessment can help all pupils, it yields particularly good results with low achievers by concentrating on specific problems with their work and giving them a clear understanding of what is wrong and how to put it right. Pupils can accept and work with such messages, provided that they are not clouded by overtones about ability, competition, and comparison with others. In summary, the message can be stated as follows: feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparisons with other pupils.

Self-assessment by pupils. Many successful innovations have developed self- and peer-assessment by pupils as ways of enhancing formative assessment, and such work has achieved some success with pupils from age 5 upward. This link of formative assessment to self-assessment is not an accident; indeed, it is inevitable.

To explain this last statement, we should first note that the main problem that those who are developing self-assessments encounter is not a problem of reliability and trustworthiness. Pupils are generally honest and reliable in assessing both themselves and one another; they can even be too hard on themselves. The main problem is that pupils can assess themselves only when they have a sufficiently clear picture of the targets that their learning is meant to attain. Surprisingly, and sadly, many pupils do not have such a picture, and they appear to have become accustomed to receiving classroom teaching as an arbitrary sequence of exercises with no overarching rationale. To overcome this pattern of passive reception requires hard and sustained work. When pupils do acquire such an overview, they then become more committed and more effective as learners. Moreover, their own assessments become an object of discussion with their teachers and with one another, and this discussion further promotes the reflection on one's own thinking that is essential to good learning.

Thus self-assessment by pupils, far from being a luxury, is in fact an essential component of formative assessment. When anyone is trying to learn, feedback about the effort has three elements: recognition of the desired goal, evidence about present position, and some understanding of a way to close the gap between the two.15 All three must be understood to some degree by anyone before he or she can take action to improve learning.

Such an argument is consistent with more general ideas established by research into the way people learn. New understandings are not simply swallowed and stored in isolation; they have to be assimilated in relation to preexisting ideas. The new and the old may be inconsistent or even in conflict, and the disparities must be resolved by thoughtful actions on the part of the learner. Realizing that there are new goals for the learning is an essential part of this process of assimilation. Thus we conclude: if formative assessment is to be productive, pupils should be trained in self-assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve.

The evolution of effective teaching. The research studies referred to above show very clearly that effective programs of formative assessment involve far more than the addition of a few observations and tests to an existing program. They require careful scrutiny of all the main components of a teaching plan. Indeed, it is clear that instruction and formative assessment are indivisible.

To begin at the beginning, the choice of tasks for classroom work and homework is important. Tasks have to be justified in terms of the learning aims that they serve, and they can work well only if opportunities for pupils to communicate their evolving understanding are built into the planning. Discussion, observation of activities, and marking of written work can all be used to provide those opportunities, but it is then important to look at or listen carefully to the talk, the writing, and the actions through which pupils develop and display the state of their understanding. Thus we maintain that opportunities for pupils to express their understanding should be designed into any piece of teaching, for this will initiate the interaction through which formative assessment aids learning.

Discussions in which pupils are led to talk about their understanding in their own ways are important aids to increasing knowledge and improving understanding. Dialogue with the teacher provides the opportunity for the teacher to respond to and reorient a pupil's thinking. However, there are clearly recorded examples of such discussions in which teachers have, quite unconsciously, responded in ways that would inhibit the future learning of a pupil. What the examples have in common is that the teacher is looking for a particular response and lacks the flexibility or the confidence to deal with the unexpected. So the teacher tries to direct the pupil toward giving the expected answer. In manipulating the dialogue in this way, the teacher seals off any unusual, often thoughtful but unorthodox, attempts by pupils to work out their own answers. Over time the pupils get the message: they are not required to think out their own answers. The object of the exercise is to work out -- or guess -- what answer the teacher expects to see or hear.

A particular feature of the talk between teacher and pupils is the asking of questions by the teacher. This natural and direct way of checking on learning is often unproductive. One common problem is that, following a question, teachers do not wait long enough to allow pupils to think out their answers. When a teacher answers his or her own question after only two or three seconds and when a minute of silence is not tolerable, there is no possibility that a pupil can think out what to say.

There are then two consequences. One is that, because the only questions that can produce answers in such a short time are questions of fact, these predominate. The other is that pupils don't even try to think out a response. Because they know that the answer, followed by another question, will come along in a few seconds, there is no point in trying. It is also generally the case that only a few pupils in a class answer the teacher's questions. The rest then leave it to these few, knowing that they cannot respond as quickly and being unwilling to risk making mistakes in public. So the teacher, by lowering the level of questions and by accepting answers from a few, can keep the lesson going but is

actually out of touch with the understanding of most of the class. The question/answer dialogue becomes a ritual, one in which thoughtful involvement suffers.

There are several ways to break this particular cycle. They involve giving pupils time to respond; asking them to discuss their thinking in pairs or in small groups, so that a respondent is speaking on behalf of others; giving pupils a choice between different possible answers and asking them to vote on the options; asking all of them to write down an answer and then reading out a selected few; and so on. What is essential is that any dialogue should evoke thoughtful reflection in which all pupils can be encouraged to take part, for only then can the formative process start to work. In short, the dialogue between pupils and a teacher should be thoughtful, reflective, focused to evoke and explore understanding, and conducted so that all pupils have an opportunity to think and to express their ideas.

Tests given in class and tests and other exercises assigned for homework are also important means of promoting feedback. A good test can be an occasion for learning. It is better to have frequent short tests than infrequent long ones. Any new learning should first be tested within about a week of a first encounter, but more frequent tests are counterproductive. The quality of the test items -- that is, their relevance to the main learning aims and their clear communication to the pupil -- requires scrutiny as well. Good questions are hard to generate, and teachers should collaborate and draw on outside sources to collect such questions.

Given questions of good quality, it is essential to ensure the quality of the feedback. Research studies have shown that, if pupils are given only marks or grades, they do not benefit from the feedback. The worst scenario is one in which some pupils who get low marks this time also got low marks last time and come to expect to get low marks next time. This cycle of repeated failure becomes part of a shared belief between such students and their teacher. Feedback has been shown to improve learning when it gives each pupil specific guidance on strengths and weaknesses, preferably without any overall marks. Thus the way in which test results are reported to pupils so that they can identify their own strengths and weaknesses is critical. Pupils must be given the means and opportunities to work with evidence of their difficulties. For formative purposes, a test at the end of a unit or teaching module is pointless; it is too late to work with the results. We conclude that the feedback on tests, seatwork, and homework should give each pupil guidance on how to improve, and each pupil must be given help and an opportunity to work on the improvement.

All these points make clear that there is no one simple way to improve formative assessment. What is common to them is that a teacher's approach should start by being realistic and confronting the question "Do I really know enough about the understanding of my pupils to be able to help each of them?"

Much of the work teachers must do to make good use of formative assessment can give rise to difficulties. Some pupils will resist attempts to change accustomed routines, for any such change is uncomfortable, and emphasis on the challenge to think for yourself (and not just to work harder) can be threatening to many. Pupils cannot be expected to believe in the value of changes for their learning before they have experienced the benefits of such changes. Moreover, many of the initiatives that are needed take more class time, particularly when a central purpose is to change the outlook on learning and the working methods of pupils. Thus teachers have to take risks in the belief that such investment of time will yield rewards in the future, while "delivery" and "coverage" with poor understanding are pointless and can even be harmful.

Teachers must deal with two basic issues that are the source of many of the problems associated with changing to a system of formative assessment. The first is the nature of each teacher's beliefs about learning. If the

teacher assumes that knowledge is to be transmitted and learned, that understanding will develop later, and that clarity of exposition accompanied by rewards for patient reception are the essentials of good teaching, then formative assessment is hardly necessary. However, most teachers accept the wealth of evidence that this transmission model does not work, even when judged by its own criteria, and so are willing to make a commitment to teaching through interaction. Formative assessment is an essential component of such instruction. We do not mean to imply that individualized, one-on-one teaching is the only solution; rather we mean that what is needed is a classroom culture of questioning and deep thinking, in which pupils learn from shared discussions with teachers and peers. What emerges very clearly here is the indivisibility of instruction and formative assessment practices.

The other issue that can create problems for teachers who wish to adopt an interactive model of teaching and learning relates to *the beliefs teachers hold about the potential of all their pupils for learning*. To sharpen the contrast by overstating it, there is on the one hand the "fixed I.Q." view -- a belief that each pupil has a fixed, inherited intelligence that cannot be altered much by schooling. On the other hand, there is the "untapped potential" view -- a belief that starts from the assumption that so-called ability is a complex of skills that can be learned. Here, we argue for the underlying belief that all pupils can learn more effectively if one can clear away, by sensitive handling, the obstacles to learning, be they cognitive failures never diagnosed or damage to personal confidence or a combination of the two. Clearly the truth lies between these two extremes, but the evidence is that *ways of managing formative assessment that work with the assumptions of "untapped potential" do help all pupils to learn and can give particular help to those who have previously struggled*.

Policy and Practice

Changing the policy perspective. The assumptions that drive national and state policies for assessment have to be called into question. The promotion of testing as an important component for establishing a competitive market in education can be very harmful. The more recent shifting of emphasis toward setting targets for all, with assessment providing a touchstone to help check pupils' attainments, is a more mature position. However, we would argue that *there is a need now to move further, to focus on the inside of the "black box" and so to explore the potential of assessment to raise standards directly as an integral part of each pupil's learning work.*

It follows from this view that several changes are needed. First, policy ought to start with a recognition that the prime locus for raising standards is the classroom, so that the overarching priority has to be the promotion and support of change within the classroom. Attempts to raise standards by reforming the inputs to and measuring the outputs from the black box of the classroom can be helpful, but they are not adequate on their own. Indeed, their helpfulness can be judged only in light of their effects in classrooms.

The evidence we have presented here establishes that a clearly productive way to start implementing a classroom-focused policy would be to improve formative assessment. This same evidence also establishes that in doing so we would not be concentrating on some minor aspect of the business of teaching and learning. Rather, we would be concentrating on several essential elements: the quality of teacher/pupil interactions, the stimulus and help for pupils to take active responsibility for their own learning, the particular help needed to move pupils out of the trap of "low achievement," and the development of the habits necessary for all students to become lifelong learners. Improvements in formative assessment, which are within the reach of all teachers, can contribute substantially to raising standards in all these ways.

Four steps to implementation. If we accept the argument outlined above, what needs to be done? The proposals outlined below do not

follow directly from our analysis of assessment research. They are consistent with its main findings, but they also call on more general sources for guidance.[16]

At one extreme, one might call for more research to find out how best to carry out such work; at the other, one might call for an immediate and large-scale program, with new guidelines that all teachers should put into practice. Neither of these alternatives is sensible: while the first is unnecessary because enough is known from the results of research, the second would be unjustified because not enough is known about classroom practicalities in the context of any one country's schools.

Thus the improvement of formative assessment cannot be a simple matter. There is no quick fix that can alter existing practice by promising rapid rewards. On the contrary, if the substantial rewards promised by the research evidence are to be secured, each teacher must find his or her own ways of incorporating the lessons and ideas set out above into his or her own patterns of classroom work and into the cultural norms and expectations of a particular school community.[17] This process is a relatively slow one and takes place through sustained programs of professional development and support. This fact does not weaken the message here; indeed, it should be seen as a sign of its authenticity, for lasting and fundamental improvements in teaching and learning must take place in this way. A recent international study of innovation and change in education, encompassing 23 projects in 13 member countries of the Organisation for Economic Co-operation and Development, has arrived at exactly the same conclusion with regard to effective policies for change.[18] Such arguments lead us to propose a four-point scheme for teacher development.

1. *Learning from development.* Teachers will not take up ideas that sound attractive, no matter how extensive the research base, if the ideas are presented as general principles that leave the task of translating them into everyday practice entirely up to the teachers. Their classroom lives are too busy and too fragile for all but an outstanding few to undertake such work. What teachers need is a variety of living examples of implementation, as practiced by teachers with whom they can identify and from whom they can derive the confidence that they can do better. They need to see examples of what doing better means in practice.

   So changing teachers' practice cannot begin with an extensive program of training for all; that could be justified only if it could be claimed that we have enough "trainers" who know what to do, which is certainly not the case. The essential first step is to set up a small number of local groups of schools -- some primary, some secondary, some inner-city, some from outer suburbs, some rural -- with each school committed both to a school-based development of formative assessment and to collaboration with other schools in its local group. In such a process, the teachers in their classrooms will be working out the answers to many of the practical questions that the evidence presented here cannot answer. They will be reformulating the issues, perhaps in relation to fundamental insights and certainly in terms that make sense to their peers in other classrooms. It is also essential to carry out such development in a range of subject areas, for the research in mathematics education is significantly different from that in language, which is different again from that in the creative arts.

   The schools involved would need extra support in order to give their teachers time to plan the initiative in light of existing evidence, to reflect on their experience as it develops, and to offer advice about training others in the future. In addition, there would be a need for external evaluators to help the teachers with their development

work and to collect evidence of its effectiveness. Video studies of classroom work would be essential for disseminating findings to others.

2. *Dissemination.* This dimension of the implementation would be in low gear at the outset -- offering schools no more than general encouragement and explanation of some of the relevant evidence that they might consider in light of their existing practices. Dissemination efforts would become more active as results and resources became available from the development program. Then strategies for wider dissemination -- for example, earmarking funds for inservice training programs -- would have to be pursued.

   We must emphasize that this process will inevitably be a slow one. To repeat what we said above, *if the substantial rewards promised by the evidence are to be secured, each teacher must find his or her own ways of incorporating the lessons and ideas that are set out above into his or her own patterns of classroom work.* Even with optimum training and support, such a process will take time.

3. *Reducing obstacles.* All features in the education system that actually obstruct the development of effective formative assessment should be examined to see how their negative effects can be reduced. Consider the conclusions from a study of teachers of English in U.S. secondary schools.

   Most of the teachers in this study were caught in conflicts among belief systems and institutional structures, agendas, and values. The point of friction among these conflicts was assessment, which was associated with very powerful feelings of being overwhelmed, and of insecurity, guilt, frustration, and anger. . . . This study suggests that assessment, as it occurs in schools, is far from a merely technical problem. Rather, it is deeply social and personal.[19]

The chief negative influence here is that of short external tests. Such tests can dominate teachers' work, and, insofar as they encourage drilling to produce right answers to short, out-of-context questions, they can lead teachers to act against their own better judgment about the best ways to develop the learning of their pupils. This is not to argue that all such tests are unhelpful. Indeed, they have an important role to play in securing public confidence in the accountability of schools. For the immediate future, what is needed in any development program for formative assessment is to study the interactions between these external tests and formative assessments to see how the models of assessment that external tests can provide could be made more helpful.

All teachers have to undertake some summative assessment. They must report to parents and produce end-of-year reports as classes are due to move on to new teachers. However, the task of assessing pupils summatively for external purposes is clearly different from the task of assessing ongoing work to monitor and improve progress. Some argue that these two roles are so different that they should be kept apart. We do not see how this can be done, given that teachers must have some share of responsibility for the former and must take the leading responsibility for the latter.[20] However, teachers clearly face difficult problems in reconciling their formative and summative roles, and confusion in teachers' minds between these roles can impede the improvement of practice.

The arguments here could be taken much further to make the case that teachers should play a far greater role in contributing to summative assessments for accountability. One strong reason for giving

teachers a greater role is that they have access to the performance of their pupils in a variety of contexts and over extended periods of time.

This is an important advantage because sampling pupils' achievement by means of short exercises taken under the conditions of formal testing is fraught with dangers. It is now clear that performance in any task varies with the context in which it is presented. Thus some pupils who seem incompetent in tackling a problem under test conditions can look quite different in the more realistic conditions of an everyday encounter with an equivalent problem. Indeed, the conditions under which formal tests are taken threaten validity because they are quite unlike those of everyday performance. An outstanding example here is that collaborative work is very important in everyday life but is forbidden by current norms of formal testing.[21] These points open up wider arguments about assessment systems as a whole -- arguments that are beyond the scope of this article.

4. *Research.* It is not difficult to set out a list of questions that would justify further research in this area. Although there are many and varied reports of successful innovations, they generally fail to give clear accounts of one or another of the important details. For example, they are often silent about the actual classroom methods used, the motivation and experience of the teachers, the nature of the tests used as measures of success, or the outlooks and expectations of the pupils involved.

   However, while there is ample justification for proceeding with carefully formulated projects, we do not suggest that everyone else should wait for their conclusions. Enough is known to provide a basis for active development work, and some of the most important questions can be answered only through a program of practical implementation.

   Directions for future research could include a study of the ways in which teachers understand and deal with the relationship between their formative and summative roles or a comparative study of the predictive validity of teachers' summative assessments versus external test results. Many more questions could be formulated, and it is important for future development that some of these problems be tackled by basic research. At the same time, experienced researchers would also have a vital role to play in the evaluation of the development programs we have proposed.

Are We Serious About Raising Standards?

The findings summarized above and the program we have outlined have implications for a variety of responsible agencies. However, it is the responsibility of governments to take the lead. It would be premature and out of order for us to try to consider the relative roles in such an effort, although success would clearly depend on cooperation among government agencies, academic researchers, and school-based educators.

The main plank of our argument is that standards can be raised only by changes that are put into direct effect by teachers and pupils in classrooms. There is a body of firm evidence that formative assessment is an essential component of classroom work and that its development can raise standards of achievement. We know of no other way of raising standards for which such a strong prima facie case can be made. Our plea is that national and state policy makers will grasp this opportunity and take the lead in this direction.

[1.] James W. Stigler and James Hiebert, "Understanding and Improving Classroom Mathematics Instruction: An Overview of the TIMSS Video Study," *Phi Delta Kappan*, September 1997, pp. 19-20.

[2.] There is no internationally agreed-upon term here. "Classroom evaluation," "classroom assessment," "internal assessment," "instructional assessment," and "student assessment" have been used by different authors, and some of these terms have different meanings in different texts.

[3.] Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education*, March 1998, pp. 7-74.

[4.] Lynn S. Fuchs and Douglas Fuchs, "Effects of Systematic Formative Evaluation: A Meta-Analysis," *Exceptional Children*, vol. 53, 1986, pp. 199-208.

[5.] See Albert E. Beaton et al., *Mathematics Achievement in the Middle School Years* (Boston: Boston College, 1996).

[6.] Lynn S. Fuchs et al., "Effects of Task-Focused Goals on Low-Achieving Students with and Without Learning Disabilities," *American Educational Research Journal*, vol. 34, 1997, pp. 513-43.

[7.] OFSTED (Office for Standards in Education), *Subjects and Standards: Issues for School Development Arising from OFSTED Inspection Findings 1994-5: Key Stages 3 and 4 and Post-16* (London: Her Majesty's Stationery Office, 1996), p. 40.

[8.] Nicholas Daws and Birendra Singh, "Formative Assessment: To What Extent Is Its Potential to Enhance Pupils' Science Being Realized?," *School Science Review*, vol. 77, 1996, p. 99.

[9.] Clement Dassa, Jesús Vazquez-Abad, and Djavid Ajar, "Formative Assessment in a Classroom Setting: From Practice to Computer Innovations," *Alberta Journal of Educational Research*, vol. 39, 1993, p. 116.

[10.] D. Monty Neill, "Transforming Student Assessment," *Phi Delta Kappan*, September 1997, pp. 35-36.

[11.] *Task Group on Assessment and Testing: A Report* (London: Department of Education and Science and the Welsh Office, 1988).

[12.] Richard Daugherty, *National Curriculum Assessment: A Review of Policy, 1987-1994* (London: Falmer Press, 1995).

[13.] Terry A. Russell, Anne Qualter, and Linda McGuigan, "Reflections on the Implementation of National Curriculum Science Policy for the 5-14 Age Range: Findings and Interpretations from a National Evaluation Study in England," *International Journal of Science Education*, vol. 17, 1995, pp. 481-92.

[14.] Phillipe Perrenoud, "Towards a Pragmatic Approach to Formative Evaluation," in Penelope Weston, ed., *Assessment of Pupils' Achievement: Motivation and School Success* (Amsterdam: Swets and Zeitlinger, 1991), p. 92.

[15.] D. Royce Sadler, "Formative Assessment and the Design of Instructional Systems," *Instructional Science*, vol. 18, 1989, pp. 119-44.

[16.] Paul J. Black and J. Myron Atkin, *Changing the Subject: Innovations in Science, Mathematics, and Technology Education* (London: Routledge for the Organisation for Economic Co-operation and Development, 1996); and Michael G. Fullan, with Suzanne Stiegelbauer, *The New Meaning of Educational Change* (London: Cassell, 1991).

[17.] See Stigler and Hiebert, pp. 19-20.

[18.] Black and Atkin, op. cit.

[19.] Peter Johnston et al., "Assessment of Teaching and Learning in Literature-Based Classrooms," *Teaching and Teacher Education*, vol. 11, 1995, p. 359.

[20.] Dylan Wiliam and Paul Black, "Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment," *British Educational Research Journal*, vol. 22, 1996, pp. 537-48.

[21.] These points are developed in some detail in Sam Wineburg, "T. S. Eliot, Collaboration, and the Quandaries of Assessment in a Rapidly Changing World," *Phi Delta Kappan*, September 1997, pp. 59-65.

PAUL BLACK is professor emeritus in the School of Education, King's College, London, where DYLAN WILIAM is head of school and professor of educational assessment.

Revision of version seen in the October 26, 2005 issue of Education Week

Where in the World Are Formative Tests? Right under Your Nose!

Stuart Kahl, Ph.D.
Founding Principal
Measured Progress

If you're shopping around from vendor to vendor for a formative assessment tool, the tests you're seeing probably aren't optimal for formative use. Formative assessment is actually a process that teachers use during instruction to assess student grasp of the specific topics and skills they are teaching. It involves, among other instructional steps, evidence gathering, feedback, and instructional adjustment. The data gathering is a "midstream" check to identify specific student misconceptions and mistakes while the material is being taught and thus to guide subsequent learning activity.

Formative assessment evidence gathering is not designed to measure post-instruction mastery or to contribute to students' course grades. It is conducted before the teacher has moved on to new topics or more advanced skills. Therefore, when formative assessment evidence is gathered, students will not necessarily have reached the level of mastery they will achieve after further instruction.

The fact is that formative assessment is what teachers do on a daily basis. Teachers design and apply many evidence-gathering tools and techniques, which can take the form of classroom quizzes, worksheets, homework, projects, and portfolios, not to mention the teacher's observations of class work. A test vendor would have to take up permanent residence in the classroom in order to provide many formative assessment tools.

So what are the vendors selling? The preponderance of externally developed tests being offered are really summative, whether they are used as end-of-year assessments, early-warning tests, or benchmark assessments covering material taught in recent weeks or months. Typically, periodic summative testing helps to identify relative strengths and weaknesses of instructional programs, as well as students "at risk," to pinpoint the need for additional help before those students take subsequent, high-stakes summative assessments.

Many of the tests rely on multiple-choice questions, which many educators contend do not provide the kind of diagnostic information they need to adjust instruction to address the misconceptions and errors of individual students. Experience has shown that constructed-response questions, which enable teachers to see actual student work, lead to a better understanding of student reasoning and help teachers to detect misconceptions and erroneous practices.

When evaluating the formative assessment practices of teachers, there are three important things to consider. First, timing is critical. Formative assessment evidence gathering must take place during the time the content is being taught. The second key consideration is whether the assessment tools and techniques yield accurate, pertinent diagnostic information; does the student's work reveal specific misunderstandings with respect to the content and skills being taught? The third consideration might seem obvious, but it is at the heart of where attention concerning formative assessment should be directed. Is the diagnostic information being used to inform and adjust instruction for individual students?

Teachers are formative assessors. It's what they do every day. But if the results don't guide instructional practices and learning activities, then the exercise is virtually meaningless. Given that, it is just as important to invest in professional development that helps teachers gather and use diagnostic information as it is to purchase still more summative tests.

Formative Assessment and Professional Development: A Question of Mind(set) over (Subject) Matter

Stuart Kahl, Ph.D.
Founding Principal

Federal mandates, accountability requirements, increased NCLB flexibility, superior content standards, myriad testing tools, countless data management systems, and mountains of data—each can play a role in the noble effort to improve student achievement. However, all of them combined won't make much difference if we short-change the powerful link between student learning and the quality of classroom interactions.

Unfortunately in some quarters, certain attributes of good teaching are neglected because of a focus on teachers' academic backgrounds. People with this focus might assume that a nuclear physicist would make a great middle school physical science teacher just because of his or her expert content knowledge. A fundamental problem with this is that it fails to acknowledge that teaching is equal parts content, pedagogy, and relationships. The latter two are just as important as, and no easier to master than, the former. All three are absolutely critical to formative assessment, the instructional process that research shows can have a dramatic positive impact on achievement.

Clearly, professional development is the best vehicle for changing instruction, moving toward better formative assessment practices. Professional development comes in many shapes and sizes; but much of what's out there is not helpful. When kids get a day off so their teachers can hear a feel-good message from a motivational educator, little is accomplished. While it's no secret that such one-shot experiences are ineffective, more substantial professional development can be equally limited. In order to significantly alter teachers' interactions with students and assure that teachers successfully apply the principles of assessment for learning or formative assessment (not "testing"), two things have to happen: a change of mindset and whole-school involvement in the transformation.

Let's start with **mindset**. For teachers who are not already using effective formative assessment practices, the transformation they must undertake is significant. These changes pertain to everything from grading practices to the way they and their students spend instructional time. This is not an easy sell to teachers whose time is at a premium and whose instructional behaviors are well established. A strong support system is essential for teachers to learn that change in their instructional practice is both desirable and necessary and that change does not mean more work, just working differently.

Whole-school involvement is important for several reasons. First, teachers can't change some things on their own. Improving grading practices, time allocation, and the like requires the action and involvement of school leadership. Second, efficiencies and improved practices come from collaboration. For example, learning communities can offer teachers the opportunity to learn from one another as they share ideas for lessons, activities, and approaches.

Ongoing teacher and administrator training that allows educators to apply what they learn to improve their own practices is formative in and of itself. Year-long institutes, principal coaching models, and other approaches that sustain this support are the only ways to significantly change what's happening in many classrooms—and that's where achievement levels are raised.

It's time to moderate our obsession with standards, assessments, and accountability systems. We now need to shift our focus to providing teachers and administrators with appropriate and effective professional development.

What do you think?
Let us know at twocents@measuredprogress.org

ATTRIBUTES OF EFFECTIVE FORMATIVE ASSESSMENT

A WORK PRODUCT COORDINATED[1] BY SARAH MCMANUS
NC DEPARTMENT OF PUBLIC INSTRUCTION

Paper prepared for the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO)

[1] *Grateful thanks go to the various members of the FAST SCASS who contributed text and edits, and provided feedback on various iterations of this document.*

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Formative Assessment for Students and Teachers
State Collaborative on Assessment and Student Standards

The Council's State Collaborative on Assessment and Student Standards (SCASS) strives to provide leadership, advocacy and service in creating and supporting effective collaborative partnerships through the collective experience and knowledge of state education personnel to develop and implement high standards and valid assessment systems that maximize educational achievement for all children.

John Tanner, Director Center for Innovative Measures
Douglas Rindone and Duncan MacQuarrie, Co-Coordinators, FAST SCASS

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700

Washington, DC 20001-1431

Phone (202) 336-7000

Fax (202) 408-8072

www.ccsso.org

ATTRIBUTES OF EFFECTIVE FORMATIVE ASSESSMENT

A work product coordinated by Sarah McManus, NC Department of Public Instruction, for the Formative Assessment for Students and Teachers (FAST) Collaborative

Background of a Definition

There has been substantial interest in *formative assessment* among U.S. educators during recent years. Increasing numbers of educators regard formative assessment as a way not only to improve student learning, but also to increase student scores on significant achievement examinations. To promote the use of formative assessment, the Council of Chief State School Officers (CCSSO) created a national initiative. The initiative formally began in January 2006, when CCSSO formed the Formative Assessment (FA) Advisory Group consisting of measurement and education researchers including Jim Popham, Lorrie Shepard, Rick Stiggins, and Dylan Wiliam and state agency leaders from across the nation. (A complete list of FA Advisory Group members is at end of document.)

CCSSO also formed a new State Collaborative on Assessment and Student Standards (SCASS) to implement the vision of the FA Advisory Group. The first challenge for the Formative Assessment for Students and Teachers (FAST) SCASS was to work with the FA Advisory Group to review the various definitions of formative assessment and related research. The FA Advisory Group and FAST SCASS devoted substantial effort to clarify the meaning of "formative assessment," based on current literature, and determine how formative assessment may best be used by the nation's educators.

In October 2006, FAST SCASS educators representing approximately 25 states agreed on the definition of formative assessment presented in this document and it was subsequently approved by the FA Advisory Group. In the year following, the FAST SCASS and FA Advisory Group isolated the attributes that, based on the research and current literature, would render formative assessment most effective. This document presents the definition of formative assessment and identifies and explains the five attributes of effective formative assessment.

The Definition of Formative Assessment

During the October 2006, inaugural FAST SCASS meeting in Austin, Texas, the following definition of formative assessment was adopted, without dissent:

> ***Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.***

A Closer Look at the Definition

The primary purpose of the formative assessment process, as conceived in this definition, is to provide evidence that is used by teachers and students to inform instruction and learning during the teaching/learning process. Effective formative assessment involves collecting evidence about how student learning is progressing during the course of instruction so that necessary instructional adjustments can be made to close the gap between students' current understanding and the desired goals. Formative assessment is not an adjunct to teaching but, rather, integrated into instruction and learning with teachers and students receiving frequent feedback.

One key feature of this definition is its requirement that formative assessment be regarded as *a process* rather than a particular kind of assessment. In other words, there is no such thing as "a formative test." Instead, there are a number of formative assessment strategies that can be implemented during classroom instruction. These range from informal observations and conversations to purposefully planned instructionally embedded techniques designed to elicit evidence of student learning to inform and adjust instruction.

A second important part of the definition is its unequivocal requirement that the formative assessment process involve both teachers *and* students. The students must be actively involved in the systematic process intended to improve their learning. The process requires the teacher to share learning goals with students and provide opportunities for students to monitor their ongoing progress.

Attributes

There are five attributes that have been identified from the literature as critical features of effective formative assessment. No one of the following attributes should be regarded as a *sine qua non*, that is, an attribute without which the assessment would not be formative.

1.

Learning Progressions: Learning progressions should clearly articulate the sub-goals of the ultimate learning goal.

Learning progressions describe how concepts and skills build in a domain, and show the trajectory of learning along which students are expected to progress. From a learning progression teachers have the big picture of what students need to learn, as well as sufficient detail for planning instruction to meet short-term goals. They are able to connect formative assessment opportunities to the short-term goals to keep track of how well their students' learning is moving forward.

For example, at the earliest stages of a progression for historical inquiry students must learn how to investigate the past from a range of sources of information, (e.g., stories, eyewitness accounts, pictures, photographs, artifacts, historic buildings, museums, galleries, and technology-based sources). Students build on this learning in later stages of the progression to develop an understanding that people represent and interpret the past in different ways (e.g., through pictures, plays, films, reconstructions, museum displays, and fiction and nonfiction accounts), and that the interpretations reflect the intentions of those who make them (e.g., writers, archaeologists, historians, and filmmakers). A goal for students at each level of the progression would be to investigate a set of artifacts in increasingly sophisticated ways to extract information about a particular period or event in history. Not only would such investigations support the students' development of historical reasoning, they would also provide evidence of the students' ability to reason in increasingly complex ways. This involves moving from the early stages of reasoning based on simple observation to the more complex stages based on indirect observation and the synthesis of multiple sources of information. Using the evidence elicited from such tasks connected to the goals of the progression, a teacher could identify the "just right gap" – a growth point in learning that involves a step that is neither too large nor too small – and make adjustments to instruction accordingly.

2.

Learning Goals and Criteria for Success: Learning goals and criteria for success should be clearly identified and communicated to students.

Because the formative assessment process helps students achieve intended learning outcomes based on explicit learning progressions, teachers must first identify and then communicate the instructional goal to students. In addition to communicating the nature of the instructional goal, teachers must provide the criteria by which learning will be assessed so that students will know whether they are successfully progressing toward the goal. This information should be communicated using language readily understood by students, and may be accompanied by realistic examples of those that meet and do not meet the criteria.

For example, suppose the goal of a social studies instructional unit was to have students "prepare a written critique of the quality of arguments in political essays in a local newspaper's editorial pages." The teacher might first offer students a paraphrased version of that goal such as, "You will be able to judge the strengths and weaknesses of arguments in the editorials you find in our daily newspapers." The teacher would discuss the criteria for evaluating arguments and then provide several examples of critiques of political essays. This will provide students with a reasonably clear idea of the analytic skills they are to develop and also provide them with the tools required to assess their own written analyses.

3.

Descriptive Feedback: Students should be provided with evidence-based feedback that is linked to the intended instructional outcomes and criteria for success.

Descriptive feedback should be about the particular qualities of student learning with discussion or suggestions about what the student can do to improve. It should avoid comparisons with other pupils. Specific, timely feedback should be based on the learning goal and criteria for success. It should help the student answer three basic questions: Where am I going? Where am I now? How can I close the gap?

For example, in an eighth grade writing class the students are learning how to construct an argument. They are focusing specifically on speech-writing and have examined several effective speeches, both from prominent speech-makers in history and from previous years' eighth grade students. In this particular lesson, students have been asked to write an opening paragraph to their speech with the success criteria of introducing their topic in a way that engages the audience. The feedback the teacher gives to one student is, "The opening paragraph does not capture the audience's attention because it does not clearly state what the speech is about. However, the opening sentence of the second paragraph states your position with an effective contrast. What can you do to improve or strengthen your opening paragraph?" With this kind of descriptive feedback and collaboration, the teacher clarifies the goal for the student, provides specific information about where the student is in relation to meeting the criteria, and offers enough substantive

information to allow the student an opportunity to identify ways to move learning forward.

Similarly, in a sixth grade math class students working in groups have been asked to review an example of the steps a student from a previous year took to solve a problem. They must decide if the work is correct or incorrect and provide an explanation for their view. The success criterion that the teacher gives them is, "Include any properties or rules that may apply in your explanation." When the groups report back after their discussions, the teacher listens for the rules or properties in the explanations, and this becomes the focus of her feedback. To one group she says, "Your explanation shows me that you understand that the steps the student took to solve the problem were incorrect. Remember the success criterion. You must also relate your explanation to one of the properties we have been discussing in class to indicate the reason the steps were incorrect." Again, the students know the goal, where their response differed from the criteria, and how they can improve their explanations.

4.

Self- and Peer-Assessment: Both self- and peer-assessment are important for providing students an opportunity to think meta-cognitively about their learning.

Formative assessment is a process that directly engages both teachers and students. In addition to teacher feedback, when students and their peers are involved there are many more opportunities to share and receive feedback. Helping students think meta-cognitively about their own learning fosters the idea that learning is their responsibility and that they can take an active role in planning, monitoring, and evaluating their own progress. To support both self- and peer-assessment, the teacher must provide structure and support so students learn to be reflective of their own work and that of their peers, allowing them to provide meaningful and constructive feedback.

In self-assessment, students reflect on and monitor their learning using clearly explicated criteria for success. In peer-assessment, students analyze each others' work using guidelines or rubrics and provide descriptive feedback that supports continued improvement. For example, students can work in pairs to review each other's work to give feedback. A teacher needs to have modeled good feedback with students and talked about what acceptable and unacceptable comments look like in order to have created a safe learning environment. Students can use a rubric to provide feedback to a peer by articulating reasons why a piece of work is at one level and discussing how it could be improved to move it to the next level. Alternatively, feedback could be given using a format such as "two stars and a wish," which provides a structure for a student to identify two aspects of the work that are particularly strong (stars) and one aspect the peer might improve (a wish). Students then need time to reflect on the feedback they have received to make changes or improvements. In addition, students can be encouraged to be self-reflective by thinking about their own work based on what they learned from giving feedback to others. A further benefit of providing feedback to a peer is that it can help deepen the student's own learning. However, student- and peer-assessment should not be used in the formal grading process.

5.

Collaboration: A classroom culture in which teachers and students are partners in learning should be established.

Sharing learning goals and criteria for success with students, supporting students as they monitor and take responsibility for their own learning, helping students to provide constructive feedback to each other, and involving students in decisions about how to move learning forward are illustrations of students and teachers working together in the teaching and learning process.

However, for students to be actively and successfully involved in their own learning, they must feel that they are bona fide partners in the learning process. This feeling is dependent on a classroom culture characterized by a sense of trust between and among students and their teachers; by norms of respect, transparency, and appreciation of differences; and by a non-threatening environment. Creating such a culture requires teachers to model these behaviors during interactions with students, to actively teach the classroom norms, and to build the students' skills in constructive self- and peer-assessment. In this type of classroom culture, students will more likely feel they are collaborators with their teacher and peers in the learning process.

While evidence exists in varying degrees to support the five attributes presented, there is clearly no one best way to carry out formative assessment. The way these attributes are implemented depends on the particular instructional context, the individual teacher, and—perhaps most importantly—the individual students.

For examples on how to incorporate the five attributes into practice refer to the document Formative Assessment: Examples of Practice. [2]

Suggested Readings

Heritage, M. (February, 2008). Learning Progressions: Supporting Instruction and Formative Assessment. Council of Chief State School Officers: Washington DC.

[2] *Formative Assessment: Examples of Practice. Council of Chief State School Officers: Washington, DC 2008. A work product initiated and led by E. Caroline Wylie, ETS, for the Formative Assessment for Students and Teachers (FAST) Collaborative.*

CCSSO Formative Assessment Advisory Group[3]

ANNETTE BOHLING, AZ, SR. VICE PRESIDENT OF ACCREDITATION, ADVANCED, ARIZONA STATE UNIVERSITY

WILLIAM BUSHAW, IN, EXECUTIVE DIRECTOR, PHI DELTA KAPPA

DOUG CHRISTENSEN, NE, COMMISSIONER OF EDUCATION, NEBRASKA DEPARTMENT OF EDUCATION

ANGELA FAHERTY, ME DEPUTY ASSOCIATE SUPEINTENDENT, STANDARDS AND ASSESSMENT SECTION, MAINE DEPARTMENT OF EDUCATION

MARGARET HERITAGE, CA, ASSISTANT DIRECTOR FOR PROFESSIONAL DEVELOPMENT, NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS AND STUDENT TESTING, UNIVERSITY OF CALIFORNIA

GERUNDA HUGHES, DC, ASSOCIATE PROFESSOR, CURRICULUM AND INSTRUCTION PROGRAM COORDINATOR, SECONDARY EDUCATION, HOWARD UNIVERSITY

HENRY JOHNSON, FORMER ASSISTANT SECRETARY FOR ELEMENTARY AND SECONDARY EDUCATION, U.S. DEPARTMENT OF EDUCAITON

STUART KAHL, NH, PRESIDENT AND CEO, MEASURED PROGRESS

KEN KAY, AZ, PRESIDENT, PARTNERSHIP FOR 21ST CENTURY SKILLS

SARAH MCMANUS, NC, SECTION CHIEF, TESTING POLICY AND OPERATIONS, NORTH CAROLINA DEPARTMENT OF PUBLIC INSTRUCTION

BOB NIELSEN, IL, SUPERINTENDENT, BLOOMINGTON, ILLINOIS PUBLIC SCHOOLS

SCOTT NORTON, LA, DIRECTOR, STUDENT STANDARDS & ASSESSMENTS, LOUISIANA DEPARTMENT OF EDUCATION

JIM POPHAM, HI, EMERITUS PROFESSOR, UNIVERSITY OF CALIFORNIA

DORIS REDFIELD, WV, PRESIDENT/CEO, EDVANTIA

WENDY ROBERTS, DE, ASSESSMENT DIRECTOR, ELAWARE DEPARTMENT OF EDUCATION

LORRIE SHEPARD, CO, DEAN, SCHOOL OF EDUCATION, PROFESSOR OF EDUCATION, UNIVERSITY OF COLORADO AT BOULDER

RICK STIGGINS, OR, CEO, ASSESSMENT TRAINING INSTITUTE (AT), ETS

MARTHA THURLOW, MN, PROFESSOR, NATIONAL CENTER ON EDUCATION OUTCOMES, UNIVERSITY OF MINNESOTA

DYLAN WILIAM, UK, DEPUTY DIRECTOR, INSTITUTE OF EDUCATION, UNIVERSITY OF LONDON

[3] *The CCSSO Formative Assessment Advisory Group was formed in March 2006. This is a list of the original members who were responsible for developing and approving the definition and attributes of effective formative assessment.*

Are Good Grading Practices Like Putting Your Thumb in Your Navel?

Stuart Kahl, Ph.D.
Founding Principal

New tennis players often have a weak backhand swing, using a flicking, "thumb-up" motion similar to dealing cards. The most effective advice I've heard for that backhand problem is to hold your hand in the position you would if you were sticking your thumb in your navel–thumb pointing backwards. Following that simple tip takes care of several shortcomings of a poor tennis stroke and produces amazing results.

With so much information proliferating about assessment FOR learning, perhaps professional development providers and coaches of teachers and school administrators might benefit from a "thumb-in-the-navel" focus.

In recent years, we've learned of the tremendous positive impact of formative classroom assessment, or "assessment FOR learning," on student achievement. Yet we're finding that changing classroom practice isn't easy. Despite the overwhelming research evidence of the effectiveness of instructional practices associated with formative assessment, teachers are more than a little reluctant to change some habits that may well be counterproductive to student learning.

Recently, I ran across an article by Bill Schafer on assessment literacy for teachers*, in which he cited research identifying twelve common grading practices that "interfere with accurate assessment and student learning." The list includes entries about grading work completed during the learning process while feedback from the work is still being used for instruction, using unannounced quizzes, assigning zeros for missing or incomplete work, and many more.

It occurs to me that grading practices might be education's thumb–the one problem we can fix that will take care of many shortcomings. A school that successfully addresses the twelve items on Schafer's list will likely transform its teachers into effective practitioners of formative assessment. Granted, grading practices may not be easy to change, but tackling the twelve practices in phases is doable, and changing those practices is absolutely necessary if we want to bring about significant gains in achievement. We must focus on what's happening in the classroom to advance student learning, and we know a lot about what works.

Remember, "thumbs back" for tennis and "thumbs up" for changing grading practices in favor of proven formative assessment techniques.

* Schafer, W.D. (1993). Assessment literacy for teachers. *Theory into Practice, 32(2),118-126.* Copyright 1993, College of Education, The Ohio State University.

What do you think?
Let us know at twocents@measuredprogress.org

Formative Assessment: What Do Teachers Need to Know and Do?

To many of today's teachers, assessment is synonymous with high-stakes standardized tests. But there is an entirely different kind of assessment that can actually transform both teaching and learning. Ms. Heritage describes what the skillful use of formative assessment would look like.

By Margaret Heritage

FORMATIVE assessment, if used effectively, can provide teachers and their students with the information they need to move learning forward. But after more than a hundred years of exhortations and a significant body of research on the topic, the idea that assessment and teaching are reciprocal activities is still not firmly situated in the practice of educators. Instead, assessment is often viewed as something in competition with teaching, rather than as an integral part of teaching and learning.

In our current accountability environment, assessment is not regarded as a source of information that can be used during instruction. Instead, it has become a tool solely for summarizing what students have learned and for ranking students and schools. In the process, the reciprocal relationship between teaching and assessment has been lost from sight. In a context in which assessment is overwhelmingly identified with the competitive evaluation of schools, teachers, and students, it is scarcely surprising that classroom teachers identify assessment as something external to their everyday practice.

Educators recognize that annual state tests provide too little information that arrives too late for planning instruction, and this has prompted districts and schools to supplement state assessments with interim or benchmark assessments. These typically consist of item banks, administration tools, and customized reports, and they usually are administered uniformly to all students three to four times a year. Their greater frequency notwithstanding, these assessments still do not provide teachers with information they can use for ongoing instruction. Despite the enthusiasm for these assessments at the district level and the considerable resources that are being expended on them, the fact remains that they cover too long a period of instruction and provide too little detail for effective use in ongoing instructional planning. At best, they function more as snapshots of student progress and as predictors of student performance on the end-of-year, statewide tests. Indeed, Dylan Wiliam and Marnie Thompson suggest that they might better be described as "early warning summative" tools rather than as tools that can be formative to instruction.[1] Furthermore, teachers do not control how or when these tests occur, what the purpose of the assessment is, or who

■ *MARGARET HERITAGE is assistant director for professional development at the National Center for Research on Evaluation, Standards and Student Testing (CRESST) at the University of California, Los Angeles.*

is assessed. Yet these are the assessments that "count," even though they offer little help to teachers in their daily practice.

Compounding these difficulties is the fact that assessment has traditionally not been a focus of preservice and inservice courses. As Richard Stiggins laments, U.S. educators are "a national faculty unschooled in the principles of sound assessment."[2] Teachers learn how to teach without learning much about how to assess. Moreover, their administrators also lack training in assessment and therefore do not have the skills to support the development of assessment competencies.

What is missing in assessment practice in this country is the recognition that, to be valuable for instructional planning, assessment needs to be a moving picture — a video stream rather than a periodic snapshot. If assessment is used to inform effective instruction, then that assessment is quickly rendered out of date. Student learning will have progressed and will need to be assessed again so that instruction can be planned to extend the students' new growth.

Formative assessment practices, if implemented effectively, can provide teachers and their students with the data that they need. Moreover, there is empirical evidence that formative assessment, unlike benchmark assessments, is effective in improving student achievement.[3] However, in a profession that already feels burdened by the amount of assessment, there is a danger that teachers will see formative assessment as yet another external demand that takes time away from teaching.

WHAT IS FORMATIVE ASSESSMENT?

Formative assessment is a systematic process to continuously gather evidence about learning. The data are used to identify a student's current level of learning and to adapt lessons to help the student reach the desired learning goal. In formative assessment, students are active participants with their teachers, sharing learning goals and understanding how their learning is progressing, what next steps they need to take, and how to take them.

Formative assessment involves a variety of strategies for evidence gathering, which can be categorized into three broad types: on-the-fly assessment, planned-for interaction, and curriculum-embedded assessment.[4]

On-the-fly assessment. On-the-fly assessment occurs spontaneously during the course of a lesson. For example, a teacher listening to group discussions hears students expressing misconceptions about the science concept she has been teaching. She then changes the direction of her lesson to provide a quick "pop-up" lesson.[5] The pop-up lesson enables the teacher to clear up the misconceptions before proceeding with her planned instructional sequence.

Planned-for interaction. In planned-for interaction, teachers decide beforehand how they will elicit students' thinking during the course of instruction. For example, teachers plan the questions they will ask during the course of the lesson in order to enable students to explore ideas, and these questions can elicit valuable assessment information.[6]

Curriculum-embedded assessments. There are two kinds of curriculum-embedded assessments, those that teachers and curriculum developers embed in the ongoing curriculum to solicit feedback at key points in a learning sequence and those that are part of ongoing classroom activities. For example, student mathematical representations created during lessons can function as formative assessments, as can students' science notebooks that are also part of students' regular classroom activity.[7]

ELEMENTS OF FORMATIVE ASSESSMENT

There are four core elements of formative assessment: 1) identifying the "gap," 2) feedback, 3) student involvement, and 4) learning progressions. Teachers need to have a clear understanding of each of these elements.

Identifying the gap. In a seminal paper in 1989, Royce Sadler established the essential purpose of formative assessment as the means to identify the gap between a student's current status in learning and some desired educational goal. He stressed that this gap will vary from student to student and spelled out the consequence for pedagogy: "If the gap is perceived as too large by a student, the goal may be unattainable, resulting in a sense of failure and discouragement on the part of the student. Similarly, if the gap is perceived as too 'small,' closing it might not be worth any individual effort. Hence, to borrow from Goldilocks, formative assessment is a process that needs to identify the 'just right gap.'"[8]

Educational psychologists call this "just right gap" the zone of proximal development (ZPD). Originating with Lev Vygotsky's still-influential formulation, the ZPD is defined as the distance between what the child can accomplish during independent problem solving and the level of problem solving that can be accomplished under the guidance of an adult or in collaboration with a more expert peer.[9] The teacher's task is to identify and build on immature but maturing structures and, through collaboration and guidance, to fa-

cilitate cognitive growth. In the process, the child internalizes the resources required for solving a particular problem, and these resources become part of the child's independent developmental achievement. The term "scaffolding" characterizes the support that adults give to learners in the ZPD to move them from what they already know to what they can do next.[10] Effective formative assessments, then, should identify what a student might achieve in his or her ZPD and enable teachers to adapt teaching to close the gap between the student's current state of learning and the desired state.

Feedback. Formative assessment is designed to provide feedback at multiple levels. First, it provides feedback to the teacher about current levels of student understanding. This feedback also informs what the next steps in learning should be.

Feedback also is central to guiding students through their own next steps. Sadler's model strongly emphasizes feedback to students through the use of the feedback loop. This loop involves teachers and their students in an ongoing process. Effective feedback from teachers provides clear, descriptive, criterion-based information that indicates to the students where they are in a learning progression (defined below), how their understanding differs from the desired learning goal, and how they can move forward. The teacher takes steps to close the gap between the students' current learning and the goal by modifying instruction, assessing again to give further information about learning, modifying instruction again, and so on. In formative assessment, learners must be able to *use* feedback to improve their learning.[11] Another important aspect of the relationship between feedback and learning is that feedback has a strong effect on students' motivation and their sense of self-efficacy — how they feel about their various abilities — both of which are major influences on learning.

Student involvement. Improving learning through formative assessment also depends on the active involvement of students in their own assessment.[12] In formative assessment, students learn the skills of self- and peer assessment so that, as Sadler suggests, they collaborate with their teachers in developing a shared understanding of their current learning status and what they need to do to move forward in their learning. In doing so, they are using metacognitive processes. They reflect on their learning, monitoring what they know and understand and determining when they need more information. They also develop self-regulation strategies and are able to adapt their learning tactics to meet their own learning needs. Students must also collaborate with their teachers to determine the criteria for success for each step along the learning progression.

Learning progressions. If formative assessment is to provide guidance to teachers and students, it must be linked to a learning progression. The learning progression should clearly articulate the subgoals that constitute progress toward the ultimate goal. Most state standards, by themselves, do not provide a clear progression for understanding where students are relative to desired goals. In fact, many state standards do not even provide a clear picture of what learning is expected. Developing learning progressions toward standards is a critical element of formative assessment. Learning progressions provide the big picture of what is to be learned, and they help teachers locate students' current learning status on the continuum along which students are expected to progress.

Most state standards, by themselves, do not provide a clear progression for understanding where students are relative to desired goals.

Students also need to have short-term goals, which are derived from the learning progression and described in terms of success criteria. Success criteria are the guide to learning while the student is engaged in the learning tasks. The success criteria provide the framework within which formative assessment takes place and make possible the interpretation of evidence.[13]

THE KNOWLEDGE TEACHERS NEED

To use formative assessment successfully in the classroom, teachers need specific knowledge and skills. Four basic elements of teacher knowledge are critical: 1) domain knowledge, 2) pedagogical content knowledge, 3) knowledge of students' previous learning, and 4) knowledge of assessment.

Domain knowledge. Teachers must know the concepts, knowledge, and skills to be taught within a domain, the precursors necessary for students to acquire them, and what a successful performance in each looks like. With this knowledge, they are able to define a learning progression of subgoals toward the desired learning that will act as the framework to guide assessment and instruction. A sufficiently detailed progression will also supply the success criteria for recognizing when students have demonstrated a successful performance and when they have not and for providing students with substantive feedback.

Teachers also need to understand student metacognition as it relates to assessment. As noted earlier, students develop the ability to monitor and assess their own learning so that they recognize when they are learning and when they are not. Linked to self-assessment is self-regulation, the will to act in ways that result in learning. And when students recognize they are not learning, they have the strategies to do something about it. Finally, teachers need to know that students' motivational beliefs — for example, beliefs about their general level of competence or self-efficacy — may influence their learning.[14]

Pedagogical content knowledge. To effectively adapt instruction to student learning, teachers' pedagogical content knowledge must include familiarity with multiple models of teaching for student achievement in a specific domain and knowledge of which model of teaching is appropriate for what purpose. As already noted, the gap between current status and learning goals will differ from student to student, so teachers will need differentiated instructional strategies and a knowledge of how to use them in the classroom. To support student self-assessment, teachers will also need to be familiar with multiple models of teaching metacognitive processes and self-assessment skills.

Students' previous learning. If teachers are to build on students' previous learning, they need to know what that previous learning is. Students' previous learning includes: 1) their level of knowledge in a specific content area, 2) their understanding of concepts in the content area (i.e., the degree to which they can make generalizations through a process of abstraction from a number of discrete examples), 3) the level of their skills specific to the content area (i.e., the capacity or competence to perform a task), 4) the attitudes the students are developing (e.g., the value the students place on the subject, the interest they display, and their levels of initiative and self-reliance), and 5) their level of language proficiency.

Assessment knowledge. Teachers must know about the range of formative assessment strategies so that they can maximize the opportunities for gathering evidence. In addition, even though formative assessment strategies will not always meet accepted standards of validity and reliability, teachers need to understand that the quality of the assessment is an important concern. The overriding issue is consequential validity. Because the purpose of formative assessment is to promote further learning, its validity hinges on how effectively learning takes place in subsequent instruction. Teachers also need to know how to align formative assessments with instructional goals, and they need to ensure that the

evidence from the formative assessment and the inferences they draw from it are of sufficient quality to enable them to understand where the learner is along a learning progression.[15] Finally, teachers need to know that their own assessments of learning are not the only available sources of evidence; students' self- and peer assessments provide important opportunities for establishing their current learning status.

THE SKILLS TEACHERS NEED

In addition to an appropriate knowledge base, the successful implementation of formative assessment requires specific teacher skills. Teachers need to be able to 1) create classroom conditions that allow for successful assessment, 2) teach the students to assess their own learning and the learning of others, 3) interpret the evidence, and 4) match their instruction to the gap.

Creating the conditions. If students are going to be involved in assessment, two things need to happen. First, teachers must create a classroom culture that supports self- and peer assessment. This means that the classroom is a place where all students feel that they are respected and valued and that they have an important contribution to make. Second, teachers must have the skills to build a community of learners, characterized by a recognition and appreciation of individual differences. Classroom norms of listening respectfully to one another, responding positively and constructively, and appreciating the different skill levels among peers will enable all students to feel safe in the learning environment and to learn with and from one another. Above all, teachers will need the skills to model the "safety" norms of the classroom in their own behavior.

Student self-assessment. Teachers must teach students to assess their own learning and the learning of others. This involves helping students to set goals and criteria for success, to reflect on their own and others' understanding, and to evaluate learning according to the criteria. Strategies to involve students in self-assessment can be as simple as asking students to reflect on their performance through such questions as "Do you think that your response demonstrated understanding? If so, why do you think this? If not, why do you think you did not demonstrate understanding?" From this basis, students can learn to be more independent and can recognize when they do not understand, when they need to do something about it, and what they can do to improve.

Teacher skills also include helping students learn to give constructive feedback to their peers that can provide for future growth. From simple beginnings like saying, "It wasn't clear to me when. . ." or "I didn't understand your point about . . . ," students can progress to a detailed analysis of their peers' performance against specific criteria. Once again, the teacher must model all of this in the classroom so that students see that they are collaborators with their teacher and peers in developing a shared understanding of their current learning status and what they need to do to move forward.

Interpreting evidence. Teachers' skills in drawing inferences from students' responses are crucial to the effectiveness of formative assessment. No matter what the assessment strategy — observation, dialogue, asking for a demonstration or a written response — teachers must examine students' responses from the perspective of what they show about their conceptions, misconceptions, skills, and knowledge. This involves a careful analysis of the responses in relation to the criteria for success. In essence, teachers need to infer what the "just right gap" is between the current learning and desired goals, identifying students' emerging understanding or skills so that they can build on these by modifying instruction to facilitate growth.

The analysis of student responses takes place in different time frames, depending on the method of assessment. In on-the-fly assessments, teachers have to make inferences on a moment-by-moment basis. A curriculum-embedded analysis of student work might take place after the lesson and will provide more time for close examination. In both instances the importance of domain knowledge to analysis cannot be overstated; the success of the analysis is wholly dependent on it. Without a strong base of domain knowledge there is a danger that teachers' analyses will focus on the surface aspects of learning at the expense of deeper levels of understanding. An inaccurate analysis of the students' learning status will lead to errors in what the next instructional steps will be.

The analysis of student responses also provides the substance for feedback to students. Teachers need the skills to translate their analyses into clear and descriptive feedback, matched to the criteria for success, that can be used by students to further their learning.

Matching instruction to the gap. It is axiomatic to formative assessment that, if the next instructional steps to close the gap are too hard for the student, frustration will almost certainly result, and if they are too easy, boredom and disaffection are potential outcomes. Therefore, teachers need the skills to translate their interpretations of the assessment results into instructional actions that are matched to the learning needs of their students. This involves selecting the learning experiences that will place appropriate demands on the student and

ordering these experiences so that each successive element leads the student toward realizing the desired outcome. Having matched the next steps in learning to the gap, teachers' scaffolding skills come into play. Their skills in deciding on the appropriate strategy must be complemented by their skills in executing the strategy. Their job is to ensure that the student receives appropriate support so that new learning is incrementally internalized and ultimately becomes part of the student's independent achievement.

Matching the instruction to the gap cannot be done successfully without differentiating classroom instruction. In any classroom, one student's "just right gap" will not always be the same as another's. Clearly it is not practical for teachers to engage in one-on-one instruction with each student. However, strategic questioning in a whole-class lesson can provide scaffolding for a range of learning levels, while forming subgroups for instruction, assigning individual activities, and employing a combination of didactic and exploratory approaches help accommodate differences.

CONCLUSION

Even if teachers have all the required knowledge and skills for formative assessment, without the appropriate attitudes toward the role that formative assessment can play in teaching and learning, their knowledge and skills will lie dormant.

Teachers must view formative assessment as a worthwhile process that yields valuable and actionable information about students' learning. If they do not, formative assessment will be seen as "yet another thing" that is being externally imposed on them. Teachers must view formative assessment and the teaching process as inseparable and must recognize that one cannot happen without the other.

Also, if students are going to be successfully involved in monitoring and assessing their own and their peers' learning, then they need to be regarded by their teachers as partners in learning. This is not an attitude that has traditionally been prevalent in the profession.

If formative assessment is to be an integral part of professional practice, there needs to be a major investment made in teachers. This investment must begin with changes in preservice training. No teacher should exit a professional training program without the knowledge to assess student learning. Furthermore, beginning teachers must have opportunities to develop and practice the skills of assessing before they are responsible for a class of students. Teacher educators have a significant role to play in ensuring that teacher education programs equip their students with the knowledge and skills necessary to integrate teaching and assessment in classroom practice.

The investment in teachers must continue with inservice professional development that involves a commitment by leaders at all levels of the education system. Rather than providing teachers with more tests, leaders at the state, district, and school levels should invest in a coordinated effort to establish structures and provide resources that support effective professional development.

This investment is a long-term project that should not be shortchanged. The payoff will be improved teacher practices and improved student learning, and that is surely worth it.

---

1. Dylan Wiliam and Marnie Thompson, "Integrating Assessment with Learning: What Will It Take to Make It Work?," in Carol A. Dwyer, ed., *The Future of Assessment: Shaping, Teaching and Learning* (Mahwah, N.J.: Erlbaum, 2006).

2. Richard J. Stiggins, "Assessment Crisis: The Absence of Assessment FOR Learning," *Phi Delta Kappan*, June 2002, pp. 758-65.

3. Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy and Practice*, vol. 5, November 1998, pp. 7-73.

4. Richard J. Shavelson, "On the Integration of Formative Assessment in Teaching and Learning with Implications for Teacher Education," paper prepared for the Stanford Education Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group, 2006, available at www.stanford.edu/dept/SUSE/SEAL.

5. H. Margaret Heritage, Norma Silva, and Mary Pierce, "Academic Language: A View from the Classroom," in Alison L. Bailey, ed., *Language Demands of Students Learning English in School: Putting Academic Language to the Test* (New Haven, Conn.: Yale University Press, 2006).

6. Paul Black et al., *Assessment for Learning: Putting It into Practice* (New York: Open University Press, 2003).

7. H. Margaret Heritage and David Niemi, "Toward a Framework for Using Student Mathematical Representations as Formative Assessments," *Educational Assessment*, vol. 11, 2006, pp. 265-82; and Pamela Aschbacher and Alicia Alonzo, "Examining the Utility of Elementary Science Notebooks for Formative Assessment Purposes," *Educational Assessment*, vol. 11, 2006, pp. 179-203.

8. D. Royce Sadler, "Formative Assessment and the Design of Instructional Systems," *Instructional Science*, vol. 18, 1989, p. 130.

9. Lev Vygotsky, *Mind in Society* (Cambridge, Mass.: Harvard University Press, 1978); and idem, *Thought and Language* (Cambridge, Mass.: MIT Press, 1986).

10. David Wood, Jerome Bruner, and Gail Ross, "The Role of Tutoring in Problem Solving," *Journal of Child Psychology and Psychiatry*, vol. 17, 1976, pp. 89-100.

11. Assessment Reform Group, *Assessment for Learning: Beyond the Black Box* (Cambridge: University of Cambridge, School of Education, 1999).

12. Ibid.

13. Shirley Clarke, *Formative Assessment in the Secondary Classroom* (London: Hodder Murray, 2005).

14. Wynne Harlen, "The Role of Assessment in Developing Motivation for Learning," in John Gardner, ed., *Assessment and Learning* (London: SAGE, 2006), pp. 61-80.

15. Gordon Stobart, "The Validity of Formative Assessment," in Gardner, pp. 133-46.

Technology takes formative assessment to a whole new level

Student response system (SRS) technology has caught on in classrooms nationwide as a tool for boosting class participation, as well as helping teachers ensure that students understand what's being taught before they move on to another concept. But the current generation of the technology has its limitations.

For one thing, the lag time between student responses kills the pace of learning, says Promethean Director Tony Cann. In a typical use of the technology, the teacher poses a question to the entire class, then pauses as students answer the question on their personal "clicker" devices. This results in a lot of waiting around—time that could be put to better use.

Another problem is that students see, and answer, the same question as their peers. For students who already understand the material and are ready to move on, this can be a tedious process—and teachers risk losing their interest.

Promethean thinks it has developed a solution to these problems. The company has unveiled a brand-new version of software that could take SRS technology to a whole new level—something the company calls "real-time personalized intervention" (RTPI).

The new technology can send a question directly to each student's ActivExpression unit—Promethean's version of a "clicker" device. Once the student answers a question, he or she immediately gets a new question to answer. Best of all, the system is adaptive, meaning it can quickly hone in on each student's abilities and deliver personalized questions that target these abilities.

"If you believe interactive whiteboards are a level 4 improvement [in education], this is a level 8," Cann said, adding: "It will have an enormous effect on teaching and learning."

Cann said the technology gives teachers the ability to do handheld formative assessment in real time—assessment that adapts to the pace of each student, while also filling the time for each student. Students receive immediate feedback on their responses, and teachers can see how the entire class is progressing on their own computer screen.

In the teacher's view, each student response is color-coded according to the question's degree of difficulty. Wrong answers have a red "X" next to them, so teachers easily can see which students are progressing quickly through the questions and which need more help. The teacher then can work individually with students as they need it.

Ron Clark, founder of the Ron Clark School in Atlanta, called the development “shockingly effective, exciting, and brilliant.”

“As a teacher, I am able to send questions to each student’s ActivExpression, and they are able to work on the problems at their own pace. I can instantly track the progress of each student on the ActivBoard, and I can go directly to the students who aren’t on task or who are having problems,” he said.

“The best part about RTPI is that it places a sparkle on every student’s face. They love the thrill of texting their answers, and the competitive component of seeing how quickly you can answer all of the questions in each series is something that appeals to kids a great deal.”

After class, Clark said, “I always review the responses again to see where there are overall weaknesses in the class. I also take note of material that the students understand fully, and I realize that there is no need to review the content that has been mastered.”

Cann said the system is capable of delivering 15,000 to 20,000 self-paced questions per minute. It is available as a free software upgrade for current ActivExpression users, and the ActivExpression units themselves—which turn on in just two seconds and have a battery life that lasts for about 25,000 questions, he said—work with any whiteboard system, not just the ActivBoard.

Alignment of Assessments

The key to creating effective and efficient assessments

Entire books address effective item writing and test development. At the risk of over-simplifying this topic, ensuring alignment (using one or more from a series of criteria) lies at core of making assessments effective and efficient tools. Alignment covers many factors, which increase in number and complexity as the stakes associated with the assessments rise.

Traditional educational practice called for aligning curriculum, instruction, and assessment. The birth of the standards movement added another factor: state standards. At face value alignment among all four factors makes sense. The likelihood of student success increases when curricula align with state standards, when instruction aligns with curricula, and when assessments align with the other three. Contrary to what seems to be a too frequent misconception these days, tests shouldn't drive curriculum and instruction—standards should. They should also drive assessments. We wouldn't be surprised if students performed poorly on a classroom test covering material they were never taught. Similarly, if the math curriculum in a school weren't aligned with the state standards, we wouldn't be surprised if the school failed to make AYP.

Achieving acceptable alignment of assessments involves the following activities:

- Defining the purpose of the assessment and what information would be most useful
- Determining what type(s) of items or other activities will generate that information
- Ensuring a match between the assessment and the expectations (reflecting the standards, curricula, and instruction) in four key areas:
    - the content
    - the level of knowledge and skills
    - the breadth of knowledge and skills
    - the distribution of items
- As appropriate, ensuring that the assessment reflects applicable measurement principles

Well-established frameworks, processes, and tools exist to carry out these activities, with the complexity and intensiveness geared to the stakes involved. As educators gain experience in aligning their assessments, they will be able to do so with increasing ease and effectiveness. To begin, the process need not be overly complicated—just considering these factors will likely

improve item and assessment content. Over time each assessment will more effectively and efficiently provide the information needed to achieve the intended purpose.

The following examples illustrate common mistakes—opportunities where assessment literacy can help educators do more with less:

Vignette: An administrator believes that because items in a commercial item bank are aligned to her state standards, that a random selection of items from that bank will produce a test that is aligned to the state test.

Vignette: A state legislator, influenced by a group of school administrators, urges his colleagues to endorse the use of an already existing adaptive general achievement measure for NCLB-required testing.

Vignette: A state board chairperson is surprised to hear that a commercial NRT test form that he approved for a statewide testing program was already being used by many schools in the state.

Aligning assessments and standards

Eight reviewers sit at a long table in a conference room. Their notebook computers communicate with each other and with a database via a wireless network. Papers are scattered across the table.

The silence is deafening. The tension in the room is palpable. This project requires extended thinking, over a period of some hours, with multiple ways of solving problems. High-level reasoning is required; developing an argument is necessary.

Other resources

A similar alignment resource, WCER's Surveys of Enacted Curriculum project, encourages teacher reflection and conversation about classroom practice and instructional content. Teachers can compare their own practice and instructional content to responses by other teachers around the country and within their school or district. Participating states, schools and districts use aggregated teacher reports to develop a baseline of information about teacher practice in mathematics, science and English language arts, or to inform professional development or school improvement planning efforts.

But these people are not taking a test. They are evaluating a test. The evaluators are reading sample items from a state's learning standards, then assigning Depth of Knowledge levels to each content objective, each assessment item, and to each objective targeted by each assessment item.

"I gave this item a Level 3, but I was torn," one says. "There's potential for higher order thinking. But it could be a 2, because often we don't prompt students to that level. So this item is halfway between a 2 and a 3."

The No Child Left Behind (NCLB) Act requires states to align their performance standards and assessments. But correspondence between state-level standards and assessments tends to be only moderate, particularly in terms of 'depth of knowledge' and 'range of knowledge.'

Based on years of working with standards and assessments, WCER Senior Scientist Norman Webb has designed a system for measuring the degree of alignment. Too often, Webb says, education systems are fragmented, so teachers and students get mixed messages about goals and expectations. In the absence of clear principles of alignment, learning expectations can be lowered for some students while being raised for others. That creates potential inequities.

Webb emphasizes that alignment is the degree to which expectations and assessments work together to improve and measure students learning. As such, alignment is a quality of the relationship between expectations and assessments and not a specific attribute of either of these system components. “These parts of the education system must work together to help students achieve at higher levels of understanding,” Webb says.

Whether it’s in language arts, mathematics, social studies, or science, expectations and assessments should agree on the underlying concepts and what it means to “know” these concepts, Webb says. Aligned expectations and assessments describe and represent how students link concepts and how their instructional experiences should be organized.

The degree of alignment of expectations and assessments can be determined using four criteria:

• **Categorical concurrence** measures the extent to which the same or consistent categories of content appear in the standards and the assessments. The criterion is met for a given standard if there are more than five assessment items targeting that standard. Six items are assumed as a minimum for an assessment measuring content knowledge related to a learning goal and as a basis for making some decisions about students' knowledge of that learning goal.

• **Range-of-knowledge correspondence** determines whether the span of knowledge expected of students on the basis of a standard corresponds to the span of knowledge that students need in order to correctly answer the corresponding assessment items/activities. The criterion is met for a given standard if more than half of the objectives that fall under that standard are targeted by assessment items.

• **Balance of representation** measures whether objectives that fall under a specific standard are given relatively equal emphasis on the assessment. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a learning goal are equally distributed among the objectives for the given learning goal. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit.

• **Depth-of-knowledge consistency** measures the degree to which the knowledge elicited from students on the assessment is as complex within the context area as what students are expected to know and do as stated in the standards. The criterion is met if more than half of targeted objectives are hit by items of the appropriate complexity. For example, assume an assessment included six items related to one learning goal and that students are required to answer correctly four of those items to be judged proficient – i.e., 67% of the items. If three items, 50% of the six items were at or above the depth of knowledge level of the corresponding Standards, then for a

student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth of knowledge level of one learning goal.

Depth of knowledge levels

The descriptions for each of four Levels for mathematics help to clarify what the different levels represent in each subject area.

Level 1 (recall and reproduction) is the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple science process of procedure. A student answering a Level 1 item either knows the answer or does not.

Level 2 (skills and concepts) includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in Level 1. Keywords that generally distinguish a Level 2 item include 'classify,' 'organize,' 'estimate,' 'make observations,' 'collect and display data,' and 'compare data.'

Level 3 (strategic thinking) requires reasoning, planning, using evidence, and higher level of thinking than the previous two levels. The complexity results because the multistep task requires more demanding reasoning.

Level 4 (extended thinking) Tasks at this level have high cognitive demands and are very complex. Students are required to make several connections, to related ideas within the content area or among content areas—and have to select or devise one approach among many solution alternatives. This level requires complex reasoning, experimental design and planning, and probably will require an extended period of time, either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item.

The online alignment tool

A recently developed web-based tool by Brian Vesperman moves the alignment process away from paper and pencil and brings it into the electronic age, allowing states to more quickly evaluate that alignment. Some 25 states have used WCER's online Web Alignment Tool (WAT) http://www.wcer.wisc.edu/wat/index.aspx to guide and automate the process. District and state education staff, for example, use results from alignment analysis to refine standards and to identify more appropriate assessment items. For each alignment criterion, an acceptable level is defined by what would be required to assure that a student meets the standard(s).

The WAT is free and available to all. It requires some training and effort to learn how to use it, but anyone can register as a group leader and use the tool to conduct a study.

Maine leads once again with Common Core pilot

With Common Core State Standards (CCSS) now on 46 state agendas, including the District of Columbia (D.C.) it's time to start thinking implementation. But leaders are saying it takes more than vendor press releases and simple classroom curriculum supplements—it takes research and a focus on teaching—and one state is leading the way in 21st century learning…again.

Maine, already known for its Maine Learning Technology Initiative (MLTI) and successful 1-to-1 computing program, is now in a unique collaboration with the University of Southern Maine's research center and a Portland-based educational software company, AcademicMerit.

The collaboration will explore the use of technology-based solutions to help schools meet the demand of the CCSS through a pilot involving more than 23 schools, more than 30 teachers, and almost 1,500 students from districts all over Maine.

"Last year [2010] was frenetic, with states and districts making promises and submitting proposals left and right," said Ogden Morse, chief executive of AcademicMerit and an English teacher on leave from Falmouth High School in Maine. "This year [2011] is what I refer to as 'the pregnant pause,' with many of those same folks asking, 'So how are we actually going to do what we promised?' As a result, right now, there are a lot of people in search of 'what works.' This pilot is an opportunity for the rest of the country to point in our direction and say 'Well, look at what's going on in Maine…'"

And though ultimately the goal of CCSS implementation is to improve student learning outcomes, one of the pilot's main professional development goals is to help teachers assess student work.

During the pilot, which began in February 2011 and ended in early June the same year, classes use AcademicMerit's products, called Literary Companion (LC) and Assessments21 (A21)—a suite of online tools targeting English/Language Arts in grades 7-12.

LC aims to help students deepen their understanding of literature and non-fiction, vocabulary, and reading and writing skills by offering interactive, technology-based supplements to the curriculum. Classes use LC for an entire literature unit, and the program also can be used outside of class for homework.

Assessments21 generates classroom-based formative and summative assessment data that help teachers inform instruction. It is this assessment process that truly makes the program unique, company reps said.

“In the CCSS era, we teachers essentially have to know the anchor standards for both reading and writing (and their grade-specific benchmarks)—and then make sure we are improving student performance on them,” Morse said. “Well, the only way to know whether students are improving is to be able to measure their performance on an ongoing basis against those anchor standards. That’s a challenge.”

Double-blind method

The revolutionary part of the pilot comes from the double-blind scoring and A21 assessments.

Both students and teachers are asked to create an account through AcademicMerit, with a login name and password. When the first essay is assigned, students submit the essay through their accounts.

The student essay is then graded by scorers who are part of AcademicMerit and who do not know specific student information. A first scorer grades the essay based on the CCSS specifications of thinking, content, organization, style, and mechanics. Each category has a six-point scale, and a four or better is a passing grade.

Then the essay goes to a second scorer, who doesn’t know student information or the first score given to the essay. The essay is then delivered to the teacher. Each score for each specification includes a detailed explanation as to why the score was given.

After this first round of assessments, teachers are required to go through the online score-calibration system, or AcademicMerit’s beta PD program. Here, the teacher is asked to grade an essay provided by AcademicMerit based on the accompanying rubric. The teacher’s scores must generally align with other authorized scorers’ assessments of the essay.

Once the teacher’s scoring of an essay matches those of AcademicMerit’s authorized scorers, the teacher then becomes an authorized scorer for AcademicMerit.

Each teacher is required to score the number of student essays equal to twice the number of students he or she has participating in the study.

According to Peter Vose, English chair at Falmouth High School and pilot participant, seeing the scores of the other readers helps to maintain fair and objective standards by which to judge and grade student work.

“I found that I did not agree with all of the scores, and my eMail to the company asking for clarification was answered quickly and persuasively. I asked my student to reflect on their work and their scores, and nearly all agreed with the scorers’ assessments,” said Vose.

He continued: “The other assessments we have used, such as SAT’s NWEA’s and Accuplacer, simply provide a number with little of the detail that AcademicMerit scores provide.”

“The consortia working on the common assessments due to launch in 2014 have stated clearly that they see classroom-based interim assessments as a key element of the larger system,” said Jeff Mao, learning technology policy director at the Maine Department of Education. “AcademicMerit’s tools have the potential to be a model of just that for 7-12 ELA. What better place than Maine to evaluate that potential?”

Research and beyond

Throughout the pilot, student and teacher progress will be monitored and teachers will be asked to complete comprehensive surveys and possibly participate in focus groups.

The aim of the research will be to understand how technology can best improve learning in conjunction with the CCSS, and, also, to provide a working example of how states and schools can collaborate with private companies to embrace the CCSS.

The results of the study are expected in late June or early July 2011.

“Since the data in our programs generate…practical value, I foresee opportunities for us to collaborate on a more formal basis with states, like Maine, seeking to test or implement the sort of system we represent,” said Morse. “Already, we are in discussions with large districts like Cobb County, Ga., about implementation plans that could be a model for others.”

“Simply put, this is the next step,” said former Maine governor Angus King, whose administration championed the state’s laptop initiative. “Computers in the classroom alone were never going to transform learning and instruction. These programs will, because they use that technology to deliver outstanding academic content that is better than anything I’ve seen to date.”

Mao said this type of pilot is inspiring because AcademicMerit began with a teacher–Morse–who saw an opportunity.

“Part of MLTI’s mission is to spur economic development for the state, and if other student-centered, comprehensive resources develop, and the data is good, we’ll of course, try to leverage these types of resources,” he said.

Already, vendors around the country are promoting their resources for the implementation of the CCSS: iParadigms’ Turnitin, dataMetrics’ TestWiz, Ascend Education’s Ascend Math, and Key Curriculum Press’ Key Math Strategists.

What exactly do "fewer, clearer, and higher standards" really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and implement assessments

Karin K. Hess, Dennis Carlock, Ben Jones, and John R. Walkup

Abstract

With the ever-increasing call for more rigorous curriculum, instruction, and assessment in the United States, the National Governors Association and the Council of Chief State School Officers are aiming to define the rigorous skills and knowledge that students need in order to succeed academically in college-entry courses and in workforce training programs (Glod, 2009). The proposed Common Core Standards will require high-level cognitive demand, such as asking students to demonstrate deep conceptual understanding through the application of content knowledge and skills to new situations. Using two widely accepted measures of describing cognitive rigor — Bloom's Taxonomy of Educational Objectives and Webb's Depth-of-Knowledge Levels — this article defines cognitive rigor and presents a matrix that integrates these models as a strategy for analyzing instruction and influencing teacher lesson planning. Using Hess' Cognitive Rigor Matrix (CRM), a density plot illustrates how the preponderance of curricular items (e.g., assignment questions and problem solving tasks) might align to cells in the matrix. Research results applying the matrix in two states' large-scale collection of student work samples are presented, along with a discussion of implications for curriculum planning in order to cultivate twenty-first century skills.

Beginning with Bloom

In 1956, a group of educational psychologists headed by Benjamin Bloom developed a classification of levels of intellectual behavior important in learning. Bloom created this taxonomy for categorizing the levels of abstraction of questions that commonly occur in educational settings. Using these levels for analysis, Bloom found that over 95% of test questions students encounter at the college level required them to think only at the lowest possible level: the recall of information. Bloom's committee identified three domains of educational activities: cognitive (*knowledge*), affective (*attitude*), and psychomotor (*skills*). Within the cognitive domain, which is tied directly to mental skills, Bloom identified a hierarchy of six levels that increased in complexity and abstraction – from the simple recall of facts, *knowledge,* to the highest order of thinking, *evaluation*. In practice, educators assigned Bloom's Taxonomy levels according to the main action verb associated with a level in the taxonomy. For example, examining the meaning of a metaphor and categorizing geometric shapes would both align to Bloom's Taxonomy, *Analysis* level. While educators have found such verb cues of Bloom's Taxonomy levels to be useful in guiding teacher questioning, verbs often appear at more than one level in the taxonomy (e.g., *appraise, compare, explain, select, write*); and often the verb alone is inadequate for determining the actual cognitive demand required to understand the content addressed in a test question or learning activity. (See Table 1.)

Building upon Bloom's early work, many educational and cognitive psychologists have since developed various schemas to describe the cognitive demand for different learning and assessment contexts. In 2001, Anderson, Krathwohl, et al. presented a structure for rethinking Bloom's Taxonomy. Whereas the original taxonomy applied one dimension, the revised taxonomy table employs two dimensions—cognitive processes *and* knowledge.

The cognitive processes dimensions resemble those found in the original taxonomy, but placement on the taxonomy continuum has changed slightly (e.g., evaluation no longer resides at the highest level) and descriptions have been expanded and better differentiated for analyzing educational objectives. The revised descriptors consider both the processes (the verbs) and the knowledge (the nouns) used to articulate educational objectives. This restructuring of the original taxonomy recognizes the importance of the interaction between the content taught — characterized by factual, conceptual, procedural, and metacognitive knowledge — and the thought processes used to demonstrate learning.

Table 1: A Comparison of Descriptors: Bloom's Original Taxonomy and the Revised Bloom's Taxonomy of Cognitive Process Dimensions

Bloom's Taxonomy (1956)	The Revised Bloom Process Dimensions (2001)
Knowledge Define, duplicate, label, list, memorize, name, order, recognize, relate, recall, reproduce, state	**Remember** Retrieve knowledge from long-term memory, recognize, recall, locate, identify
Comprehension Classify, describe, discuss, explain, express, identify, indicate, locate, recognize, report, restate, review, select, translate	**Understand** Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models (e.g., cause-effect)
Application Apply, choose, demonstrate, dramatize, employ, illustrate, interpret, practice, schedule, sketch, solve, use, write	**Apply** Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task
Analysis Analyze, appraise, calculate, categorize, compare, criticize, discriminate, distinguish, examine, experiment, explain	**Analyze** Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)
Synthesis Rearrange, assemble, collect, compose, create, design, develop, formulate, manage, organize, plan, propose, set up, write	**Evaluate** Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique
Evaluation Appraise, argue, assess, choose, compare, defend, estimate, explain, judge, predict, rate, core, select, support, value, evaluate	**Create** Put elements together to form a coherent whole, reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce for a specific purpose

Webb's Depth-of-Knowledge (DOK) Levels

Depth of knowledge forms another important perspective of cognitive complexity. Probably the best-known work in the area of depth of knowledge is that of Norman Webb (1997, 1999). Webb's work has forced states to rethink the meaning of test alignment to include both the content assessed in a test item and the intended cognitive demand, or the depth to which we expect students to demonstrate understanding of that content. In other words, the complexity of both the content (e.g., simple vs. complex data displays; interpreting literal vs. figurative language) and the task required (e.g., solving routine vs. non-routine problems) are used to determine DOK levels. Webb describes his depth-of-knowledge levels as "nominative" rather than as a taxonomy, meaning that DOK levels name (or describe) four different and deeper ways a student might interact with content (2002).

Webb's Depth-of-Knowledge Levels

DOK-1 – **Recall & Reproduction** - Recall of a fact, term, principle, concept, or perform a routine procedure

DOK-2 - **Basic Application of Skills/Concepts** - Use of information, conceptual knowledge, select appropriate procedures for a task, two or more steps with decision points along the way, routine problems, organize/display data, interpret/use simple graphs

DOK-3 - **Strategic Thinking & Reasoning** - Requires reasoning, developing a plan or sequence of steps to approach problem; requires some decision making and justification; abstract, complex, or non-routine; often more than one possible answer

DOK-4 - **Extended Thinking** - An investigation or application to real world; requires time to research, problem solve, and process multiple conditions of the problem or task; non-routine manipulations, across disciplines/content areas/multiple sources

Identifying the DOK levels of questions in tests or class assignments can help to articulate how deeply students must understand the related content to complete the necessary tasks. Unlike Bloom's Taxonomy, Webb's model dictates that depth-of-knowledge levels do not necessarily correlate to the commonly understood notion of "difficulty." That is, an activity that aligns to a particular level is not always "easier" than an activity that aligns to a DOK level above it. For example, a DOK-1 activity might ask students to restate a simple fact or a much more abstract theory, the latter being much more difficult to memorize and restate. Neither of these DOK-1 tasks asks for much depth of understanding of the content. On the other hand, greater depth is required to explain how or why a concept or rule works (DOK-2), to apply it to real-world phenomena with justification or supporting evidence (DOK-3), or to integrate a given concept with other concepts or other perspectives (DOK-4).

Interpreting and assigning intended DOK levels to both the standards and the related assessment items are now essential requirements in any alignment analyses. Webb's depth-of-knowledge levels have been applied across all content areas (Hess, 2004, 2005a, 2005b, 2006a; Petit & Hess, 2006) and many states and districts utilize the concept of depth of knowledge to designate the depth and complexity of state standards in order to align the state's large-scale assessments or to revise existing standards to achieve higher cognitive levels for instruction. Consequently, teachers need to develop the ability to design instruction, and create units of study/curriculum and classroom assessments for a greater range of cognitive demand.

Cognitive Rigor and the CR Matrix

Although related through their natural ties to the complexity of thought, Bloom's Taxonomy and Webb's depth-of-knowledge differ in scope and application. Bloom's Taxonomy categorizes the cognitive skills required of the brain to perform a task, describing the "type of thinking processes" necessary to answer a question. Depth of knowledge, on the other hand, relates more closely to the depth of content understanding and scope of a learning activity, which manifests in the skills required to complete the task from inception to finale (e.g., planning, researching, drawing conclusions). Both the thinking processes and the depth of content knowledge have direct implications in curricular design, lesson delivery, and assessment development and use.

While there is no simple one-to-one correspondence between these complexity schemas to articulate cognitive rigor, the superposition of Bloom's Taxonomy and Webb's Depth-of-Knowledge Levels was originally expressed in matrix form by Hess (2006b) for use in states where the conversation about cognitive complexity as part of the test design and item development process was just beginning. The

CR matrix has been helpful in explaining to teachers how the two conceptual models—Bloom's Taxonomy and Webb's DOK levels —are alike, yet different (Table 2). More importantly, the CR matrix allows educators to examine the depth of understanding required for different tasks that might seem at first glance to be at comparable levels of complexity. Finally, the CR matrix allows educators to uniquely categorize and examine selected assignments/ learning activities that appear prominently in curriculum and instruction. For example, the rote completion of single-step mathematical routines, often derided by the moniker "plug and chug," lies positioned within the (DOK-1, Bloom-3/Apply) or the (1, 3) cell of the CR matrix. Using the CR matrix to plot typical mathematics assignments from a unit of study, a teacher might discover to what extent this level of cognitive rigor is being assessed compared to (DOK-3, Bloom-3) or the (3, 3) cell of the CR matrix, using strategic thinking/reasoning (DOK-3) and application (Bloom-3). **When used to plot multiple assignments over time, the CR matrix can graphically display a unique view of instructional emphasis and ultimately reveal the focus of learning within a classroom, a grade level, or a school system.**

Table 2: Hess' Cognitive Rigor Matrix with Curricular Examples: Applying Webb's Depth-of-Knowledge Levels to Bloom's Cognitive Process Dimensions

Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 **Recall & Reproduction**	**Level 2** **Skills & Concepts**	**Level 3** **Strategic Thinking/ Reasoning**	**Level 4** **Extended Thinking**
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	Recall, recognize, or locate basic facts, ideas, principles Recall or identify conversions: between representations, numbers, or units of measure Identify facts/details in texts			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models	Compose & decompose numbers Evaluate an expression Locate points (grid/, number line) Represent math relationships in words pictures, or symbols Write simple sentences Select appropriate word for intended meaning Describe/explain how or why	Specify and explain relationships Give non-examples/examples Make and record observations Take notes; organize ideas/data Summarize results, concepts, ideas Make basic inferences or logical predictions from data or texts Identify main ideas or accurate generalizations	Explain, generalize, or connect ideas using supporting evidence Explain reasoning when more than one response/approach is possible Explain phenomena in terms of concepts Compose full composition to meet specific purpose and audience Identify theme(s) using text evidence	Explain how concepts or ideas specifically relate to other content domains or concepts Develop generalizations of the results obtained or strategies used and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	Follow simple/routine procedure (recipe-type directions) Solve a one-step problem Calculate, measure, apply a rule Apply an algorithm or formula (area, perimeter, etc.) Represent in words or diagrams a concept or relationship Apply rules or use resources to edit spelling, grammar, punctuation, conventions	Select a procedure according to task needed and perform it Solve routine problem applying multiple concepts or decision points Retrieve information from a table, graph, or figure and use it solve a problem requiring multiple steps Use models to represent concepts Write paragraph using appropriate organization, text structure, and signal words	Use concepts to solve non-routine problems Design investigation for a specific purpose or research question Conduct a designed investigation Apply concepts to solve non-routine problems Use reasoning, planning, and evidence Revise final draft for meaning or progression of ideas	Select or devise an approach among many alternatives to solve a novel problem Conduct a complex project that specifies a problem, identifies solution paths, solves the problem, and reports results Illustrate how multiple themes (historical, geographic, social) may be interrelated
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	Retrieve information from a table or graph to answer a question Identify or locate specific information contained in maps, charts, tables, graphs, or diagrams	Categorize, classify materials Compare/ contrast figures or data Select appropriate display data Organize or interpret (simple) data Extend a pattern Identify use of literary devices Identify text structure of paragraph Distinguish: relevant-irrelevant information; fact/opinion	Compare information within or across data sets or texts Analyze and draw conclusions from more complex data Generalize a pattern Organize/interpret data: complex graph Analyze author's craft, viewpoint, or potential bias	Analyze multiple sources of evidence or multiple works by the same author, or across genres, or time periods Analyze complex/abstract themes Gather, organize, and analyze information from multiple sources Analyze discourse styles across texts
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			Cite evidence and develop a logical argument for concepts Describe, compare, and contrast solution methods Verify reasonableness of results Justify conclusions made	Gather, analyze, & evaluate relevancy & accuracy Draw & justify conclusions Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	Brainstorm ideas, concepts, or perspectives related to a topic or concept	Generate conjectures or hypotheses based on observations or prior knowledge	Synthesize information within one source, data set, or text Formulate an original problem, given a situation or data set Develop a complex conceptual model for a given situation	Synthesize information across multiple sources or texts Design a model to inform and solve a real-world, complex, or abstract situation

Research Results Applying the Hess CR Matrix

In two recent large-scale studies of the enacted (or taught) mathematics and English language arts curricula, teachers from 200 Nevada and Oklahoma public schools submitted over 200,000 samples of student work, which encompassed homework samples, tests, quizzes, and worksheets, completed during the period from February – May, 2008 (The Standards Company LLC, 2008a, 2008b). Using a process developed by The Standards Company LLC, curriculum specialists analyzed: (a) each item on each work sample for the Bloom's Taxonomy levels needed to formulate an adequate response to the prompt; and (b) the overall depth-of-knowledge of each content-specific assignment. The analysts then assigned to each collected work sample the highest Bloom's Taxonomy level appearing on the assignment in addition to its overall depth-of-knowledge level.

The CR density plots in Figure 1 illustrate some sample results for the two studies, relying on the same two-dimensional layout as the CR matrix shown in Table 2, but incorporating the percentage of assignments as a color shade. In this sense, the reader can liken student work to semi-transparent sheets stacked vertically on each cell of the appropriate CRM cell. As the stack increases in height it becomes more opaque and, therefore, darkens. (The studies encompass hundreds of such plots, disaggregated according to grade level, subject area/course, socioeconomic status, etc. The two plots shown here are cumulative examples for each subject area.)

Results for grade 3 English language arts indicate a preponderance of assignments correlating to the (DOK-2, Bloom-2/Understand) cell of Cognitive Rigor. Mathematics assignments, on the other hand, sampled the (DOK-1, Bloom-3/Apply) cell to a greater extent than other types/depths of thinking (as shown in Fig. 1). Examples of assignments correlating to the lowest level of depth of knowledge and the “apply-level” of Bloom's Taxonomy included one-step solutions of algebraic equations, non-rote multiplication, and long division. Although mathematics assignments associated with the (1,3) cell are necessary for students to practice their numeracy and fluency skills in mathematics, the result nonetheless may point to an over-reliance on instructional activities corresponding to straightforward applications of learned or routine steps. This emphasis would not prepare students for non-routine applications or transfer of the same mathematics skills.

Readers should note that the collection period for the entirety of both studies spanned roughly three months, but this only included five consecutive days at each school site. For this reason, the analysis of assignments is likely to have failed to capture instances of DOK-4 activities, which typically require multiple class sessions over extended time to complete.

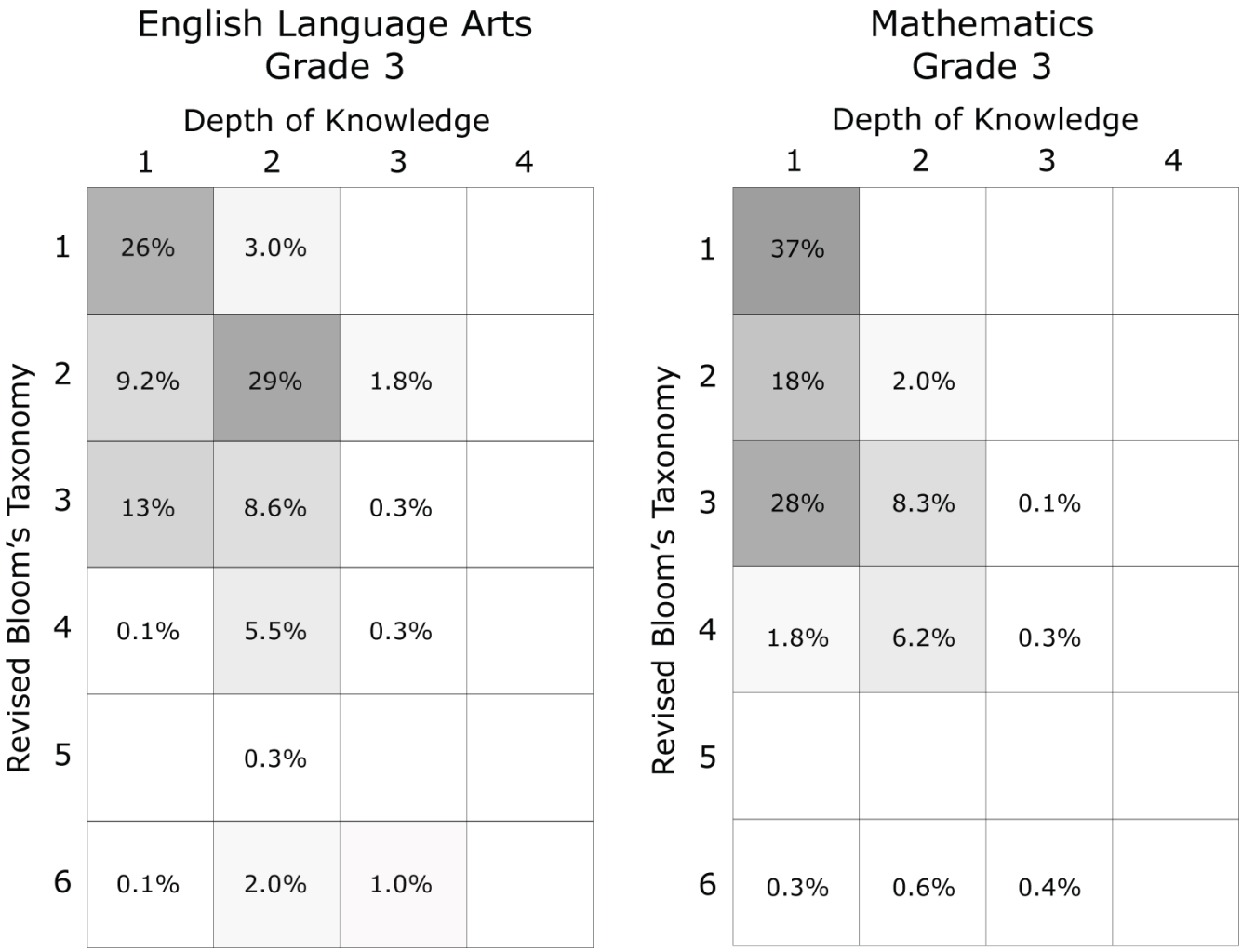


Figure 1: Density plots comparing the cognitive rigor of the English language arts and mathematics enacted curriculum (The Standards Company LLC 2008a, The Standards Company LLC 2008b). To generate the results shown, student work was collected from 205 schools across two states. Although all public school grade levels were analyzed, only grade 3 is shown here. The English language arts data corresponds to 12,060 samples of student work; 8,428 for mathematics.

Discussion

One conclusion the researchers have drawn from this work is that both measures of cognitive complexity can serve useful purposes in education reform at the state level (standards development and large-scale assessment alignment) and at the school and classroom levels (lesson design and teaching and assessment strategies). Because cognitive rigor encompasses the complexity of content, the cognitive engagement with that content, and the scope of the planned learning activities, the CR matrix has significant potential to enhance instructional and assessment practices at the classroom level. Superimposing the two cognitive complexity measures produces a means of analyzing the emphasis placed on each intersection of the matrix in terms of curricular materials, instructional focus, and classroom assessment.

Ensuring that curriculum is aligned to "rigorous" state content standards is, in itself, insufficient for preparing students for the challenges of the twenty-first century. Current research on the factors

influencing student outcomes and contributing to academic richness supports the concept that learning is optimized when students are involved in activities that require complex thinking and the application of knowledge. Expert teachers provide *all* students with challenging tasks and demanding goals, structure learning so that students can reach high goals, and know how to enhance both surface and deep learning of content (Hattie, 2002). Students learn skills and acquire knowledge more readily when they understand concepts more deeply, recognize their relevance, and can transfer that learning to new or more complex situations. Transfer is more likely to occur when learners have developed deep understanding of content and when initial learning focuses on underlying principles and cause-effect relationships (NRC, 2001).

As educators become more skilled at recognizing the elements and dimensions of cognitive rigor and analyzing its implications for instruction and assessment, they can provide learning opportunities that benefit all students, across all subject areas and grade levels. In essence, the role of a school system is to prepare students by providing them with an aligned curriculum with differentiated emphasis on each of the four depth of knowledge levels. The cognitive rigor matrix might serve as a constant reminder to educators that students need exposure to novel and complex activities every day.

Authors

Karin K. Hess is a Senior Associate, at the National Center for the Improvement of Educational Assessment, Dover, NH [khess@nciea.org]

Dennis Carlock, Ben Jones, and John R. Walkup are curriculum research specialists at The Standards Company LLC, Clovis, CA.

References Cited

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (Eds) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Addison Wesley Longman, Inc.

Bloom B. S. (Ed.) Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain.* New York: David McKay.

Glod, M. (Monday, June 1, 2009). "46 states, D.C. plan to draft common education standards." *Washington Post* [retrieved June 15, 2009] http://www.washingtonpost.com/wp-dyn/content/article/2009/05/31/AR2009053102339.html?referrer=emailarticle

Hattie, J. (October 2002). "What are the attributes of excellent teachers?" Presentation at the New Zealand Council for Educational Research Annual Conference, University of Auckland.

Hess, K. (2004). "Applying Webb's Depth-of-Knowledge (DOK) Levels in reading." [online] available: http://www.nciea.org/publications/DOKreading_KH08.pdf

Hess, K. (2005a). "Applying Webb's Depth-of-Knowledge (DOK) Levels in social studies." [online] available: http://www.nciea.org/publications/DOKsocialstudies_KH08.pdf

Hess, K. (2005b). "Applying Webb's Depth-of-Knowledge (DOK) Levels in writing." [online] available: http://www.nciea.org/publications/DOKwriting_KH08.pdf

Hess, K. (2006a). "Applying Webb's Depth-of-Knowledge (DOK) Levels in science." [online] available: http://www.nciea.org/publications/DOKscience_KH08.pdf

Hess, K. (2006b). "Exploring cognitive demand in instruction and assessment." [online] available: http://www.nciea.org/publications/DOK_ApplyingWebb_KH08.pdf

National Research Council. (2001). Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.) *Knowing what student know: The science and design of educational assessment*. Washington, D.C.: Academy Press.

Petit, M. & Hess, K. (2006). "Applying Webb's Depth-of-Knowledge (DOK) and NAEP levels of complexity in mathematics." [online] available: http://www.nciea.org/publications/DOKmath_KH08.pdf

The Standards Company LLC. (2008a). "Study of the alignment of student assignments to the academic standards in the state of Nevada pursuant to Senate Bill 184, Chap. 420, Statutes of Nevada 2007." Retrieved April 13, 2009, from Legislative Counsel Bureau, Nevada State Legislature, technical report, http://www.leg.state.nv.us/lcb/fiscal/Final_Report-Curriculum_Study.pdf .

The Standards Company LLC. (2008b). "Analysis of the enacted curriculum for the Oklahoma State Department of Education for the collection period February – March, 2008." Oklahoma State Department of Education, unpublished technical report.

Webb, N. (March 28, 2002) "Depth-of-Knowledge Levels for four content areas," unpublished paper.

Webb, N. (August 1999). Research Monograph No. 18: "Alignment of science and mathematics standards and assessments in four states." Washington, D.C.: CCSSO.

Webb, N. (1997). Research Monograph Number 6: "Criteria for alignment of expectations and assessments on mathematics and science education. Washington, D.C.: CCSSO.

Hess' Cognitive Rigor Matrix & Curricular Examples: Applying Webb's Depth-of-Knowledge Levels to Bloom's Cognitive Process Dimensions - *ELA*

Revised Bloom's Taxonomy	Webb's DOK Level 1 Recall & Reproduction	Webb's DOK Level 2 Skills & Concepts	Webb's DOK Level 3 Strategic Thinking/ Reasoning	Webb's DOK Level 4 Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	o Recall, recognize, or locate basic facts, details, events, or ideas explicit in texts o Read words orally in connected text with fluency & accuracy			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion), predict, compare/contrast, match like ideas, explain, construct models	o Identify or describe literary elements (characters, setting, sequence, etc.) o Select appropriate words when intended meaning/definition is clearly evident o Describe/explain who, what, where, when, or how o Define/describe facts, details, terms, principles o Write simple sentences	o Specify, explain, show relationships; explain why, cause-effect o Give non-examples/examples o Summarize results, concepts, ideas o Make basic inferences or logical predictions from data or texts o Identify main ideas or accurate generalizations of texts o Locate information to support explicit-implicit central ideas	o Explain, generalize, or connect ideas using supporting evidence (quote, example, text reference) o Identify/ make inferences about explicit or implicit themes o Describe how word choice, point of view, or bias may affect the readers' interpretation of a text o Write multi-paragraph composition for specific purpose, focus, voice, tone, & audience	o Explain how concepts or ideas specifically relate to *other* content domains or concepts o Develop generalizations of the results obtained or strategies used and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	o Use language structure (pre/suffix) or word relationships (synonym/antonym) to determine meaning of words o Apply rules or resources to edit spelling, grammar, punctuation, conventions, word use o Apply basic formats for documenting sources	o Use context to identify the meaning of words/phrases o Obtain and interpret information using text features o Develop a text that may be limited to one paragraph o Apply simple organizational structures (paragraph, sentence types) in writing	o Apply a concept in a new context o Revise final draft for meaning or progression of ideas o Apply internal consistency of text organization and structure to composing a full composition o Apply word choice, point of view, style to impact readers' /viewers' interpretation of a text	o Illustrate how multiple themes (historical, geographic, social) may be interrelated o Select or devise an approach among many alternatives to research a novel problem
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	o Identify whether specific information is contained in graphic representations (e.g., map, chart, table, graph, T-chart, diagram) or text features (e.g., headings, subheadings, captions) o Decide which text structure is appropriate to audience and purpose	o Categorize/compare literary elements, terms, facts/details, events o Identify use of literary devices o Analyze format, organization, & internal text structure (signal words, transitions, semantic cues) of different texts o Distinguish: relevant-irrelevant information; fact/opinion o Identify characteristic text features; distinguish between texts, genres	o Analyze information within data sets or texts o Analyze interrelationships among concepts, issues, problems o Analyze or interpret author's craft (literary devices, viewpoint, or potential bias) to create or critique a text o Use reasoning, planning, and evidence to support inferences	o Analyze multiple sources of evidence, or multiple works by the same author, or across genres, time periods, themes o Analyze complex/abstract themes, perspectives, concepts o Gather, analyze, and organize multiple information sources o Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			o Cite evidence and develop a logical argument for conjectures o Describe, compare, and contrast solution methods o Verify reasonableness of results o Justify or critique conclusions drawn	o Evaluate relevancy, accuracy, & completeness of information from multiple sources o Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, produce	Brainstorm ideas, concepts, problems, or perspectives related to a topic or concept	o Generate conjectures or hypotheses based on observations or prior knowledge and experience	o Synthesize information within one source or text o Develop a complex model for a given situation o Develop an alternative solution	o Synthesize information across multiple sources or texts o Articulate a new voice, alternate theme, new knowledge or perspective

Hess' Cognitive Rigor Matrix & Curricular Examples: Applying Webb's Depth-of-Knowledge Levels to Bloom's Cognitive Process Dimensions – *M-Sci*

Revised Bloom's Taxonomy	Webb's DOK Level 1 Recall & Reproduction	Webb's DOK Level 2 Skills & Concepts	Webb's DOK Level 3 Strategic Thinking/ Reasoning	Webb's DOK Level 4 Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	o Recall, observe, & recognize facts, principles, properties o Recall/ identify conversions among representations or numbers (e.g., customary and metric measures)			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models	o Evaluate an expression o Locate points on a grid or number on number line o Solve a one-step problem o Represent math relationships in words, pictures, or symbols o Read, write, compare decimals in scientific notation	o Specify and explain relationships (e.g., non-examples/examples; cause-effect) o Make and record observations o Explain steps followed o Summarize results or concepts o Make basic inferences or logical predictions from data/observations o Use models /diagrams to represent or explain mathematical concepts o Make and explain estimates	o Use concepts to solve non-routine problems o Explain, generalize, or connect ideas using supporting evidence o Make and justify conjectures o Explain thinking when more than one response is possible o Explain phenomena in terms of concepts	o Relate mathematical or scientific concepts to other content areas, other domains, or other concepts o Develop generalizations of the results obtained and the strategies used (from investigation or readings) and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	o Follow simple procedures (recipe-type directions) o Calculate, measure, apply a rule (e.g., rounding) o Apply algorithm or formula (e.g., area, perimeter) o Solve linear equations o Make conversions among representations or numbers, or within and between customary and metric measures	o Select a procedure according to criteria and perform it o Solve routine problem applying multiple concepts or decision points o Retrieve information from a table, graph, or figure and use it solve a problem requiring multiple steps o Translate between tables, graphs, words, and symbolic notations (e.g., graph data from a table) o Construct models given criteria	o Design investigation for a specific purpose or research question o Conduct a designed investigation o Use concepts to solve non-routine problems o Use & show reasoning, planning, and evidence o Translate between problem & symbolic notation when not a direct translation	o Select or devise approach among many alternatives to solve a problem o Conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct	o Retrieve information from a table or graph to answer a question o Identify whether specific information is contained in graphic representations (e.g., table, graph, T-chart, diagram) o Identify a pattern/trend	o Categorize, classify materials, data, figures based on characteristics o Organize or order data o Compare/ contrast figures or data o Select appropriate graph and organize & display data o Interpret data from a simple graph o Extend a pattern	o Compare information within or across data sets or texts o Analyze and draw conclusions from data, citing evidence o Generalize a pattern o Interpret data from complex graph o Analyze similarities/differences between procedures or solutions	o Analyze multiple sources of evidence o analyze complex/abstract themes o Gather, analyze, and evaluate information
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			o Cite evidence and develop a logical argument for concepts or solutions o Describe, compare, and contrast solution methods o Verify reasonableness of results	o Gather, analyze, & evaluate information to draw conclusions o Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	o Brainstorm ideas, concepts, or perspectives related to a topic	o Generate conjectures or hypotheses based on observations or prior knowledge and experience	o Synthesize information within one data set, source, or text o Formulate an original problem given a situation o Develop a scientific/mathematical model for a complex situation	o Synthesize information across multiple sources or texts o Design a mathematical model to inform and solve a practical or abstract situation

Performance Assessment

Providing richer, more direct evidence of what students know and can do

At a very broad level, performance assessment requires students to demonstrate their knowledge and skills through some form of "product" or presentation/demonstration rather than merely selecting a response from two or more options. Examples could range from filling in a blank to carrying out a major research project and formally presenting the results. Typically, educators define performance measures as opportunities for students to show how they can apply their knowledge and skills in disciplinary and interdisciplinary *tasks* focused on key aspects of academic learning. When people speak of performance assessment today in the context of 21st century skills, they often refer to more substantial activities—either short-term, on-demand tasks or curriculum-embedded, project-based tasks that yield reliable and valid scores. The most common example of such performance assessment is writing "prompts" that require students to produce essays or other forms of extended writing. Other scorable products or performances could include responses to constructed-response questions following some activity, research reports, oral presentations, and debates.

Performance assessment garnered considerable attention in the late 1980's as concerns rose about the dominant—and in some cases exclusive—role multiple-choice (selected-response) tests were playing at the state and district levels. Several states began introducing a range of non-multiple-choice items—from constructed-response items to real-world tasks and portfolios—to their large-scale testing programs. However, the increased testing and short turn-around times mandated by NCLB forced some states to cut back on these types of measures. Over the past several years, as the nation worried about our international competitiveness and the rigorous demands of college and career, performance measures have once again started to attract attention.

Performance measures provide richer, more direct evidence of what students know and can do, thereby enabling teachers to more accurately pinpoint and address learning challenges. In this regard, they are more effective than other, seemingly quicker and cheaper measures. Done well, they also demand deeper and higher-order thinking by students. Performance measures can take time to plan, administer, and score, which has caused some educators to shy away from them. However, you can start simply by having students show their work, and solutions exist to efficiently address the activities in more complex tasks. Performance assessment has also drawn some criticism as a *subjective*, rather than an *objective*, measure like a multiple-choice test in which answers are either right or wrong. Here too, through the considerable experience gained

over the past 20+ years, the results of performance measures can be reliable and valid—some experts would argue even more so than *objective* measures.

The following examples illustrate common mistakes—opportunities where assessment literacy can help educators do more with less.

Vignette: A state legislator believes only high-level content experts can be qualified to score constructed responses of students.

Vignette: A political activist argues that any tests requiring human scoring are being scored for personal values and opinions.

Vignette: When budgets are tight, legislators and other policy makers are quick to argue that since results of the more expensive measures of higher order skills are correlated with those of multiple-choice tests, there's no reason to implement or retain the non-multiple-choice testing.

Performance Assessment—
An Idea Whose Time Has Come (Again)

Stuart Kahl, Ph.D.
Founding Principal

No matter where you stand on common-core standards, Race to the Top, 21st Century skills, or ESEA reauthorization, one clear benefit of all the focus on reform is the education community's reinvigorated interest in the promise of performance assessment as a component of a balanced assessment system.

As envisioned here, performance assessment is a testing approach that relies on extended activities that yield scorable products or performances. Current discussion supports the practice of using interim performance approaches, along with more traditional summative assessments, to address accountability requirements. In addition, it's clear that "readiness" skills such as problem solving, critical thinking, and communication, are difficult to evaluate with many existing state tests and better measured by performance assessment.

In the pre-NCLB days, many states undertook pioneering efforts in large-scale performance assessment. Much was done right, but there was plenty of room for improvement, too. The lessons learned during that period were invaluable. We now pay great attention to the alignment of tasks to standards, and we have approaches to scoring, standard setting, and other psychometric tasks that are better suited to performance assessment. In short, we know how to do it.

I recommend that states, districts, and schools embrace locally managed and scored, curriculum-embedded performance assessment. However, it's important to take steps to ensure that the tasks and measurement are of high quality:

- Students should be engaged in projects that are closely tied to content standards and, for accountability purposes, that undergo review and pilot procedures similar to those used for the development of traditional tests.
- Project tasks should yield multiple, individual, scorable products demonstrating each student's knowledge and skills—products such as written reports, oral presentations, or other demonstrations we know how to score reliably.
- For high-stakes programs, local scoring could be audited by central "second" scoring on a sampling basis.

If curriculum-embedded, these assessments would have immediate classroom uses, both summative and formative. In short, performance assessments not only can measure student learning, they can be a rich part of the learning itself.

I have every hope that a much broader acceptance of performance assessment will be an enduring legacy of all the programs, initiatives, and reforms that currently dominate conversations in the education environment.

What do you think?

Let us know at twocents@measuredprogress.org

measured progress

The Measured Progress Difference
It's all about student learning. Period.

Briefing Paper Prepared for Members of
The Congress of
The United States

Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

George H. Wood
Director, The Forum for Education and Democracy
Principal, Federal Hocking High School, Stewart, Ohio

Linda Darling-Hammond
Charles E. Ducommun Professor, Stanford University
Co-Director, School Redesign Network

Monty Neill
Co-Director, Fair Test (National Center for Fair & Open Testing)

Pat Roschewski
Director of Statewide Assessment
Nebraska Department of Education

May 16, 2007

For More Information Contact
George Wood, Forum for Education and Democracy
740-448-4941
www.forumforeducation.org

Executive Summary

Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

By George Wood, Linda Darling-Hammond, Monty Neill and Pat Roschewski

Performance based assessments, often locally controlled and involving multiple measures of achievement, offer a way to move beyond the limits and negative effects of standardized examinations currently in use for school accountability. While federal legislation calls for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding" (NCLB, Sec. 1111, b, 2, I, vi), most assessment tools used for federal reporting focus on lower-level skill that can be measured on standardized mostly multiple-choice tests. High stakes attached to them have led schools to not engage in more challenging and engaging curriculum but to limit school experiences to those that focus on test preparation.

Performance assessments that are locally managed and involve multiple sources of evidence assist students in learning and teachers in teaching for higher order skills. These tools engage students in the demonstration of skills and knowledge through the performance of tasks that provide teachers with an understanding of student achievement and learning needs. Large scale examples involving the use of such performance-based assessments come from states such as Nebraska, Wyoming, Connecticut and New York, as well as nations such as Australia and Singapore. The evidence from research on these and other systems indicate that through using performance assessments schools can focus instruction on higher order skills, provide a more accurate measure of what students know and can do, engage students more deeply in learning, and provide for more timely feedback to teachers, parents, and students in order to monitor and alter instruction.

Research evidence suggests that in order for performance assessment systems to work, governments must make significant investments in both teacher development and the development of performance tasks. However, this investment is often no greater than the cost of standardized measures. More important, it strengthens teacher quality and student learning. Performance assessment systems can be reliable and valid, having both content and predictive validity when appropriately utilized.

Based on the evidence that performance based assessment better meets the federal agenda of teaching for higher-level skills, reauthorization of NCLB should support and encourage state and local education agencies in developing performance assessments. Congress can amend Section 1111 (b)(3) of NCLB with a new paragraph (D) that authorizes and encourages states to move to performance based assessments and multiple measures incorporated into a system combining state and local assessments. Authorization for adequate funding to support this move should be included in the legislation.

Refocusing Accountability:
Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

Over the past decade, educators, policymakers, and the public have begun to forge a consensus that our public schools must focus on better preparing all children for the demands of citizenship in the 21st century. This has resulted in states developing 'standards-based' educational systems and assessing the success of districts and schools in meeting these standards measured through more systematic testing. However, most of these tests are multiple choice, standardized measures of achievement, which have had a number of unintended consequences, including: narrowing of the academic curriculum and experiences of students (especially in schools serving our most school-dependent children); a focus on recognizing right answers to lower-level questions rather than on developing higher-order thinking, reasoning, and performance skills; and growing dissatisfaction among parents and educators with the school experience. The sharp differences between the forms of testing used in the United States and the assessments used in other higher-achieving countries also suggest that low international rankings may be related to over-reliance on standardized testing in the U.S.

These unfortunate consequences have occurred despite language in NCLB calling for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding" (NCLB, Sec. 1111, b, I, vi). Changing what counts as assessment evidence, coupled with other significant changes in NCLB's accountability structure (e.g., adequate yearly progress and sanctions), could help to overcome these problems and contribute toward school improvement

Performance Assessment: A Definition

Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers into an automobile with a DMV official for a spin around the block and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel parking technique. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV. We understand the value of this performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the test tell us some important things about potential drivers' skills, we also know that preparing for the test helps improve those skills as potential drivers practice to get better. The test sets a standard toward which everyone must work. Without it, we'd have little assurance about what people can actually *do* with what they know about cars and road rules, and little leverage to improve actual driving abilities.

Performance assessments in education are very similar. They are tools that allow teachers to gather information about what students can actually do with what they are learning – science experiments that students design, carry out, analyze, and write up; computer programs that students create and test out; research inquiries that they pursue, seeking and assembling evidence about a question, and presenting in written and oral

form. Whether the skill or standard being measured is writing, speaking, scientific or mathematical literacy, or knowledge of history and social science research, students actually perform tasks involving these skills and the teacher observes, gathers information about, and scores the performance based upon a set of pre-determined criteria. As in our driving test example, these assessments typically consist of three parts; a task, a scoring guide or rubric, and a set of administration guidelines. The development, administration, and scoring of these tasks requires teacher development to insure quality and consistency. The research suggests that such assessments are better tools for showing the extent to which students have developed higher order thinking skills, such as the abilities to analyze, synthesize, and evaluate information. They lead to more student engagement in learning and stronger performance on the kinds of authentic tasks that better resemble what they will need to do in the world outside of school. They also provide richer feedback to teachers, leading to improved learning outcomes for students.

Extensive research and experience, both here and abroad, have demonstrated that the use of *performance assessments* which are *locally administered* and use *multiple* sources of evidence offer the opportunity to turn assessment systems to serve their primary purpose—**assisting students in learning and teachers in teaching for higher order intellectual skills**. In fact, the assessment systems of most of the highest-achieving nations in the world are a combination of centralized assessments that use mostly open-ended and essay questions and local assessments given by teachers which are factored into the final examination scores. These local assessments--which include research papers, applied science experiments, presentations of various kinds, and projects and products that students construct--are mapped to the syllabus and the standards for the subject and are selected because they represent critical skills, topics, and concepts. Central authorities often determine curricular areas and skills to assess, but the assessments are generally designed, administered, and scored locally.

The *local management* of such assessments refers to both their use and scoring. While not all performance assessments are locally developed many are; and decisions about when to use them in the learning process and how to adapt them to particular content are made at the school or classroom level. This is vital as assessment must be responsive to emerging student needs and enable fast and specific teacher response, something that standardized examinations with long lapses between administration and results cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about the standards and how to teach to them and about what their students' learning needs are. The process improves their teaching. These rich assessment tasks can also be utilized as formative or benchmark assessments, which help teachers' gauge ongoing progress, while avoiding the reduction of such assessments to commercially available multiple-choice formats.

Using *multiple sources of evidence* refers to the way in which performance assessments provide multiple ways to view student learning. For example, multiple samples of actual writing taken over time can best reveal to a teacher the progress a student is making in the development of composition skills. This provides ongoing feedback to learners as well, as they see how they are developing as writers and what

they have yet to master. In addition, different kinds of writing tasks – persuasive essays, research papers, journalistic reports, responses to literature – encourage students to develop the full range of their writing and thinking skills in ways that writing a five-paragraph essay over and over again do not.

These features of performance, local administration, and multiple sources of evidence are used in many assessment systems. Let's think back to the state driver's license exam. This involves both a written test and a performance assessment on the road. Everyone knows precisely what to expect in terms of the skills to be demonstrated —for example, whether or not the applicant can parallel park—as the examination is not a total secret. The fact that the assessment is open and transparent is not a problem, because the point is to see whether drivers have developed these real-world abilities. The performance is scored by the instructor, working from a rubric, and if the driver is sufficiently successful in all aspects of the examination (as determined by a state cut-off score), a license is conferred. The task is so well defined that instructional programs (driver's education) which include both hands on and classroom instruction clearly demonstrate their effectiveness in preparing students to perform. (This is reflected in the reduced insurance rates we grant to graduates of driver's education programs.) Imagine what life on our roads would be like if we did not require prospective drivers to demonstrate what they know before taking the wheel.

Some states, districts, and schools have constructed a similarly rich set of assessments of competence that measure the higher-order thinking called for by new standards. In many cases they are explicitly intended to augment and complement more traditional tests.

Illinois' assessments provide a good example of the contrast between classroom performance assessment and a state multiple-choice test. The state's grade 8 science learning standard 11B reads: "Technological design: Assess given test results on a prototype; analyze data and rebuild and retest prototype as necessary." The multiple choice example on the state test simply asks what "Josh" should do if his first prototype sinks, with the wanted answer "Change the design and retest his boat." The classroom assessment, however says: "Given some clay, a drinking straw, and paper, design a sailboat that will sail across a small body of water. Students can test and retest their designs." In the course of this activity, students can explore significant physics questions such as displacement in order to understand why what was a ball of clay can be made to float. Such activities combine hands-on inquiry with reasoning skills, have visible real-world applications, are more engaging, and enable deeper learning. They also enable the teacher to assess student learning along multiple dimensions, including the ability to frame a problem, develop hypotheses, reflect on outcomes and make reasoned and effective changes, demonstrate scientific understanding, use scientific terminology and facts, persist in problems solving, and organize information, as well as develop sound concepts regarding the scientific principles in use.

Many states – including Connecticut, New York, and Vermont -- have developed and use such hands-on assessments as part of their state testing systems. Indeed, the

National Science Foundation provided millions of dollars for states to develop such hands-on science and math assessments as part of its Systemic Science Initiative in the 1990s, and prototypes exist all over the country.

Perhaps the most important benefit to utilizing performance assessments is that they assist in learning and teaching. They are *formative* in that they provide teachers and students with the feedback they need from authentic tasks to see if they have actually mastered content. They can also be *summative* in that they can serve as a final assessment of student capabilities with respect to state and local standards. Because of their numerous positive features, they are more sensitive to instruction and more useful for teaching than standardized examinations, while providing richer evidence of student learning that can be used by those outside the classroom or school.

Performance Assessment: Large Scale Examples

As we have noted, it is possible to create and implement assessment systems that include multiple sources of evidence which are performance based and locally managed. Some U.S. states and many countries have developed extensive performance-based assessment systems that engage teachers, parents, and students in thinking carefully about what students have learned and how to measure that learning. Examples include:

- Nebraska utilizes a system of assessments created and scored by local educators. These systems are peer-reviewed in a system supported by assessment experts and include a check on the validity of such assessments through the use of a state-wide writing examination and the administration of one norm-referenced test.
- Wyoming uses a "body of evidence" approach that is locally developed in order to determine whether students have mastered standards required for graduation.
- Connecticut uses rich science tasks as part of its statewide assessment system. For example, students design and conduct science experiments on specific topics, analyze the data, and report their results to prove their ability to engage in science reasoning. They also critique experiments and evaluate the soundness of findings.
- Maine, Vermont, New Hampshire, and Rhode Island have all developed systems that combine a jointly constructed reference exam with locally developed assessments that provide evidence of student work from performance tasks and portfolios.
- In New York, the New York Performance Assessment Consortium is a network of 47 schools in the state that rely upon performance assessments to determine graduation. (Because of the quality of their work, they have a state waiver from some of the Regents Examinations). Research from their work indicates that New York City students who graduate from these schools (which have a much higher graduation rate than the City although they serve more low-income students, students of color, and recent immigrants) are more successful in college than students with a traditional Regents diploma which relies upon standardized tests.
- In Silicon Valley, CA, many school districts use the Mathematics Assessment Resource System (MARS), an internationally developed program which requires students to learn complex knowledge and skills to do well on a set of performance-based tasks. The evidence is that students do as well on traditional

tests as peers who are not in the MARS program, while MARS students do far better at solving complex problems.

- Australia, New Zealand, Hong Kong, Singapore, England, and Canada operate systems of assessment that include local performance-based assessments that count toward the total examination score (typically at least 50%). In Queensland, Australia the state's "New Basics" and "Rich Tasks" approach to standards and assessment, which began as a pilot in 2003, offers extended, multi-disciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally-oriented curriculum. They are, says Queensland, "specific activities that students undertake that have real-world value and use, and through which students are able to display their grasp and use of important ideas and skills." Extensively researched, this system has had excellent success as a tool for school improvement. Studies found stronger student engagement in learning in schools using the Rich Tasks. Similar to MARS, on traditional tests, New Basics students scored about the same as students in the traditional program, but they performed notably better on assessments designed to gauge higher order thinking. The Singapore government has employed the developers of the Queensland system to focus their school improvement strategies upon performance assessments. High-scoring Hong Kong has also begun a process of expanding its already-ambitious school-based assessment system.

Clearly there is extensive experience available for designing and implementing assessment systems that include performance assessments, require multiple sources of evidence, and include local assessments. There is also an extensive research literature on performance assessments. The examples above are all examples of performance assessment *systems*; that is, assessment systems that use primarily or exclusively performance tasks, offering a strong existence proof for the viability of such systems.

Perhaps the most complex question surrounding these assessments when they are locally developed or scored is how to ensure comparability. Many of the systems described earlier, both in the U.S. and abroad, use common scoring guides. Queensland's system, like those in a number of countries, also employs "moderation," a process of bringing samples from different schools to be rescored, with results sent back to the originating schools. This process leads to stronger comparability across schools and is part of building a strong performance assessment system. The Learning Record, at one time used in dozens of U.S. schools, established very high inter-rater agreement (reliability) using moderation because the instrument is high quality and the training is effective.

Nebraska, through its peer review process, verifies that scorers within each district participate in extensive scorer training on common rubrics. Although districts may be using different tools, consistency and comparability within classrooms, buildings, and districts is supported in this way. Valid comparison across districts is achieved through external validation checks such as the statewide writing assessment, the ACT and other

commonly administered standardized tests. Each district's assessment system is evaluated and approved through a review process conducted by measurement experts.

Performance Assessment: Evidence

The research and work that has been done on performance assessment has uncovered a number of benefits, challenges, and criteria for making such assessment systems successful. Among the benefits of performance assessment systems are that they:

- Elevate the focus of instruction to higher order thinking skills;
- Provide a more accurate and comprehensive assessment of what students know and can do;
- Lead to more student engagement in both the learning and assessment process;
- Invite more teacher buy-in and encourage collaborative work;
- Support improvement of teaching practices;
- Provide clearer information to parents as to student development, accomplishments, and needs; and
- Allow instruction to be altered in a timely fashion to meet student learning needs.

From the research and evidence on performance assessment, there are a number of lessons learned that should be considered when designing a system that substantially incorporates performance-based assessments:

- Although some methods of managing performance assessments can cost more then machine scoring of multiple choice tests (i.e. when such assessments are treated as traditional external tests and shipped out to separately paid scorers), the cost calculus changes when assessment is understood as part of teachers' work and learning – built into teaching and professional development time. Much evidence suggests that developing and scoring these assessments is a high-yield investment in teacher learning and a good use of professional development resources. In addition, performance assessment systems are not necessarily more costly than accountability systems that rely upon standardized measures of achievement. For example, Nebraska, which utilizes a locally managed assessment system, spends only $.03 per child (or $9,000) on outside assessment contracts while Ohio, relying upon standardized measures, spends $50.00 per child (or $92,000,000). In most European and Asian systems, and in those used in several U.S. states, scoring of assessments is conducted by teachers and time is set aside for this aspect of teachers' work and learning. While teacher time to create and score the assessments can be substantial, these activities lead to more skilled and engaged teachers. In contrast, external standardized tests provide teachers with little guidance on how to improve student learning when they simply receive numerical scores on secret tests months after the students have left school. Hence the professional development that seeks to help teachers improve achievement in this system is under-informed and ineffective.

- Extensive professional development is necessary for educators to learn to build, use, and score assessments that will inform and guide their teaching. Few teachers now have that knowledge, but they can and will develop it when given the opportunity, as has been demonstrated in many systems. The system must engage the adult learners in curriculum alignment, performance task development, scoring processes, and data analysis so that they 'own' the system and do not feel bypassed. This includes developing a peer review, audit, or moderation system that provides for a feedback loop, checks on quality, and includes directions for staff development.

- Productive use of performance assessments, like proper use of standardized tests, should be aimed at revealing areas needing improvement and should lead to curriculum and professional learning supports rather than punishments. Only if schools or districts show themselves unwilling to take advantage of support should sanctions be undertaken.

- Personnel in departments of education and legislatures at the state and federal levels must understand that only classroom teachers can directly impact instruction and learning. Therefore, their task is to provide assistance to teachers to make the system work.

- Careful attention must be paid to the performance tasks. They should be developed in response to criteria that establishes the technical quality of assessments (including checking for bias and fairness), high proficiency standards, consistent administration of assessment, and opportunity to learn what is assessed. They should also be constructed to allow students with special needs and those who are learning English opportunities to demonstrate their knowledge appropriately.

<u>Performance Assessment: Federal Legislative Initiatives</u>

In the reauthorization of NCLB, consideration should be given to how federal legislation could support these more sophisticated forms of assessment that support students in developing higher order thinking and reasoning skills. Congress should provide support for states to design accountability systems that use multiple performance measures of student achievement that include locally administered performance assessments. To that end, we would suggest that legislative language capturing the following items be located in the reauthorization of NCLB.

1. Allow for and encourage the use of locally administered performance assessments as part of a balanced system for reporting on school and student achievement, in keeping with the existing requirement in Section 1111 (b) (3) (vi) that multiple measures be used to assess higher-order thinking and understanding.

2. Provide funding to states and localities to develop such systems that meet criteria which include:
    i. Assurance of the technical quality of assessments used for state reporting so that the evidence of learning derived from the classroom, school or district performance assessments is accurate, valid and reliable for the purposes for which it will be used;
    ii. Assurance that the assessments are valid measures of state standards as well as local curricula;
    iii. Assurance that assessment measures are free from bias;
    iv. Demonstration of validation and verification processes, such as peer review, assessor training, and moderation or auditing.
3. Appropriation of funds for any state that chooses to undertake the development of school based performance assessments, in an amount no less than $10 million per state and scaled to the size of the state, to support professional development activities for teachers and school leaders associated with developing, implementing, and scoring such assessments and integrating their results in plans for improving instruction. Such funds could also be used for states to work in collaboration in the design and validation of performance-based assessment systems, the development of performance tasks or other materials, and the design of professional development.

A fuller detailing of these proposals is available.

Appendices:

1. Criteria for locally-based performance assessments to use in comprehensive state assessment systems.
2. Validation and Verification of Locally-based Performance Assessments
3. Performance Assessment: A Short Bibliography

APPENDIX 1: Criteria for locally-based performance assessments to use in comprehensive state assessment systems

State proposals for funding in a grant or pilot project should ensure that the assessments they propose to develop and use meet the following criteria:

- are performance-based [see definition, below];
- assess higher order thinking skills [as required in current law – see definition below];
- provide multiple sources of evidence of student learning [see definition below];
- are locally-based [see definition below] – (This may include the use of tasks or assessments that are locally developed or locally-selected or adapted from a bank of tasks and used when appropriate for evaluating student learning);
- are fair and unbiased;
- are based on local curriculum as well as state standards;
- are able to be integrated with curriculum and instruction in schools and classrooms;
- provide timely, diagnostically-useful information;
- employ principles of universal design, while allowing adaptation to specific needs of students, particularly English language learners and students with disabilities;
- meet technical requirements of validity and reliability for the uses to which they are put;
- can be used to demonstrate progress toward proficiency; and
- are accompanied by or integrated with extensive professional development (and, professional development supported by the Act may be used to develop, use, and score locally-based performance assessments, provided the funds are not simply used for scoring large-scale assessments)

Performance-based assessment refers to assessments that evaluate applications of knowledge to real-world tasks. Such assessments may include, for example, students' oral or written responses to questions or prompts, as well as products such as essays or research papers, mathematical problems or models; science demonstrations or experiments; or exhibitions in the arts. They may be specific tasks they may be compilations of a number of such tasks within or across subject areas.

Higher order thinking and performance skills refer to the abilities to frame and solve problems; find, evaluate, analyze, and synthesize information; apply knowledge to new problems or situations; develop and test complex ideas; and communicate ideas or solutions proficiently in oral or written form.

Multiple sources of evidence (sometimes termed "multiple measures") involve different sources and kinds of evidence of student learning in a subject or across subject areas. Multiple measures allow multiple opportunities to demonstrate achievement, are accessible to students at varying levels of proficiency, and utilize different methods for demonstrating achievement.

Locally-based assessments may include both common assessments, which are assessments developed for use at the school or district level, and classroom-based evidence obtained from curriculum-embedded schoolwork by students.

Appendix 2: Validation and Verification of Locally-based Performance Assessments

Local performance assessments, including classroom assessments, are commonly used in the instructional process in order to provide feedback to students and to improve instruction. When such assessments are used for accountability purposes they need to be *validated* as appropriately measuring the knowledge and skills they intend to measure and *verified* as being evaluated in non-biased, consistent ways.

There are several widely-used means that schools, districts, states, and other nations have developed to validate and verify the scoring of state and local performance assessments. These include:

- expert and peer review,
- concurrent validation studies and "benchmark checks,"
- assessor training and calibration
- external auditing, and
- moderation strategies.

We describe these methods briefly here and provide an example of how several of these strategies (peer review, benchmark checks, and assessor training) are used in the Nebraska assessment system, which relies on local assessment systems to complement the state's large-scale assessments.

Validation and Verification Processes

Around the world, performance tasks, projects, and collections of student work – including the Advanced Placement and International Baccalaureate examinations – are used as part of both formative assessment systems and formal examination systems that carry accountability purposes. To ensure that the assessments themselves are valid measures of the intended learning standards and appropriately evaluate what students know and can do for the intended purposes of the assessment, they are typically subjected to several kinds of ***expert review*** – both of the tasks themselves and of the scoring tools and processes used to evaluate them. This review is typically conducted by experts in the content fields being assessed and by measurement experts and may draw on pilot studies and other research evidence about student performance on the assessments.
These reviews are a means of establishing content and construct validity for the assessments.

Another form of validation is to examine outcomes on assessments in relation to those on other measures. This is done through studies of concurrent validity, which are also sometimes known as "***benchmark checks***." If there are large discrepancies between the aggregate performances of students on different measures that are not explained by

differences in the skills and content they are measuring, this is a flag for further examination of how the assessments are being designed or scored.

Research shows that performance tasks can be scored with high levels of reliability if they are well-designed and guidance for scoring is clear and well-structured. This usually involves a rubric showing the scoring dimensions and descriptions of each performance level, along with instructions for how to evaluate the tasks. Consistency is greatly strengthened when the scoring guides are clear and of high quality. Collections of student work (work samples, portfolios) can be reliably scored when students and teachers have clear guidance on the features of the work to be submitted that facilitate consistent scoring.

Training assessors also helps ensure that tasks are scored consistently and in an unbiased fashion. Assessor training typically involves learning the scoring process from an expert and reviewing benchmarks, which are assessment samples that represent responses at each score level (e.g., basic, proficient, advanced). Discussion of these examples helps bring scorers onto the same "page" – sharing a common agreement on what exemplifies work at a given level. The generally agreed-upon form of determining score consistency is inter-rater agreement: the extent to which raters agree with each other's scores. Agreement rates of .8 and above are seen as strong and generally adequate for most purposes. The supports for ensuring high levels of agreement are the proper selection of the materials submitted for moderation, high-quality scoring guides, and thorough training of the assessors. In some systems, those who are unable to regularly reach appropriate levels of agreement are not certified as assessors.

Moderation is often used to establish reliable scoring, either as part of a training process or as part of the double scoring of tasks. Moderation is a process through which tasks are scored by two or more trained readers to help readers calibrate their judgments. Sometimes, moderation is used for tasks that are just at the "cut score" for a passing or failing grade. In such moderation sessions, especially if significant stakes are attached, two readers are assigned; and if they do not agree, a third, supervising reader makes the final determination. During scoring sessions, moderators may "drift" – for example, reading a series of especially good pieces may make a reader react too negatively to an average piece. To address these kinds of problems, in the stack of pieces a reader goes through there will be samples that have already been expertly scored (not revealed to the reader) so supervisors can check on drift.

Moderation results can be used to assign a final score or to provide feedback to teachers as part of a longer-term improvement process. This process has been used for both purposes in systems in the United Kingdom and in states such as Vermont. The Advanced Placement Art assessment also uses moderation to assign scores: trained judges score student artwork, giving each student his or her final AP score. And in many other AP courses, panels of teachers grade student essays. International Baccalaureate assessments, which are open-ended essays, projects, and products, are scored in a similar fashion.

The Learning Record, a system of assessment based on a tool developed in the U.K. to collect and evaluate samples of student work, uses moderation for long-term improvement. Only a random sample of Records from each participating classroom is re-scored. The scores given by the raters and their comments are returned to the originating teacher. While this does not change the score of any student, the evidence shows that, with feedback, teachers learn to evaluate their students more accurately.

Auditing is a similar means of checking on the reliability of locally-scored assessments. This approach has been used for many years for the New York State Regents examinations which, like examinations in most European and Asian countries, are routinely scored by teachers in their local schools. A proportionate sample of tests is pulled and re-scored each year, and when a school's scores are flagged, they can be re-evaluated. If a school's assessments are not properly calibrated, additional training and guidance can be used to bring them in line. In some systems in other countries, such as the school-based assessment system in Victoria, Australia, school inspectors examine the tasks and student work samples that are scored locally and provide an overview of the quality of the work that is part of the feedback to the school and to the state agency for guiding the process of continual improvement.

Validation and Verification of Locally-Based Performance Assessments: The Case of Nebraska

The verification and validation of locally-developed performance assessments in Nebraska is conducted with two primary considerations: a peer review balanced with technical expertise, and external benchmark validation.

Peer Review and External Technical Expertise

In Nebraska each school district is visited on site by a knowledgeable and trained team of peers who are teachers or administrators in other school districts and who have experience in developing local performance assessment. The trained peers gather information about each local assessment system based upon a pre-determined set of technical assessment criteria. The review team examines the evidence available in the district and conducts conversations with local staff to determine the methodologies and processes used for establishing valid, reliably scored assessment, reviewed for fairness and appropriate level. In addition to examining the processes and the assessments, the school district must provide the validity documentation and reliability calculations, assuring that their processes have produced fair, accurate assessment results of sufficient quality for state reporting.

The Six Quality Criteria, developed in collaboration with the Buros Center for Testing at the University of Nebraska are as follows:

- The assessment items/tasks match the standards and are sufficient enough to measure the standards.
- The students are assured the opportunity to learn.

- The assessment has been reviewed for bias and insensitive language or situations.
- The assessment is at the appropriate cognitive level.
- The assessment is reliably scored.
- The mastery levels are appropriately set.

The peer review team gathers evidence from each district, but does not assign the final rating. That is left to a team of assessment experts, who are psychometricians. Each peer review team is assigned to an assessment expert. The expert and the peer review team discuss the information gathered, and draft collaboratively written feedback entered in an electronic data system for districts to receive in a timely manner. The final rating and any suggestions for improved processes are provided to the district by the technical external expert but in the language of practitioners. The validation processes provide opportunities for districts to visit with their peers, feel comfortable in sharing the evidence of their processes, and yet have the opportunity to receive understandable feedback (filtered through the peer review team) from the measurement experts.

Additional work that is required is noted, and a formal appeals process is implemented where districts indicate their intent to resubmit additional or clarified evidence within a department determined time frame. The department conducts a second review contracting with a balance of peers and the external assessment experts.

Training for the peer reviewers is extensive. The first round of training consists of two days prior to the review week. A second round of training occurs on the first full day of the review week. The training itself is a collaborative process facilitated by one expert Nebraska peer, the department of education, and one external psychometrician. In this way, the review teams have the opportunity to see the collaboration and balance of local review and technical expertise.

Scoring rubrics are detailed, thorough, and distributed well in advance to districts. These scoring guides include clear expectations by the Department of Education for the evidence to be provided. The scoring process includes orientation, practice scoring with the scoring rubric, and team scoring. Reviewers practice the scoring process with samples of district evidence of varying quality that have been selected for the training. A set of visitation guidelines are reviewed with all peer reviewers so that each district can experience a similar procedure.

External Validation – Benchmark Checks

Locally-developed assessments are not the only data source used to determine how well students are performing inside Nebraska school districts. Multiple data sources are used to not only report student performance but to serve as a source of validation, or an "audit" of local assessment processes. Among the external validation benchmark "checks" in Nebraska are the following:

- A statewide writing assessment - generated, administered, and scored on the state level to all students in grades 4,8, and 11

- A required national achievement test required once in the elementary, once in the middle school, and once in the high school
- ACT results
- NAEP results

Additionally, each year the department conducts validity studies tracking the large-scale reading, mathematics, and writing results over time. These external tests are then correlated with the local assessment results. In this state, locally developed classroom-based performance assessments are an important part of a balanced assessment system.

Appendix 3: Performance Assessment: A Short Bibliography

Information on state assessment systems:

Nebraska

Gallagher, Chris. Reclaiming Assessment (Portsmouth, N.H.: Heinemann, 2007).

Nebraska Assessment web site at www.nde.state.ne.us/stars/

New York

NY Performance Assessment Consortium at www.performanceassessment.org

Wyoming

"Wyoming Steers Clear of Exit Exams," FairTest Examiner, January 2007. (www.fairtest.org/examarts/2007%20January/Wyoming.html) and http://www.k12.wy.us/Saa/WyCAS/archive/PubsPresent/Pubs/AssessmentHandbook.pdf

Multiple States

Darling-Hammond, Rustique-Forrester, & Pecheone, Multiple Measures Approaches to High School Graduation (Stanford University: School Redesign Network, 2005)

Information on International Approaches:

Queensland, Australia

http://education.qld.gov.au/corporate/newbasics/html/richtasks/richtasks.html

Information on Performance Assessment Systems:

Mathematics Resource Assessment System

http://www.nottingham.ac.uk/education/MARS/

Learning Record

http://www.cwrl.utexas.edu/~syverson/olr/olr.html, and

http://www.fairtest.org/Learning_Record_Home.html

New test measures students' digital literacy

Employers are looking for candidates who can navigate, critically evaluate, and make sense of the wealth of information available through digital media—and now educators have a new way to determine a student's baseline digital literacy with a certification exam that measures the test-taker's ability to assess information, think critically, and perform a range of real-world tasks.

The test, *iCritical Thinking* Certification, created by the Educational Testing Service and Certiport, reveals whether or not a person is able to combine technical skills with experiences and knowledge.

Today's students need to be able to think critically and effectively solve problems while using technology, Certiport explains—going beyond simply searching for information. They also must evaluate the legitimacy of the information, put it in context, and then apply problem-solving and decision-making skills.

"The test and certification program is designed to help employers [and educators] know that a student is ready for the workforce or for academia," said Quinn Sutton, Certiport's senior vice president.

Designed for students with at least a 10th grade reading level, *iCritical Thinking* allows students to demonstrate the ability to think critically within technology-enabled academic and workplace environments. About an hour in length, the test features 14 tasks based on real-world scenarios such as extracting information from a database, drawing conclusions from a spreadsheet, or composing an eMail based on findings—tasks students would be expected to do in the 21st-century workplace.

The test simulates the use of common, vendor-neutral applications to measure students' information and communications technology (ICT) literacy skills. Each task takes about four minutes to complete. The test produces individual score and group report summaries for instructors.

Monica Brooks, Marshall University's assistant vice president for Information Technology: Online Learning and Libraries, said her school plans to use *iCritical Thinking* beginning in the fall.

Marshall University will use the certification in two different ways. A sampling of freshman will take *iCritical Thinking* as a part of their first-year seminar as a way to benchmark skills and inform instructors about the topics that need to be covered.

Marshall is also part of a state-run program that helps working adults receive a Regents Bachelor of Arts degree.

Brooks plans to use the certification at the end of her Instructional Technology of Libraries class to measure how well students learned the advanced digital literacy skills taught during the class.

“It’s perfect timing [for the certification to be released], because people need these skills. People can Google, but a lot of times they don’t know what to do with that information,” she said. “It’s important for students to be proficient so they know how to use the data.”

Sutton said the certification, which was launched last November, is unique in that it isn’t a traditional multiple-choice test, but presents test-takers with real-life scenarios.

“It’s not product training,” Sutton said. “It’s seeing if you can use and apply the skills you possess. It reflects what we do every day.”

Sutton said the certification is geared toward the business environment, adding that as the test becomes more broadly available and understood, he believes more companies will begin to look for applicants with the *iCritical Thinking* Certification.

“Based on the current economy, it’s so relevant for students to become more competitive. They need to be able to hit the ground running,” he said.

Accessible Testing

So students can accurately show what they know and can do and teachers have the insights needed to foster student growth

Whether at the state, district, school, or classroom level, assessments provide very little—if any—useful information about what students know and can do if they can't access the tests. By access we mean not only understanding and being able to engage with test items, but also being able to provide evidence of learning and development. Inaccessible assessments fail to gain insights into students' strengths and weaknesses—so teachers lack meaningful guidance in providing appropriate instructional interventions. They also can have serious negative emotional and psychological impacts on students, by making them feel "stupid" and incapable of learning anything, so they might as well give up.

For years in statewide assessment programs, creating special forms in large print or Braille and reading items to students were common accommodations, but they fell far short of meeting the entire need for accessibility supports. So did such accommodations as providing extra time or a separate testing setting. Over the past decade, however, educators and researchers have been applying a concept initially developed by architects—universal design—to both learning and assessment. We have learned a great deal. Assessments—particularly at the state level—are far more accessible than in the past. While the processes used to develop state tests are extensive (and costly), the considerations and principles underlying universal design for assessment can inform the development of district, school, and even classroom assessments.

More recently, technological advances have created new possibilities for providing accessibility supports in digital assessment content for a wide range of student needs. Some offerings are specialized and fairly narrow (providing just text-to-speech, for example), whereas others are far more encompassing. The potential offered by these technologies and the increasing prevalence of online testing prompted a group of states to collaborate in developing open standards. The initial result, released in December 2010, is the Accessible Portable Item Profile (APIP) Standards. APIP contains three key components:

1. Standards for embedding accessibility supports in digital test content
2. Standards for developing a profile of student-specific accessibility support needs that drives the delivery of the supports during testing
3. Interoperability standards permitting the efficient and accurate transfer of the assessment content and student profiles across platforms

Adoption and compliance with the APIP Standards will greatly enhance our ability for at least 99% of students to demonstrate what they know and can do through online testing. While APIP-compliant assessments will be more costly to develop, the benefits of more efficiently and accurately defining and addressing student learning needs will be far greater.

Feds to schools: Make sure ed-tech programs are accessible

K-12 schools and colleges should build accessibility into their specifications for technology, and they should evaluate whether new hardware and software can be used by all students, including those with disabilities, as part of their ed-tech procurement process, the federal Education Department (ED) said May 26, 2011.

That recommendation was part of a "Dear Colleague" letter that ED's Office for Civil Rights issued to elementary and secondary schools and higher-education institutions, along with a set of frequently asked questions (FAQs) describing the legal obligation schools have to ensure that students with disabilities aren't left behind in any ed-tech implementation.

The letter expands on a document that ED issued last year, reminding schools and colleges of their responsibility to use accessible eReader devices after more than a year of complaints from sight-impaired students attending colleges that were piloting eReader programs.

Many eReader devices have a text-to-speech function that reads words aloud, but early generations of the devices often lacked menus that sight-impaired students could navigate.

ED's May 26 letter made it clear that schools' obligations under Section 504 of the Rehabilitation Act of 1973 and Title II of the Americans with Disabilities Act extend beyond eReader devices to include any ed-tech products or services. An accompanying FAQs document gave examples to help school leaders understand how to provide equal access to ed-tech services for students with disabilities.

"The purpose of the [letter] is to remind everyone that equal access for students with disabilities is the law and must be considered as new technology is integrated into the educational environment," the document said.

It added: "… All school programs or activities—whether in a 'brick and mortar,' online, or other 'virtual' context—must be operated in a manner that complies with federal disability discrimination laws."

When evaluating new ed-tech products and services, the document said, school leaders should ask the following questions:

• What educational opportunities and benefits will your school provide through the use of the technology?

• How will the technology help you provide these opportunities and benefits?

• Does the technology exist in a format that is accessible to individuals with disabilities?

• If the technology is not accessible, can it be modified, or is there a different technological device available, so that students with disabilities can enjoy the same educational opportunities and benefits in a timely, equally effective, and equally integrated manner?

For example, the document says, suppose your school intends to establish a web-based eMail system so that students can communicate with each other and with their instructors, receive important messages from the school, and communicate with others outside of school. You must make sure that these same benefits and opportunities exist for students with disabilities "in an equally effective and equally integrated manner."

Before deciding what system to buy, you should find out whether the system is accessible to students who are blind or have low vision—that is, whether the system is compatible with screen readers and whether it gives users the option of using large fonts. If a system isn't accessible as designed, you must figure out whether another, accessible product is available—or whether the inaccessible product can be modified so it's accessible to students who are blind or have low vision.

Schools don't necessarily have to provide the same type of emerging technology to students with disabilities as they give to other students, as long as the disabled students can enjoy the same educational benefits as the other students, the document said.

For instance, say you purchase eReader devices for your school library, but you then realize the devices can't accommodate students with disabilities. You can find and supply a different type of device for students with disabilities, as long as they derive the same benefits from using this alternative device.

"Technology can be a critical investment in enhancing educational opportunities for all students," said Russlynn Ali, ED's assistant secretary for civil rights. "The department is firmly committed to ensuring that schools provide students with disabilities equal access to the benefits of technological advances."

Digital Test Delivery:

Empowering Accessible Test Design to Increase Test Validity for All Students

MICHAEL RUSSELL

Executive Summary

The development of digital test content, coupled with computer-based test delivery, provides an important opportunity to improve the accessibility of test items. By applying principles of accessible test design, the next-generation assessment systems will deliver more valid inferences about student learning based on test scores for all students. Rather than developing assessment content for the general population of students and then making post hoc changes to accommodate the needs of subgroups of students, accessible test design provides a framework for making careful decisions about the methods used to tailor test administration to maximize the measurement of targeted constructs for each student. In turn, the Accessible Portable Item Profile (APIP) standards provide a tool for implementing accessible test design. The APIP standards empower next-generation assessments to solve three challenges. First, APIP provides a structure for specifying and storing the access needs of each student. Second, APIP provides a structure for augmenting item content with a variety of supplemental and alternate accessibility information designed to ensure that a test item functions properly for students with a variety of access needs. Third, APIP provides specifications for developing test delivery systems that can use a student access profile to tailor the provision of access tools (such as magnification, color contrast, masking) and the presentation of supplemental accessibility information (audio, Braille, tactile, or signed versions of item content). Collectively, the tools provided by APIP enable next-generation assessments to capitalize on the flexibility of digital technologies to maximize test validity for all students.

Russell, Ph.D. is Vice President of Innovation at the Nimble Innovation Laboratory, a division of Measured Progress that specializes in computer-based assessment solutions that are accessible for all students including those with disabilities and special needs. Dr. Russell is also an Associate Professor at the Lynch School of Education, Boston College.

This paper was produced in partnership with Arabella Advisors. Arabella Advisors is a philanthropy consulting firm supporting the efforts of individual, family, institutional, and corporate donors worldwide. We are committed to unbiased analysis that helps donors support issues and nonprofits with confidence. Our expertise and insights transform philanthropic goals into results. For more information on our firm, please visit our website at www.arabellaadvisors.com.

RECOMMENDATIONS

1. Adopt the Accessible Portable Item Profile (APIP) standards or a similar set of item profile standards that provide a comprehensive accessibility solution

2. Establish a best practices working group to develop business rules for implementing the APIP (or others as designed) accessibility elements

3. Require that an access needs profile be formed for each student

4. Carefully define the construct intended to be measured by each item or task

5. Require that accessibility information be embedded in each test item

6. Require that the test delivery system is compliant with the APIP standard or a similar set of item profile standards that provide a comprehensive accessibility solution

Introduction

For the past 40 years, the educational testing community has struggled to accurately and validly measure the achievement of students in the margins, including English language learners and students with disabilities or special needs. For a long time, many of these students were simply excluded from testing programs. Over time, however, legal action and advocacy led to the provision of test accommodations that allowed students in the margins to participate in testing programs. Initially, test accommodations were highly controversial and many programs either excluded or flagged scores for students who were provided test accommodations. It is only during the past decade, with the adoption of the No Child Left Behind Act, that students in the margins have been fully included in large-scale educational testing programs. Nonetheless, the equity and quality with which test accommodations are provided, the expense associated with the provision of test accommodation, and the effects that test accommodations have on the validity of test-based inferences continue to raise concerns.

The adoption of computer-based testing holds promise for addressing some of these concerns. In fact, the flexibility with which the delivery of digital content can be tailored for each individual student provides a unique and powerful opportunity to increase test validity for all students. This paper explores how digital test delivery empowers test developers to purposefully design items and tasks that maximally access the intended construct within each student. It begins by reviewing how test items are designed to function and then explores how digital content and digital delivery can be designed to create flexible test taking experiences that maximize the accuracy with which intended constructs are measured.

It is only during the past decade that students in the margins have been fully included in large-scale educational testing programs.

The Challenge: Designing Tests that Work for All Students

A test is designed to measure a specific construct or set of constructs. For ease of reference, the specific construct or set of constructs measured by a test is referred to as the intended construct. To provide a measure of the intended construct, a test is composed of a set of items or tasks or both. Each item or task is specifically constructed to create a measurement experience that requires the examinee (for purposes of this paper, we will assume the examinees are students) to apply the intended construct. Each measurement experience involves a carefully crafted three-step process. During the first step, the student is presented with information (or content) that is designed to establish a problem or question that stimulates the construct of interest. During the second step, the student is provided an opportunity to interact with content contained in the item or task while applying the intended construct. Since the application of the construct of interest cannot be directly observed, the third step requires the student to produce a response that represents the product or outcome of the application of the intended construct. It is through this three-step process that a test item or task attempts to access the intended construct as it currently operates within the student. Through observations of the student's application of the intended construct provided by multiple items or tasks, a test allows the outcome of each observation to be accumulated to produce a test score that is used to make an inference about the extent to which the intended construct operates within the student.

When creating a test item or task, item writers generally design items that function optimally for the general population of students. When considering the general population of students, several assumptions about students are often made. As an example, it is often assumed that the student does not experience any challenges with fine or gross motor skills, does not have any visual or auditory needs, is able to read near or above grade level, and is proficient in English. It is also assumed that executive functioning skills are adequate to allow the student to remain on task, identify key words or phrases in an item, work fluidly with multiple pieces of information, and break complex tasks into discrete elements. Finally, it is often assumed that the student is able to produce responses using either a pencil, keyboard, or mouse. While these assumptions may apply for the majority of students, they tend not to hold for students in the margins.

For these students, the knowledge and skills that item writers typically assume test takers possess present barriers to accurately and reliably accessing and measuring the intended construct. As depicted in figure 1, each phase of an item's functioning presents potential access barriers. As an example, most test items present content that is designed to stimulate the intended construct in narrative English form. For students who are English language learners, have visual impairments, have difficulty decoding text, read below grade level, or are unfamiliar with words or phrases that appear within an item (e.g., names of people or objects), the text-based narrative format of an item stem may not stimulate the intended construct as designed. Similarly, some items and tasks that present multiple pieces of information, contain multiple steps, or mix the format in which information is presented (through narrative text, tables, graphs, graphics, or video, among others) may be difficult to interact with for students who experience challenges with executive functioning, task prioritization, or information processing. Finally, items that require either written responses or the use of a mouse to select an option or manipulate digital objects may yield inaccurate responses for stu-

dents with physical disabilities or limited fine or gross motor skills. Across these three phases of an item's functioning, inaccurate stimulation of the intended construct, difficulty interacting with item content, or difficulty producing responses that accurately reflect the application of the construct produce challenges to an item's ability to measure the intended construct.

Test Accommodations: Why Post Hoc Changes Are Inadequate

For paper-based tests, accessibility challenges have traditionally been addressed by providing test accommodations. By definition, test accommodations are changes to test administration that address a specific access need. Traditionally, test accommodations are considered after a test form is developed and involve creating different versions of the test form (e.g., a large-print or Braille version), allowing an access assistant to translate a test into another form (e.g., reading an item aloud or presenting item content in American Sign Language), or making other changes to test administration conditions (e.g., allowing a student to use masks or reading guides, allowing a proctor to assist with managing test materials, or allowing the student to record answers in the test booklet). As noted above, this approach to enhancing access through post hoc test accommodations presents at least four challenges.

1. Creating alternate versions of test forms increases the expense for a testing program.

2. Because post hoc alternate versions are often made in the absence of item writers, the construct measured by an item may be changed inadvertently.

3. In addition to adding cost to a program, relying on access assistants to make post hoc translations of item content leads to differences in translations and the quality with which translations are delivered to students.

4. Inequity in the resources (both physical and human) available across schools produces inequities with respect to the opportunity to alter test administration conditions to meet specific needs.

As a result of these challenges, opportunities to receive test accommodations and the quality with which accommodations are provided vary across schools. Collectively, these challenges limit the effectiveness of traditional post hoc accommodations for improving the ability of test items to access the intended construct.

Digital test content and delivery provide an important opportunity to overcome these challenges. This opportunity results from two features of digital content and delivery: the ability to embed additional accessibility information into digital content files as an item is developed and the ability of a digital delivery system to selectively present subsets of that information to individual users based on their specific need.

Default and Supplemental Item Content

As Mislevy and his colleagues explain, different representational forms can be used to present item content to a student.[1] To enable a student to recognize and process content, the form used to present that content may need to be tailored based on the student's representational form need. As an example, content presented in print-based form will not adequately stimulate the intended construct for a student who is blind. However, when that same content is presented in Braille, the content is able to access the intended construct within a student who is a Braille reader. Similarly, for a student who is deaf or hard of hearing, content presented in audio form may not adequately stimulate the intended construct. However, when presented in a signed form, the content is able to stimulate the intended construct within a student who communicates in sign. Forms of alternate representation include reading content aloud, presenting text-based content in sign language, Braille, tactile representations of graphical images, symbolic representations of text-based information, narrative representations of chemical compounds (e.g., "sodium chloride" instead of "NaCl") or mathematical formulas, and translating to a different language.

Related to the notion of alternate representations are distinctions among default content, alternate content, and supplemental content. Default content is item information that is presented to a student who does not have defined access needs. Typically, default content includes text, graphics, and/or tables that form the item as developed for the general population of students.

Alternate content presents a different version of the item to students with specific needs. In essence, some or all of the original content is replaced by other content. Examples include using a translated version of an item (e.g., Spanish instead of English) or wording that is more easily understood (i.e., simplified English).

In contrast, supplemental content provides additional content to address access needs. For example, text might be supplemented with an audio, Braille, or signed form of the content. Similarly, to assist in identifying important aspects of content, supplementary information may be presented, such as highlighting of key words, translation or definitions for key words, or flags that point the student to key information.

As is described in greater detail below, digital content files allow item developers to specify default content and supplemental content for each item or task. In addition, a digital content file can provide a pointer to separate files that contain alternate versions of the default content (e.g., Spanish translation). Within a digital file containing an alternate version of an item, default and supplemental content for that alternate version can be provided (e.g., default content presented in Spanish accompanied by supplemental content specifying how to read aloud that content in Spanish). It's important to note that specifying default, supplemental, and alternate item information during the item development phase allows item writers to carefully consider whether supplemental and alternate information alters the intended construct measured by the item. If it does, the item writer can either modify the supplemental or alternate information or determine that supplemental or alternate information cannot be provided. By making these decisions during item writing, a test program can ensure that the same high-quality supplemental and alternate information is available for all students while assuring that the item still measures the intended construct.

Digital Item Delivery

Capitalizing on the flexibility of computer-based technologies, computer-based test delivery interfaces can tailor the presentation of default, alternate, and supplemental item content, interactions with that content, and response modes based on each individual's needs. To do so, developers should employ principles of universal design when creating systems that can personalize the testing experience based on each individual student's needs.

The concept of universal design focuses on "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design."[2] Rather than creating a single solution, universal design has come to embrace the concept of allowing users to select from multiple alternatives. As Rose and Meyer emphasize, "Universal design does not imply 'one sizes fits all' but rather acknowledges the need for alternatives to suit many different people's needs … the essence of [universal design] is flexibility and the inclusion of alternatives to adapt to the myriad variations in learner needs, styles, and preferences."[3]

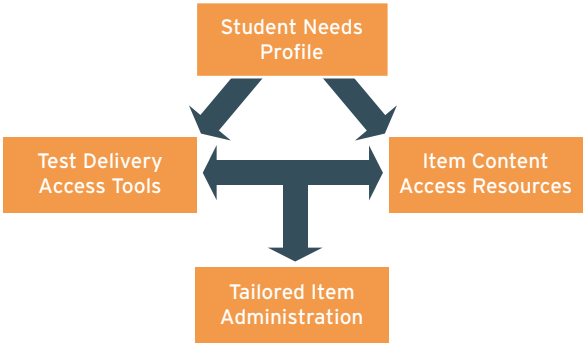
Technology allows developers to apply principles of universal design to educational assessments in a way that improves access for all users. When building a universally designed educational assessment, three important aspects must be considered:

1. To access the intended construct within each student, it must be acknowledged that a single default version of an item is not adequate. Instead, supplemental and alternate information that is carefully designed to access the intended construct within specific subgroups of students must be embedded into item content.

2. A user access needs profile must be developed for each student, and that profile must specify the method(s) of presenting, interacting, and responding to item content that are expected to best access the intended construct. The profile defines access needs for a given student and indicates which tools or representational forms should be made available for that student. The profile might also give specific settings, such as magnification levels, color contrasts, or default representational forms preferred by the student. Once defined, the access profile interacts with both the delivery interface and the item content.

3. The interface used to deliver items and tasks must be able to interact with each student's access needs profile and with the default, supplemental, and alternate information specified in each item. The interaction with the delivery interface focuses on specific tools or features embedded in the interface, activates whichever tools and features are defined in the profile, and, in some cases, controls the exact settings for those tools and features. The interaction with the item content focuses on which of the specific representational forms embedded in the item should be presented or activated for a given student in order to meet that student's specific need.

FIGURE 1:
ACCESSIBLE TEST IMPLEMENTATION MODEL



As depicted in figure 1, these three components must be designed to work together to tailor test delivery to meet the specific access needs of each student.

Accessible Test Design

Accessible test design addresses item content, representational forms, test delivery interface, and access profiles by specifying methods for flexibly tailoring an item so that the influence of non-targeted constructs is reduced for each individual student. Depending on a student's access needs, flexible tailoring may require an adaptation to the presentation of item content, the interaction with that content, the response mode, or the representational form in which content is communicated.

Adapted presentation may require item content to be presented in manner that assists the intake of information, such as magnifying or adjusting the contrast level with which item content is presented. Adapted interaction may require changes to the conditions under which a student applies the targeted construct, such as decreasing distractions by masking content, providing auditory calming, or highlighting key content within an item. Adapted response may require the student to use different types of tools to produce responses, such as a speech-to-text or an assistive communication device. Tailored representations may require item content to be presented using a different representational form, such as Braille, sign, audio, an alternate language (e.g., Spanish), or using simplified vocabulary.

Providing these adaptations and tailored representations in a consistent manner requires careful thought during the item development stage. To assure that adaptations and tailored representations do not negatively influence the validity of inferences based on the resulting test score, item developers must specify supplementary and alternate information associated with each item and assure that these alternate representations do not alter the measure of the intended construct. Providing adaptations and tailored representations in a consistent and valid manner also requires assessment programs to take a systems approach to accessibility. When designing a test delivery interface, an assessment program must specify the variety of accessibility tools and features that may be required for specific students. Finally, the test delivery system must be able to integrate student access information, item accessibility information, and interface accessibility tools and features to tailor item delivery to meet access needs for each individual student. Accessible test design requires a comprehensive model for test design and administration.

Accessible Test Design: An Example

For the past two years, the New England Common Assessment Program (NECAP) has implemented principles of accessible test design for its operational science tests and for its grade 10 operational mathematics and English language arts test. As part of this early adoption effort, the NECAP[4] states have allowed students for whom schools felt default item content did not accurately access the intended constructs to use a universally designed test delivery interface to perform the test. For these tests, digital versions of test items were developed. For each item, default and supplemental (but not alternate) item content was specified. Specifically, the supplemental item content focused on providing audio access to print-based content for students who had decoding challenges and for students who had vision needs. In addition, a universally designed test delivery interface was used to tailor the presentation of, interaction with, and response to items. Specifically, the test delivery interface allowed several types of tailored delivery that included:

- Presentation of content: (1) magnification; (2) reverse contrast; (3) color tint; and (4) alternate foreground and background color
- Supplemental content: (1) audio presentation of narrative content; (2) audio presentation of graphics; (c) audio presentation for nonvisual users; and (d) tactile presentation of graphics
- Interactions: (1) answer masking; (2) custom masking; (3) breaks; and (4) auditory calming
- Response: (1) mouse; (2) keyboard; (3) tab-enter controlled alternate communication devices; (4) touch screen; and (5) Intellikeys

While this early effort provides a powerful example that accessible test design can be implemented at scale for operational tests, it has also identified several challenges to the adoption of accessible test design.

Challenges to Accessible Test Design

Perhaps the most important challenge to address focuses on developing a clear definition of the intended construct measured by an item. Without a clear definition, it is difficult to determine whether supplemental and alternate item content alters the construct measured by the item.

A related challenge focuses on the need for access experts to interact with item content experts throughout item development. It is unlikely that any one individual will have sufficient knowledge of the intended construct and the various types of supplemental and alternate content required to allow the item to access that construct within all test takers. For that reason, item writing must involve an iterative team approach that includes a content expert and experts familiar with the specific needs addressed by the supplemental and alternate content (e.g., an expert in tactile representations for an item containing figures or graphs, or an expert in American Sign Language for an item that must be translated to ASL). While this may initially seem time- and labor-intensive, business rules for developing supplemental and alternate information can streamline this process.

Another challenge is that in order to adequately access the construct within each student, educators must make informed decisions about individual access needs and assign an appropriate access profile for each student. Since the concept of access needs and flexible test delivery is new to many educators, professional development and decision-making tools are required.

The development and use of a flexible test delivery engine that is able to interact with a student access profile and the default, supplemental, and alternate content associated with an item are also requisite. As new access needs are addressed by an assessment program and additional item information is incorporated into item files, it's important that the delivery engine be easy to adapt to incorporate these new techniques and tools.

Finally, there is a clear need for common expectations regarding the type of accessibility information that is included in an item file and the behaviors that should result when a test delivery system interacts with that information. Again, because the concept of accessible test design and tailored item delivery is new, consensus has not yet been reached about how some access needs are best met (e.g., how auditory support for sighted students who require decoding assistance may differ from auditory support for students who are blind). An important step toward reaching consensus about these issues is to form working groups that focus on developing business rules for providing specific types of supports (e.g., tactile and Braille representations) and guidelines for creating scripts for auditory and signed representations (e.g., how to read or sign dates, exponents, chemical equations, and so on). In addition, because this approach of developing supplemental and alternate item content is new, a standard for tagging or coding that information is needed so that items are interoperable across delivery systems.

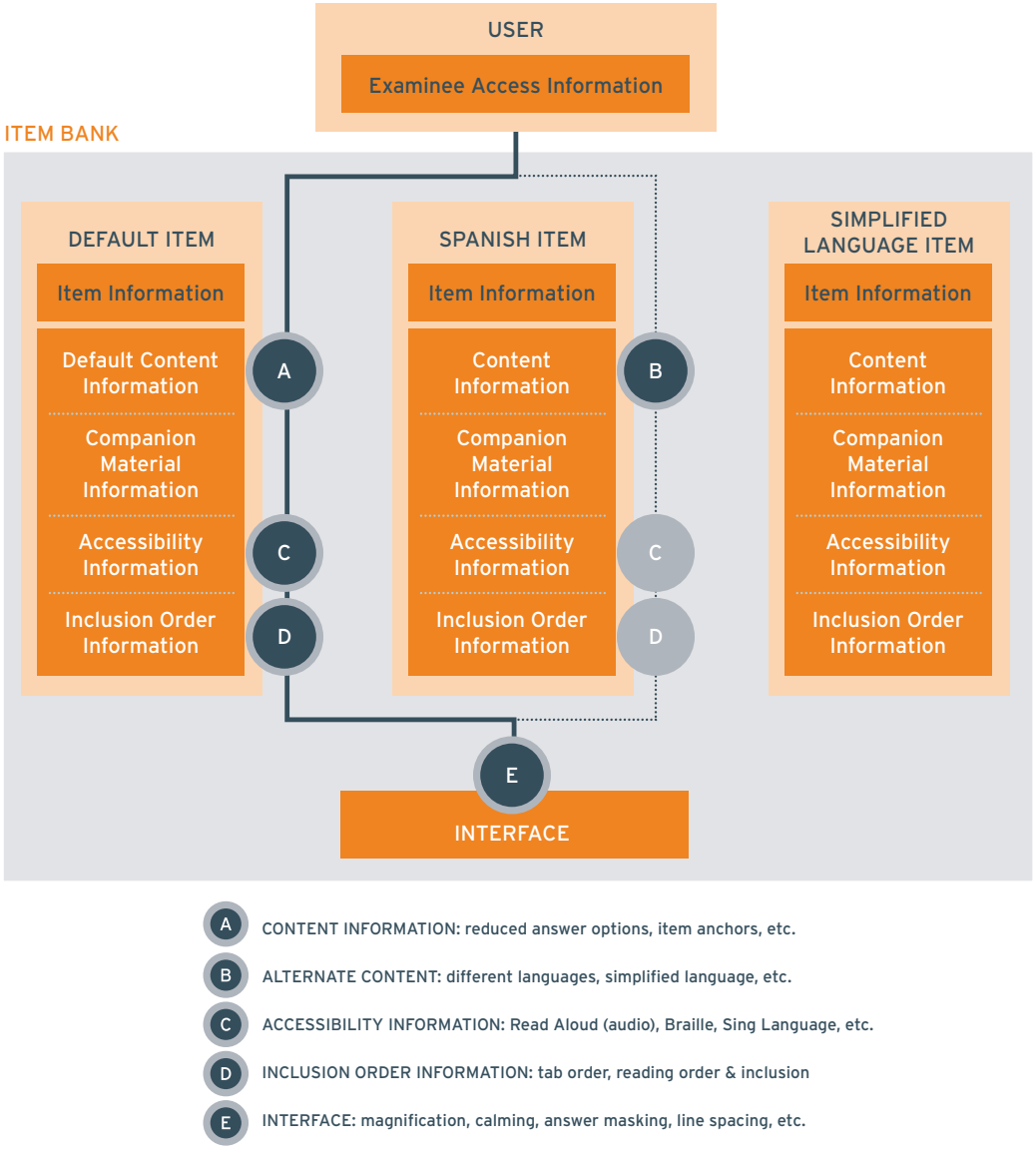
Accessible Portable Item Profile Standard

To meet many of these needs, the APIP Project has developed an open standard called the Accessible Portable Item Profile (APIP). Led by the Minnesota Department of Education, the APIP Project includes a consortium of eight states, IMS Global Learning Consortium (an interoperability standard-setting organization), and experts in testing, accessibility, and interoperability standards. Using the concept of accessible test design as a foundation, the APIP standard provides an open standard for specifying default, supplemental, and alternate item content, and for identifying the access needs for each individual student.

As depicted in figure 2, the APIP framework is composed of two information models that allow test delivery engines to tailor the presentation of items to meet the access needs of each individual student. The first information model focuses on Examinee Access Information. During test delivery, the Examinee Access Information model performs two functions. First, it provides information that allows a test delivery engine to activate specific tools that tailor the presentation of item content to the student. These embedded access features may include magnification, alternate contrast, increased white space, and answer masking. Second, it provides information that specifies which accessibility information embedded within the item model is pertinent to the student. APIP allows item developers to place a variety of types of access information within an item, including specifications for how an item is to be pre-

sented in auditory, Braille, sign, or tactile forms. In addition, the item information model allows an item developer to point to alternate versions of the item that are presented in an alternate language (e.g., Spanish) or in simplified English (e.g., with negatives removed).

FIGURE 2: APIP MODEL



As depicted in figure 2, the APIP item model has five components: Item Information, Default Content Information, Companion Material Information, Accessibility Information, and Inclusion Order Information. Each component is described briefly.

The Item Information component provides meta-information about an item. Examples of meta-information include the domain, sub-domain, intended construct, intended grade and age level, item difficulty, item discrimination, and item exposure rate. In addition, the Item Information component contains information about alternate versions of the item for which

item content has been substituted to make the item accessible in a different language or as a simplified version of the item in the default language.

The second component, Default Content Information, provides information about the contents of the item that are to be presented to a student assuming no access needs have been defined for that student. This information includes the item type, the prompt, media associated with the item (e.g., figures, tables, and graphs), response options, correct response, and scoring rule.

The third component, Companion Material Information, provides information about materials and tools that the student is expected to work with while performing the item. These materials and tools may include such things as a reading passage, a primary document, a periodic table, ruler, protractor, and calculator.

The fourth component, Accessibility Information, provides information about alternate representations of default content. The types of alternate representations specified in this component include audio, signed, or tactile presentation of item content. The Accessibility Information component may also include specifications for scaffold supports or key word translations.

The fifth and final component, Inclusion Order, specifies the order in which accessibility elements are to be presented to a student with a given category of access needs. The categories of access needs for which an inclusion order is specified include: (1) text-based audio; (2) graphic audio; (3) text-based and graphic audio; (4) nonvisual audio; (5) Braille; (6) American Sign Language; and (7) other sign language. Text-based audio access is specific to students who are able to view contents displayed on a screen but may need assistance accessing those contents. Often, students requiring text-based audio access read below grade level, have reading related disabilities, or the language in which item content is displayed is not their primary language. When audio forms are presented as core content, the item developer must specify the order in which content is presented. In addition, the item writer must consider whether all item elements are presented as core content in audio form. As an example, an item writer may opt not to present labels associated with a graphical element as part of the core content, or may opt not to read the contents of a table when the item is read from beginning to end. In these cases, item content that is not presented as core content may be accessed by the student on demand.

In sum, the APIP model empowers item developers and testing programs with a standard method for specifying the tailoring of items to meet specific accessibility needs and provides a foundation for accessible test design.

Looking to the Future

The Race to the Top (RTTT) Assessment Program holds promise to stimulate several advances in the field of assessment. Among those advances are the development of innovative item types and performance tasks that can be used at scale to assess complex, higher-order skills. Wide-scale adoption of adaptive testing is likely. In the area of formative assessment, there is potential for diagnostic instruments that help educators identify what students know and don't know and that provide information about misconceptions or other underdeveloped reasoning that may interfere with a student's conceptual understanding.

Across all forms of assessment and types of tasks used to collect information about student knowledge and understanding, it will be important that instruments adequately access the intended construct within each individual student. Accomplishing this will require a proactive *a priori* approach to developing accessible assessment content. Given the large number of instruments and accompanying items and tasks that must be developed, as well as the new ways in which students will be expected to interact with item and task content, accessibility will be a major challenge. However, the development of such tools as APIP and flexible test delivery interfaces provide an opportunity to examine issues of accessibility during prototyping and early stages of development. By doing so, the field can be proactive in identifying potential issues that specific item types or item content may present for students with specific access needs. Early identification will then allow researchers and development teams to explore alternate item and task designs that may overcome these access barriers. Or, in some cases, test designers and item developers may decide that the access need is so tightly intertwined with the intended construct (e.g., vision intertwined with using a microscope to locate and identify microscopic organisms) that the construct cannot be validly measured for students with that access need. Such decisions will have subsequent implications for the metadata associated with an item, the design of adaptive test engines, the generation of student scores, and ultimately the types of inferences that can be made about a student based on that score. While the flexibility that digital content and digital delivery offer with respect to tailoring item and task delivery to maximize access to the intended construct within every student may seem overwhelming, tools like APIP empower assessment programs to make important decisions throughout the test development process about what exactly is being measured, what is not intended to be measured, and how to tailor item and task deliver to maximize validity for all students.

Recommendations

- Take a proactive *a priori* approach to developing accessible item and task content. Rather than repeating past practices that have raised questions about the extent to which test accommodations change the measured construct, create item writing teams that are knowledgeable about the intended constructs and about the provision of supplementary accessibility information that does not alter the construct measured by the item.

- Adopt the established Accessible Portable Item Profile (APIP) standards or a similar set of item profile standards that provide a comprehensive accessibility solution, and use it as a foundation for item development, student rostering, and test delivery systems. The APIP standards provide a structure for specifying and storing student access needs and defining supplemental and alternate item accessibility information. Test delivery systems, in turn, must be developed to read and interpret a student access needs profile and to tailor the provision of accessibility options based on that profile.

- Form a working group to develop best practices for implementing APIP standards, or a similar set of item profile standards that provide a comprehensive accessibility solution. The standards adopted should be powerful tools for specifying how content is presented to students with specific access needs. To assure that the tools are applied in a consistent manner that does not violate a measured construct, business rules must be developed to help guide decisions about how to present content to meet a specific need (such as how to read aloud exponents, dates, or scientific notation, or how to present simple and complex tables in auditory and Braille form, and so on). Ideally, the best practices working group will include representatives from both RTTT assessment consortia and will comprise experts in specific content areas and in specific access needs, as well as experts with experience in implementing accessible test design.

- Clearly define the construct intended to be developed by each item and identify access-related constructs that are not intended to be measured by each item. Without a clear definition of the intended

Take a proactive *a priori* approach to developing accessible item and task content.

construct, sound decisions about the type of supplemental accessibility information embedded within an item are impossible. It is only with clear definitions of the intended construct that sound decisions can be made about whether supplemental accessibility information alters the measured construct or supports valid test-based inferences about student achievement of that construct.

- Take a team approach to item development that incorporates supplemental and alternate content into items. Provision of supplemental accessibility information requires an understanding of the intended construct and the varied accessibility needs of students. No one person possesses all of this knowledge. Hence, a team approach is needed to develop and evaluate the quality of accessibility information provided for each item.

- Be prepared to accept that some items may not be able to be made accessible for students with a specific access need and develop test specifications or adaptive algorithms that take this limitation into consideration. There will be intended constructs that overlap with access needs (e.g., using a microscope to identify microscopic objects overlap with vision). When these cases occur, it may not be possible to develop supplemental or alternate versions of the item that provide valid measures of the intended construct for students with the overlapping access need. In such cases, decision rules must be developed about whether to present the item to students and how to incorporate the item in the total test score.

- Develop an access needs profile for every student, and provide professional development and tools to support decisions about access needs. The concept of an access needs profile will be new to many educators. To help ensure that sound decisions are made about the assignment of access needs, educators will need training. Where possible, tools that help educators make informed decisions should also be developed and made available to teachers in all schools.

- Use test delivery systems that can interpret student access profiles, flexibly activate embedded access tools, and selectively present default, supplemental, and alternate item content based on each student's need. The power of accessible test design requires a sophisticated test delivery engine that is able to read and interpret student access profiles and that can use a profile to customize the delivery interface and the presentation of supplemental and alternate accessibility information based on the profile. Early adoption by NECAP demonstrates that tailored delivery is possible for a limited number of access needs. Careful design and development must build on this early adoption to meet a wider variety of access needs.

Notes

1) Robert Mislevy, et al., "On the Roles of External Knowledge Representations in Assessment Design," *Journal of Technology, Learning, and Assessment* 8(2) (2010). Retrieved January 21, 2010 from http://www.jtla.org.

2) Center for Universal Design (CUD), "About UD: Universal Design Principles" (1997) http://www.design.ncsu.edu/cud/about_ud/udprinciplestmlformat.html (accessed February 13, 2009). Archived at http://www.webcitation.org/5eZBa9RhJ.

3) D. Rose and A. Meyer, "Universal Design for Learning," *Journal of Special Education Technology* 15 (1) (2000): 66-70.

4) NECAP states are: New Hampshire, Rhode Island, and Vermont.

Using Systematic Item Selection Methods to Improve Universal Design of Assessments

Policy Directions 18

Christopher Johnstone • Martha Thurlow • Michael Moore • Jason Altman

September 2006

Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). Using Systematic item selection methods to improve universal design of assessments (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Policy18/

Background

The No Child Left Behind Act of 2001 (NCLB) and other recent changes in federal legislation have placed greater emphasis on accountability in large-scale testing. Included in this emphasis are regulations that require assessments to be accessible. States are accountable for the success of all students, and tests should be designed in a way that provides all students an opportunity to demonstrate their knowledge and skills. With the reauthorization of the Individuals with Disabilities Education Act in 2004, states are required for the first time to incorporate universal design principles in developing and administering tests, to the extent feasible.

Applying the concept of universal design to statewide assessment means that assessments are designed from the beginning and continually refined to result in more valid inferences about performance of students with diverse characteristics. These assessments are based on the premise that each child in school is a part of the population to be tested, and that test results should not be affected by disability, gender, race, or English language ability. While universally designed assessments are not intended to eliminate individualization, they may reduce the need for accommodations by eliminating access barriers associated with the tests themselves. At the same time, the intent of the measurement—the intended content and construct of the assessment—are not changed.

Including universal design in test construction is already taking place in the majority of states. According to a survey of states conducted by the National Center on Educational Outcomes (NCEO), during the 2004–2005 school year 43 states addressed issues of universal design. More than half of the states addressed universal design at the item development and review levels, and by including it in RFPs for test development.

There are many elements involved in creating universally designed assessments. They include

making sure that students with disabilities are part of field testing, for example. But a major focus of universal design in assessments is making sure that the items included in the assessment are appropriate. There are several methods for selecting items to ensure that they optimize the characteristics of universal design. The purpose of this Policy Directions is to provide an overview of these item selection methods, and to suggest that a combination of the methods will produce the better result. Each method has strengths and weaknesses, may lead to different results, and is in different stages of current practice (see Table 1). Although each method has merits, NCEO recommends states employ all methods systematically and in conjunction with each other.

Table 1. Item Analysis Methods: Pros, Cons, and Current Practice

Method	**Strengths**	**Weaknesses**	**Current Practice**	**Improvement on Current Practice**
Expert Review	Provides structured review of items by experts	Does not provide actual student data	Unstructured "sensitivity" review panels	Provides reviewers with tools to make decisions
Statistical Analyses	Provides significance data on field tested items; flags potentially problematic items	Validity questionable for small populations	Only DIF calculations performed	Multiple analyses conducted, providing patterns of flagged items
Think Aloud Methods	Provides information on why particular items function as they do	Does not provide data across groups	Not currently in widespread use	Provides important design information

---

Expert Review

Once an assessment is designed and in a format suitable for previewing, it is important for states to have sensitivity review teams examine the assessment. The use of review teams to examine items is common practice in many states, and is generally encouraged by test vendors. When creating bias and content review teams, it is important to involve members familiar with disability and language issues. Grade level experts, representatives of major cultural and disability groups—or those who can reflect their needs—researchers, and teaching professionals all make up an effective review team.

NCEO, working closely with experts in the fields of assessment, disability, reading, mathematics, and language acquisition, developed and refined a set of considerations for test developers and item reviewers to use to ensure that tests are universally designed. The considerations are listed in Table 2 (see Technical Report 42 listed in Resources). They ask six important questions about test items. Test item developers and reviewers can use these questions

to help determine the extent to which an item provides appropriate accessibility without changing the intended construct.

There are considerations that item and test developers or reviewers can use to help determine whether a test overall—not just the individual items—is universally designed:

- Do all visuals (e.g., images, pictures) and text provide information necessary to respond to the item?
- Is information organized in a manner consistent with an academic English framework with a left-right, top-bottom flow?
- Can booklets/materials be easily handled with limited motor coordination?
- Are response formats easily matched to question?
- Is there a place for taking notes (on the screen for computer-based tests) or extra white space with paper pencil?

These considerations show that it is not just each item that is important, but also how the whole test is put together that is also an important aspect of universal design.

Table 2. Considerations for Universally Designed Assessment Items

Does the Item...
Measure what it intends to measure? • Reflect the intended content standards (reviewers have information about the content being measured)? • Minimize knowledge and skills required beyond what is intended for measurement?
Respect the diversity of the assessment population? • Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)? • Avoid content that might unfairly advantage or disadvantage any student subgroup?
Have clear format for text? • Standard typeface? • Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group), and adaptable font size for computers? • High contrast between color of text and background? • Sufficient blank space (leading) between lines of text? • Staggered right margins (no right justification)?

Have clear visuals (when essential to item)?

- Visuals are needed to answer the question?
- Visuals have clearly defined features (minimum use of gray scale and shading)?
- Sufficient contrast between colors?
- Color alone is not relied on to convey important information or distinctions?
- Visuals are labeled?

Have concise and readable text?

- Commonly used words (except vocabulary being tested)?
- Vocabulary appropriate for grade level?
- Minimum use of unnecessary words?
- Idioms avoided unless idiomatic speech is being measured?
- Technical terms and abbreviations avoided (or defined) if not related to the content being measured?
- Sentence complexity is appropriate for grade level?
- Question to be answered is clearly identifiable?

Allow changes to its format without changing its meaning or difficulty (including visual or memory load)?

- Allows for the use of braille or other tactile format?
- Allows for signing to a student?
- Allows for the use of oral presentation to a student?
- Allows for the use of assistive technology?
- Allows for translation into another language?

As states and other testers explore the use of computers for testing, it is important to have ways to judge their appropriateness and universal design features as well. There are additional considerations for computer-based tests to ensure that they are universally designed (see Table 3).

When using Expert Review considerations, NCEO recommends incorporating the following into the review:

- Conduct the review as early as possible in the stages of test development.
- Include disability, technology, and language acquisition experts in item reviews.
- Provide professional development for item developers and reviewers on use of the universal design considerations.
- Present the items in the format in which they will appear on the test.
- Include standards being tested with the items being reviewed.
- Try out items with students (use Think Aloud methods).
- Field test items in accommodated formats.

- Review computer-based items on computers.

Table 3. Considerations for Universally Designed Computer-based Tests

Layout and design

- Sufficient contrast between background and text and graphics for easy readability.
- Color alone is not relied on to convey important information or distinctions.
- Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options).
- Stimulus and response options are viewable on one screen when possible.
- Page layout is consistent throughout the test.
- Computer interfaces follow Section 508 guidelines (www.section508.gov).

Navigation

- Students have received adequate training on use of test delivery system.
- Navigation is clear and intuitive; it makes sense and is easy to figure out.
- Navigation and response selection is possible by mouse click or keyboard.
- Option to return to items and return to place in test after breaks.

Screen reader considerations

- Item is intelligible when read by a text/screen reader.
- Links make sense when read out of visual context ("go to the next question" rather than "click here").
- Non-text elements have a text equivalent or description.
- Tables are only used to contain data, and make sense when read by screen reader.

Test specific options

- Access to other functions is restricted (e.g., e-mail, Internet, instant messaging).
- Pop up translations and definitions of key words/phrases are available if appropriate to the test.
- Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.
- Students are able to record their responses and read them back as an alternative to a human scribe.
- Students are allowed to create persistent marks to the extent that they are already allowed on paper-based booklets (e.g., marking items for review; eliminating multiple choice items, etc.).

Computer capabilities

- Adjustable volume.
- Speech recognition available (to convert user’s speech to text).
- Test is compatible with current screen reader software.
- Computer-based option to mask items or text (e.g., split screen).
- Computer software for test delivery is designed to be amenable to assistive technology.

Statistical Analysis

A quantitative approach to selecting items that appear to provide access to students with certain characteristics, such as disabilities, is to conduct statistical analyses on test item results. Many statistical methods exist, ranging from simple methods based on classical test theory to complex methods based on contemporary item response theories (IRT). Four statistical approaches currently used in the field by researchers have practical usefulness for identifying items that potentially violate universal design principles for groups of students.

Table 4 shows the four statistical analysis methods that are useful for flagging items to identify those that are potentially more challenging for particular students. Each of these has been used to identify items that may not be universally designed (see Technical Report 41 listed in the Resources). Statistical methods are based on various assumptions that determine when items should be flagged as producing different performance from what would be expected. The flagging of an item is taken as an indication that the item may not be as accessible as possible, and may be creating barriers that do not allow the student to demonstrate his or her knowledge and skills.

Statistical analyses are useful for understanding which items may be biased and need revision, but they do not provide information on why particular items function the way they do. Understanding why is often critical to knowing what to do in making revisions to items.

Table 4. Four Statistical Analysis Methods for Reviewing Test Item Results

Method	**Procedure**
Item Rankings	Items are ranked from most to least difficult for total population and for particular groups. It is assumed that ranks should be similar for groups and total.
Item Total Correlation (ITC)	Within-group investigation of how individual items correlate with the total score on the same test; poor correlations may signal a problem. Different ITCs for the same item across different groups of test takers may indicate that the item behaves differently across those groups.
Differential Item Functioning (DIF) – Contingency Table Methods	Performance on items is compared for large groups that perform at similar levels on the total test. In contingency table methods DIF statistics are calculated by comparing the proportion of students answering an item correctly in target and comparison groups with the same total score range; statistically significant differences may point to an item's problematic nature.
Differential Item Functioning (DIF) – Item Response Theory Approaches	In IRT, characteristics of each item are represented by an item response curve, which is a function of individual test takers' characteristics called "latent traits." Items are compared based on their item response curves between target and comparison groups.

Think Aloud Methods

Think aloud methods provide a way to answer questions about why particular items may be problematic for students. For state assessments, think aloud methods tap into the short-term memory of students who complete assessment items while they verbalize. Researchers believe that the verbalizations produced in think aloud studies provide excellent information because they are not yet in the long-term memory. Once experiences enter long-term memory, they may be tainted by personal interpretations. Therefore, an excellent way to determine whether design issues really do exist for students is to have students try out items themselves.

Think aloud methods have been used in research to identify problematic items for students with disabilities (see Technical Report 44 in Resources). When students verbalize everything they think while completing an item, it becomes easy to see how the design of the item affects the understanding of the item. If a student does have difficulty with the item, it will also be easy to determine whether the difficulty is a result of design features or a lack of curricular knowledge. The depth of understanding that results from think aloud techniques is this method's strength. Follow-up questions can supplement any unclear data derived from think aloud techniques.

Recommendations

Attaining universal design in statewide assessments is a goal for states as they continually refine and improve their assessments. With the understanding that this goal means that states are retaining the constructs and content that their tests are intended to measure, the question becomes how to identify items that may violate the principles of universal design.

It is important to have a systematic approach to reach universal design in assessments. Research is paving the way to identifying techniques that are workable. Any one technique by itself, however, may be insufficient. The methods identified here will reduce the possibility of erroneously flagging and eliminating items that reflect poor performance due to students' lack of opportunity to learn.

Specifically, using sets of considerations for expert review can make the test development process more transparent, informed, and focused on the needs of the entire population of students and ensure that the assessment results are more meaningful for the widest range of students. Statistical analysis methods can help pin-point test items that are potentially problematic and that may have universal design issues. Think aloud methods can be used with students themselves who can provide information that will help illuminate whether there are design issues that need to be addressed. These are all aspects of striving to reach universal design, which holds the promise of improved student performance. This goal can be reached without compromising the validity of the assessment.

Resources

Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners (Technical Report 44). Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Tech44/.

Analyzing Results of Large-Scale Assessments to Ensure Universal Design (Technical Report 41). Johnstone, C.J., Thompson, S.J., Moen, R.E., Bolt, S., & Kato, K. (2005). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Technical41.htm.

Consideration for the Development and Review of Universally Designed Assessments (Technical Report 42).Thompson, S.J., Johnstone, C.J., Anderson, M.E., & Miller, N.A. (2005). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Technical42.htm.

Universal Design Applied to Large-Scale Assessments (Synthesis Report 44). Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Synthesis44.html.

Universally designed assessments: Better tests for everyone! (Policy Directions No. 14). Thompson, S., & Thurlow, M. (2002). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Policy14.htm.

Considerations for the Development and Review of Universally Designed Assessments

---

NCEO Technical Report 42

Published by the National Center on Educational Outcomes

Prepared by:

Sandra J. Thompson • Christopher J. Johnstone • Michael E. Anderson • Nicole A. Miller

November 2005

Acknowledgements

NCEO extends its sincere appreciation to the expertise of the individuals who provided us with thoughts, feedback, and suggestions in order to further develop and refine the considerations for universally designed assessments:

Karen Barton, CTB McGraw Hill
Sheryl Burgstahler, DO-IT Center, University of Washington
Margo Gottlieb, Illinois Research Center
Tom Haladyna, Arizona State University
Tracey Hall, CAST, Inc.
Barbara Henderson, American Printing House for the Blind
Scott Marion, National Center for the Improvement of Educational Assessment
Ken Olsen, Mid South Regional Resource Center
Marge Petit, National Center for the Improvement of Educational Assessment
Charles Stansfield, Second Language Testing, Inc.
Gerald Tindal, University of Oregon
Carol Traxler, Gallaudet University
Tim Vansickle, Minnesota Department of Education

Executive Summary

Universal design is an approach to educational assessment based on principles of accessibility for a wide variety of end users. Thompson, Johnstone, and Thurlow described seven elements of universally designed assessments in their 2002 report entitled *Universal Design Applied to Large Scale Assessments*. Elements of universal design include inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility. Since the 2002 report, Universal Design Project staff have examined research from a variety of fields in an effort to specify how elements of universally designed assessments can be put into practice.

This report describes the development of a "considerations of universally designed assessments" form based on Thompson et al.'s original elements. Considerations are specific questions for test designers to take into account while designing assessments. This report provides the original list of considerations from Thompson et al., then describes a validation process, whereby assessment and content area experts participated in a Delphi study. The Delphi study illuminated expert consensus on some considerations and disagreement on others. All expert commentary is captured in the text of this paper and in Appendix C (in tabular form), and a revised list of considerations is found in Appendix D.

Based on the comprehensive work represented in this report, several recommendations are presented for the use of the considerations of universal design at all stages of test development:

1. Incorporate elements of universal design in the early stages of test development.
2. Include disability, technology, and language acquisition experts in item reviews.
3. Provide professional development for item developers and reviewers on use of the considerations for universal design.
4. Present the items being reviewed in the format in which they will appear on the test.
5. Include standards being tested with the items being reviewed.
6. Try out items with students.
7. Field test items in accommodated formats.
8. Review computer-based items on computers.

Introduction

The term universal design has been applied to a variety of educational approaches over the past several years. For instance, universal design for learning was first described by the Council for Exceptional Children (CEC) in a *Research Connections* article (CEC, 1999). Likewise, Thompson, Johnstone, and Thurlow (2002) of the National Center on Educational Outcomes (NCEO) described universal design approaches to large-scale assessment. In their initial paper on universal design of assessments, Thomson et al. outlined seven elements of universally

designed assessments (inclusive assessment population; precisely defined constructs; accessible, non-biased items; amenable to accommodations; simple, clear and intuitive procedures; maximum readability and comprehensibility; and maximum legibility). Although elements of universal design provide guidance to states and assessment companies about design issues, there is still a need for specific information concerning what considerations should be made in test development in order to make tests accessible to a wide range of students.

This report summarizes the process of developing and refining a list of considerations for the universal design of statewide assessments for all students, including students with disabilities and English language learners. The staff of the Universal Design Project at NCEO, working closely with experts in the fields of assessment, disability, content areas (reading and math), and language acquisition, completed this version of considerations in the summer of 2004. This revision was one of three, which followed the compilation of an initial set of considerations identified from a literature review of multiple content areas (see Thompson, et al., 2002). The first version included stakeholder input from the Council of Chief State School Officers (CCSSO) conference on large-scale assessment in 2003. Following CCSSO feedback, a second version (a Delphi review, see description later in the text) was developed by NCEO in partnership with the Minnesota Department of Education, with a primary focus on students with limited English proficiency. This report describes the process of refining the considerations during a third validation study conducted by the Universal Design Project at NCEO. This is the third version of the considerations for use by test developers and item reviewers. This report also discusses the process used to validate the considerations, the issues that arise when using these considerations, and recommendations for use.

Purpose of the Study

The purpose of this report is to describe the process of developing and refining a set of considerations for item developers and item review teams to take into account in the universal design of inclusive, standardized, statewide assessments. Although the goal of this process was to find design strategies that maximize the accessibility of tests and test items, a larger goal was to create an instrument to guide careful consideration of the elements of test design in order to discover issues in items that may be problematic.

What is Universal Design?

More than 20 years ago, Ron Mace, an architect who was a wheelchair user, began to actively promote a concept he termed “universal design.” Mace was adamant that his field did not need

more special purpose designs that serve primarily to meet compliance codes and may also stigmatize people. Instead, he promoted design that works for most people, from the child who cannot turn a doorknob to the elderly woman who cannot climb stairs to get to a door (Mace, 1998).

The term universal design is found in the newly reauthorized Individuals with Disabilities Education Act of 2004 (Public Law No: 108-446). Specifically, IDEA of 2004 states that:

> The State educational agency (or, in the case of a districtwide assessment, the local educational agency) shall, to the extent feasible, use universal design principles in developing and administering any assessments under this paragraph (§ 612(a)(16)(E).

Universal design is specifically defined in the U.S. Assistive Technology Act of 2004 (Public Law No. 108-364-ATA 2004) as follows:

> [A] concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly accessible (without requiring assistive technologies) and products and services that are interoperable with assistive technologies.

Assessments that are universally designed are designed from the beginning, and continually refined, to allow participation of the widest possible range of students, resulting in more valid inferences about performance. These assessments are based on the premise that each child in school is a part of the population to be tested, and that test results should not be influenced by disability, gender, race, or English language ability. Universally designed assessments are not intended to eliminate individualization, but they may reduce the need for accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves.

The elements of universal design, according to Thompson et al., are:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear and intuitive procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

From these elements, universal design staff constructed considerations for universally designed assessments. The considerations are a list of specific questions that help test designers locate potential design issues in items. The considerations are listed in Table 1.

Table 1: Considerations for Universally Designed Assessment Items

Does the item…
Measure what it intends to measure • Reflect the intended content standards (reviewers have information about the content being measured) • Minimize skills required beyond those being measured
Respect the diversity of the assessment population • Accessible to test takers (consider gender, age, ethnicity, socio-economic level) • Avoid content that might unfairly advantage or disadvantage any student subgroup
Have clear format for text • Standard typeface • Twelve (12) point minimum for all print, including captions, footnotes, and graphs (type size appropriate for age group) • Wide spacing between letters, words, and lines • High contrast between color of text and background • Sufficient blank space (leading) between lines of text • Staggered right margins (no right justification)
Have clear pictures and graphics (when essential to item) • Pictures are needed to respond to item • Pictures with clearly defined features • Dark lines (minimum use of gray scale and shading) • Sufficient contrast between colors • Color is not relied on to convey important information or distinctions • Pictures and graphs are labeled
Have concise and readable text • Commonly used words • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Idioms avoided unless idiomatic speech is being measured • Technical terms and abbreviations avoided (or defined) if not related to the content being measured • Sentence complexity is appropriate for grade level • Question to be answered is clearly identifiable
Allow changes to its format without changing its meaning or difficulty (including visual or memory load) • Allows for the use of braille or other tactile format • Allows for signing to a student • Allows for the use of oral presentation to a student • Allows for the use of assistive technology • Allows for translation into another language
Does the test…

Have an overall appearance that is clean and organized

- All images, pictures, and text provide information necessary to respond to the item
- Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow

In addition to the other considerations, a computer-based test should have these considerations:

Layout and design

- Sufficient contrast between background and text and graphics for easy readability
- Color is not relied on to convey important information or distinctions
- Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options)
- Stimulus and response options are viewable on one screen when possible
- Page layout is consistent throughout the test
- Computer interfaces follow Section 508 guidelines

Navigation

- Navigation is clear and intuitive; it makes sense and is easy to figure out
- Navigation and response selection is possible by mouse click or keyboard
- Option to return to items and return to place in test after breaks

Screen reader considerations

- Item is intelligible when read by a text/screen reader
- Links make sense when read out of visual context ("go to the next question" rather than "click here")
- Non-text elements have a text equivalent or description
- Tables are only used to contain data, and make sense when read by screen reader

Test specific options

- Access to other functions is restricted (e.g., e-mail, Internet, instant messaging)
- Pop up translations and definitions of key words/phrases are available if appropriate to the test
- Students are able to record their responses and read them back (and have them read back using text-to-speech) as an alternative to a human scribe, but only if student has experiences with this mode of expression and chooses it for the test

Computer capabilities

- Adjustable volume
- Speech recognition available (to convert user's speech to text)
- Test is compatible with current screen reader software
- Computer-based option to mask items or text (e.g., split screen)
- Computer software for test delivery is designed to be amenable to assistive technology

---

Delphi Review

We conducted a Delphi review to determine the usefulness of existing considerations for universally designed assessments. The intent of the Delphi review was to invite experts in the fields of assessment, special education, academic content, and language acquisition to give input on the considerations and modify them accordingly (Adler & Ziglio, 1996). The Delphi method is a structured process of using a series of questionnaires to gather the combined input from a group of persons with expertise related to a specific area or population. The method has been

used in the social science and public health fields since the mid-1970s (Adler & Ziglio, 1996). Delphi studies allow participants to give their own informed opinion on an issue. The input is then compiled and returned to the participants who can respond to further questions, respond to the input from the other participants, and revise their own comments if desired. All iterations of Delphi are anonymous.

This Delphi study took place entirely by e-mail. Participants were unaware of who was invited to participate in the study, who elected to participate, and the individuals who provided feedback (anonymity was maintained throughout the study). All suggestions and comments were given equal weight.

Participants

Universal Design Project research staff identified a group of experts to review the considerations for universally designed assessments. To ensure that important areas of expertise were represented, a chart was created and participants were recommended based on their expertise in one or more of the identified areas (see Table 2). These individuals were then invited to participate in the Delphi review before the first Delphi questionnaire was sent out. The resulting group of Delphi participants represented experts in the field of assessment, assistive technology, computer-based testing, reading, math, second language acquisition and testing, disability consultation, and special education.

Table 2: Expertise and Participants

Vision	Barbara Henderson
Computer-based testing, learning disabilities	Gerald Tindal
Item analysis	Karen Barton
Second language acquisition and testing	Margo Gottlieb
Second language acquisition, testing, and translation	Charles Stansfield
Physical disabilities	Sheryl Burgstahler
Hearing	Carol Traxler
Science	Scott Marion
Psychometrics	Tom Haladyna
Assistive technology	Tracy Hall
Math	Marge Petit
Special education assessment	Ken Olsen

State Assessment Director	Tim Vansickle

Delphi Process

The first Delphi survey (Delphi Form 1—see Appendix A) was developed to obtain specific feedback on the considerations draft presented by NCEO. Expert participants were provided ample opportunity to comment on the considerations or add to the list. The participants were asked first to rate the importance of each individual consideration on a five point Likert scale. They then were asked to comment on any of the considerations about which they felt strongly positive or negative. They could also pose questions on the form. Finally, they were asked to add any additional considerations and rate the importance of their additions. The participants were instructed to try to think about the considerations in terms of their usefulness for test developers and item reviewers.

In July 2004, the first Delphi survey (Delphi Form 1) was e-mailed to the participants. Each participant was given seven days to review the considerations and email comments back to NCEO. The comments and ratings were returned by 13 of 14 participants. These were compiled at NCEO and a second survey was developed (Delphi Form 2–see Appendix B).

The second survey (Delphi Form 2) included a list of anonymous individual ratings and the mean from all ratings assigned to each consideration. All comments made by the participants on the first form were included in the second form. Participants were asked to comment on results from the initial survey, were probed on specific issues by NCEO researchers, and were asked to comment on the 15 considerations suggested by participants (the majority relating to computer-based testing). The second survey was e-mailed out at the beginning of August 2004 and participants were again given seven days to return their comments via email. The comments were complied by the staff at NCEO in mid-August, 2004 (see Appendix C).

Response Rates

The original list of considerations (Delphi Form 1) was sent out via e-mail to 14 experts for review. Thirteen of 14 (93%) experts returned Delphi Form 1. The second survey (Delphi Form 2) was again sent out to the original 14 participants. The same thirteen participants returned the second survey (one participant did not participate in either survey). The feedback on both surveys was extensive.

Results

Using the feedback from both Delphi surveys, Universal Design Project staff revised the considerations for universally designed assessments (see Table 3). The considerations that had originally been sent to reviewers were rated as somewhat important to extremely important (from 2.67 to 5), with an average of very important (i.e., 4.3) to consider in designing and reviewing assessments. One consideration was deleted based on expert feedback, while others were added or revised. The primary additions to the considerations were the expansion of the considerations for computer-based testing. In addition, there were several additions to the discussion points for the consideration note sections. All changes to the considerations are shown in Table 3, with additions marked by underlines and deletions shown by strikethroughs.

Table 3: Summary of Consideration Ratings and Changes

Does the item…	Range	Mean
Measure what it intends to measure		
• Reflect the intended content standards (reviewers have information about the content being measured)	5–5	5.00
• Minimize knowledge and skills required beyond ~~those being~~ what is intended for ~~measured~~ measurement.	3–5	4.33
Respect the diversity of the assessment population		
• ~~Accessible~~ Sensitive to test taker~~s~~ characteristics and experiences (consider age, gender, ethnicity, ~~and~~ socio-economic level, region, disability, and language)	4–5	4.75
• Avoid content that might unfairly advantage or disadvantage any student subgroup	4–5	4.64
Have clear format for text		
• Standard typeface	3–5	4.00
• Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group)	3–5	4.09
• ~~Wide spacing between letters, words, and lines~~	2–5	3.09
• High contrast between color of text and background	3–5	4.09
• Sufficient blank space (leading) between lines of text	2–5	2.82
• Staggered right margins (no right justification)	2–5	3.36
Have clear visuals (when essential to item)		
• ~~Pictures~~ Visuals are needed to ~~respond to item~~ answer the question	3–5	4.56
• ~~Pictures~~ Visuals with clearly defined features (minimum use of gray scale and shading)	4–5	4.45
• ~~Dark lines (minimum use of gray scale and shading)~~	3–5	3.82
• Sufficient contrast between colors	1–5	3.64
• Color alone is not relied on to convey important information or distinctions	2–5	3.91
• ~~Pictures and graphs~~ Visuals are labeled	3–5	3.91
Have concise and readable text		
• Commonly used words (except vocabulary being tested)	1–5	4.18
• Vocabulary appropriate for grade level	4–5	4.83

• Minimum use of unnecessary words	1–5	4.17
• Idioms avoided unless idiomatic speech is being measured	3–5	4.67
• Technical terms and abbreviations avoided (or defined) if not related to the content being measured	4–5	4.73
• Sentence complexity is appropriate for grade level	1–5	4.45
• Question to be answered is clearly identifiable	5–5	5.00
Allow changes to its format without changing its meaning or difficulty (including visual or memory load)		
• Allows for the use of braille or other tactile format	3–5	4.67
• Allows for signing to a student	3–5	4.55
• Allows for the use of oral presentation to a student	3–5	4.36
• Allows for the use of assistive technology	3–5	4.45
• Allows for translation into another language	1–5	3.64
Does the test…		
Have an overall appearance that is clean and organized		
• All visuals (e.g., images, pictures) and text provide information necessary to respond to the item	3–5	4.50
• Information is organized in a manner consistent with an academic English framework with a left-right, top-bottom flow	4–5	4.33
• Booklets/materials can be easily handled with limited motor coordination (consideration was added)	0–5	4.00
• Response formats are easily ~~correlated~~ matched to question	0–5	3.43
• Place for student to take notes (on the screen for CBT) or extra white space with paper-pencil	0–5	3.82
In addition to the other considerations, a computer-based test should have these considerations:		
Layout and design		
• Sufficient contrast between background and text and graphics for easy readability	4-5	4.67
• Color alone is not relied on to convey important information or distinctions	2–5	3.92
• Font size and color scheme can be easily modified (through browser settings, style sheets or on-screen options)	2–5	4.08
• Stimulus and response options are viewable on one screen when possible	3–5	4.67
• Page layout is consistent throughout the test	4–5	4.75
• Computer interfaces follow Section 508 guidelines (www.section508.gov)	0–5	3.56
Navigation		
• Students have received adequate training on use of test delivery system	0–5	4.46
• Navigation is clear and intuitive; it makes sense and is easy to figure out	4–5	4.92
• Navigation and response selection is possible by mouse click or keyboard	3–5	4.67
• Option to return to items and return to place in test after breaks	3–5	4.60
Screen reader considerations		
• Item is intelligible when read by a text/screen reader	3–5	4.58
• Links make sense when read out of visual context. (“go to the next question” rather than “click here”)	4–5	4.67
• Non-text elements have a text equivalent or description	3–5	4.30
• Tables are only used to contain data, and make sense when read by screen reader	3–5	4.36
Test specific options		
• Access to other functions is restricted (e.g., e-mail, Internet, instant messaging)		

• Pop up translations and definitions of key words/phrases are available if appropriate to the test	3–5	4.55
• Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.	3–5	4.08
• Students are able to record their responses and read them back ~~(or have them read-back using text-to-speech) as alternative to human scribble, but only if student has experiences with this mode of expression and chooses it for the test~~ as an alternative to human scribe.	0–5	2.67
• Students are allowed to create persistent marks to the extent that they are already allowed to paper-based booklets (e.g., marking items for review, eliminating multiple choice items, etc.)	0–5	3.69
Computer capabilities	0–5	4.17
• Adjustable volume		
• Speech recognition available (to convert user's speech to text)	3–5	4.50
• Test is compatible with current screen reader software	1–5	3.67
• Computer-based option to mask items or text (e.g., split screen)	3–5	4.25
• Computer software for test delivery is designed to be amenable to assistive technology	0–4	3.00
	0–5	3.91

Notes that were added to the considerations address some of the anticipated issues that might arise when using the considerations. While we tried to keep the list of considerations brief and user-friendly, it was clear from participant comments that more explanation about the intent and issues surrounding the considerations needed to be presented close to the considerations in note form. The notes are not meant to be used as definitive judgment of the "good" or "bad" quality of an item or design feature. Instead, the notes are intended to add clarity to the considerations, help elucidate important issues, and help generate discussion.

Discussions About Selected Considerations

In addition to providing greater clarity to several of the considerations, many of the respondents in the Delphi review pointed out that using some of the considerations depended on the content being tested. Extensive discussion focused on issues of construct vs. content validity and the minimization of construct-irrelevant variance. There was also extensive discussion on the validity and practicality of the translation of assessments to languages other than English. In this section of the report, we present a detailed review of these discussions. Considerations about which few comments were made and no clarification was deemed necessary are not discussed. Responses to all considerations, however, can be found in Appendix C.

Consideration: *"Reflects the intended content standards (reviewers have information about the content being measured)."*

Following a discussion by Universal Design Project staff, Delphi participants were asked to comment on whether the first consideration should remain "Reflects the intended content standards (reviewers have information about the content being measured)" or whether it should be reworded "Reflects the intended *construct* (reviewers have information about the *construct* being measured)." Although opinions leaned toward changing the wording (Yes = 6, No = 3, Combination wording = 1, Did not state position but provided information to consider when making the decision = 2, Don't know = 1), only two of the participants in favor of using the term "construct" provided reasoning. One suggested that construct "would fit better with the professional terminology," while the other stated that "content is topical, constructs are conceptual. This difference in meaning is huge. Furthermore, construct is a term used in APA standards and is deeper than content."

The participants who wished the consideration to remain the same provided critical information about what to think about before a decision could be made. Specifically, one participant suggested that we consider our audience: "Construct is a formal term that theorists use. Content standards [are] what practitioners understand." Another participant suggested we consider what the terms imply: "…construct is a sort of overarching concept (i.e., reading) whereas content standards are…narrower (e.g., reproduces capital letters)…If the test is supposed to be a standards-based achievement test, then it must address standards. If not, then the item need only address the construct."

Ultimately, Universal Design Project staff decided to retain the term "content." This term appears to be consistent with the link of items to standards, and avoids the apparent confusion surrounding the term "construct." It should be noted, however, that the term "construct" may still be useful, especially if item developers (who are familiar with the concept of constructs) are using these considerations.

Consideration: *"Minimize knowledge and skills required beyond ~~those being~~ what is intended for ~~measured~~ measurement."*

The second consideration under review was altered slightly based on participant input. Initially, this consideration stated, "Minimize skills required beyond those being measured." This was changed to "Minimizes knowledge and skills required beyond what is intended for measurement" following several suggested alternate phrases. In addition to suggestions on phrasing, Delphi participants expressed concern that item writers or reviewers might interpret this consideration in such a way as to "…separate skills too much…[and thus run the risk that] we'll wind up with tests that measure isolated, basic skills." Still others expressed the belief that this consideration has direct relevance for the measurement of "higher level thinking." Yet, as another reviewer questioned, "how…the other skills (are) defined and targeted" would be important in guiding item writers and reviewers. One participant summed up the issue by saying

that it "…depends on how discrete the standards are; minimal skills can be embedded in more complex contextualized items. Ultimately, it depends on what you are measuring."

Consideration: *"~~Accessible~~ Sensitive to test taker~~s~~ characteristics and experiences (consider age, gender, ethnicity, ~~and~~ socio-economic level, region, disability, and language."*
The third consideration was changed from "Accessible to test takers (consider age, gender, ethnicity, and socio-economic level" to "Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)." When asked about including the term "bias" in this consideration, participants were somewhat divided. While some indicated that bias should be included to "reference systematic variance that interferes with making a valid inference," others clarified that "bias and accessibility are separate issues from a review standpoint, though obviously related." Keeping participants' suggestions and reasoning in mind, it was decided that the term "bias" would be included in the note portion of the consideration and that the demographic variables would be expanded from four to seven, reflecting the need for greater sensitivity to the experiences of very diverse populations.

Consideration: *"Standard typeface."*
When considering the clarity of the format for text in assessments, most participants agreed that a standard typeface was important. There was, however, confusion about the meaning of "standard." Some Delphi participants had interpreted this consideration as implying that a single standard font existed, as illustrated in the following comment: "There is no standard typeface, thus the myriad fonts used in various publisher's files, even within the same text or textbook." In order to reduce confusion over the meaning of the term, however, it was determined that additional clarification was needed. Consequently, the following was added to the note section: "Use clear, common, familiar, and consistent fonts," followed by examples of font.

Consideration: *"Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group)."*
When considering which font size to select, several Delphi participants noted the importance of considering the font style. Given the fact that a 12-point font can vary in size depending upon the font style, an additional issue was included in the note section. As suggested, one consideration (width of spacing between letters) was combined with font. One participant stated "Wide spacing is not necessarily best; proper font selection is more important." Consequently, this consideration was added to the note section of the consideration addressing font.

Consideration: *"High contrast between color of text and background."*
When considering the use of color in text or background, participants suggested going beyond the issue of contrast to consider print density. Specifically, one participant stated, "[E]ven with sufficient color contrast, color blind users may not be able to distinguish text and background. [I] suggest you further recommend high *print density* contrast. This would also avoid isoluminance effects for non-visually-impaired students." ("Isoluminance" is the point at which two colors

have an equivalent luminous intensity, or brightness.) Based on these comments, information on print density and isoluminance was added to the note section for the consideration addressing format for text.

Consideration: *"~~Pictures~~ Visuals are needed to ~~respond to item~~ answer the question."*
The use of visuals resulted in considerable discussion ranging from issues surrounding limiting visuals, the use of visuals to provide only redundant information, and the benefits/drawbacks of using visuals in relation to specific disabilities. In relation to the content of visuals, for example, it was suggested, "Pictures, line art, etc. should be related to the item [and] should enhance understanding, [but] not [be] required for understanding, with the exception of data tables like on math and science tests." Additionally, another Delphi participant stated, "often there are pictures used that are not redundant with the text but that are relevant to the item and to the construct." Consequently, it was suggested that the wording of this consideration take this idea into account. Rather than dramatically change the wording of this consideration, qualifying information was provided to the note portion below the consideration addressing the idea that clear and well-designed graphics or pictures should add value for students who need a visual cue.

Consideration: *"Commonly used words (except vocabulary being tested)."*
When considering the vocabulary used in assessments, both for directions and specific items, many Delphi participants commented on the need for greater clarity surrounding the specification that the text be comprised of "commonly used words." Several participants suggested that the term "age-appropriate" was preferable, while another suggested adding "concise and readable." Ultimately, the greatest concern with this particular consideration was that there be some acknowledgement that the words selected should be common, "with the exception of subject specific terminology…" In other words, the "item should consist of commonly understood words or vocabulary…" except when knowledge of specific vocabulary is being tested. One participant also suggested that the vocabulary be "…consistent with each specific grade level," with another suggesting "at or below grade level [when] reading is not the primary construct tested." As a result of this feedback, additional clarification was added to the wording of the consideration (i.e., the consideration was changed from "Commonly used words" to "Commonly used words (except vocabulary being tested)" as well as in the note section following the consideration.

Consideration: *"Allows for translation into another language."*
Perhaps the most controversial consideration of all was "Allows for translation into another language." One and one-half pages of initial comments, questions, and suggestions were followed by an additional one and one-half pages of responses, comments, questions, and suggestions. The response of one participant summarized a number of the issues that participants grappled with when determining the appropriateness of this consideration:

> "This is a questionable and highly controversial issue, particularly when one realizes that such a standard is impossible to meet. About 72% of our LEP students are Spanish speakers, but the other 28% represent many diverse languages. How do we accommodate and what is the theoretical rationale and what is the technology for doing this? Is it possible? Is it beneficial?"

In reference to the impracticality of translating tests into the less commonly represented language groups, some participants questioned the fairness of accommodating some students (e.g., Spanish speakers) and denying others. Another stated "What harm is done by helping the 72% of LEP students who speak Spanish? We provide accommodations to others where possible, but some would propose that a translated test is harmful. Poppycock!"

Participants also suggested some disagreement in terms of the quality of the translations/skill of the translators. A primary problem with translation, however, was clear: "The limitation is money. Translation must be cost effective like everything else in education. You can't provide translated tests for very small numbers. The Lau decision (*Lau v. Nichols*, 1974) and other civil rights decisions make it clear that numbers dictate expectations of school systems." Given the cost, customized dictionaries were suggested as a possible alternative to fully translated tests.

Besides the practicality/impracticality of translating tests, one area of considerable debate surrounded the validity of the inferences that can be made from scores derived from translated tests. Some participants expressed the belief that translated tests reduced the validity of scores ("Data analysis has shown these to be less than valid measures of student performance."), or that certain translations would result in less valid scores ("Some critical and relevant word/concepts [do] not translate into every language."). Others, however, made the argument that there are few instances where concepts do not translate:

> "Minnesota translates to Hmong and Somali. Only in these languages are there relevant words/concepts that do not translate easily into English. The other languages of state assessment (Spanish, Russian, Chinese, Korean, Haitian Creole) almost never pose a problem for translating words or concepts. Professional translators will tell you they can translate almost any word or idea, and if they encounter one they can't, they will tell you that too."

Another participant added, "Translation is no more a threat to validity than a change in option order or a change in font. Such changes might generate a miniscule change in item difficulty, but they don't affect validity... [Translation] is the exact same test stated in a different language." Yet others brought up the issue of validity in reference to a specific construct being measured. For example, two participants stated that translating English language arts (ELA) tests would invalidate the inferences that could be made from the scores. In light of NCLB legislation, a participant brought up a final important point of consideration: "A translated test is

always much less of a threat to validity and score comparability than an alternate assessment," suggesting that a translated test is preferable to alternate assessment measures for English language learners.

Two reviewers suggested that this consideration be eliminated given the controversy, at least until more research was available. Ultimately, Universal Design Project research staff decided to retain this consideration, acknowledging the issues item writers and reviewers face as they incorporate this consideration into the test construction/revision process. This information was included in the note section following the consideration.

Summary of Revisions

At the completion of the study, the Universal Design Project staff revised the original considerations based on Delphi responses (Appendix D). The most extensive revisions were made to the content and wording of the considerations. Some of the most significant changes to the considerations that resulted from the Delphi process are described here:

1. Wording of several of the considerations was revised using feedback from the Delphi review participants. For example, "Minimize skills required beyond those being measured" was changed to "Minimize knowledge and skills required beyond what is intended for measurement" and "Accessible to test takers (consider age, gender, ethnicity, and socio-economic level)" was changed and expanded to "Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)."

2. Computer-based testing considerations were expanded. Much of the useful feedback for this section came from reviewers who are familiar with the development of computer-based tests. With these revisions, the section of considerations for computer-based testing was clarified and redundancies with other considerations were eliminated.

3. Notes were added to the considerations. These notes discuss some of the anticipated issues that might arise when using the considerations. While we tried to keep the list of considerations brief and user-friendly, it was clear that more explanation about the intent and issues surrounding the considerations needed to be presented on the same page. The notes are intended to add clarity to the considerations and help elucidate important issues. Notes also provide evidence of the complexity of some of the considerations and illustrate that considerations are not static rules, but general principles that aid in flagging potentially problematic items.

4. One font-dependent consideration ("Wide spacing between letters, words, and lines") was eliminated. Instead it was included in the note section for "Have a clear format for text."

5. Relevant research citations were added to the considerations so that people wanting to investigate a certain issue in more depth would have the resource citations at hand (see Appendix E).

6. We created a review checklist of the considerations for item reviewers and developers (see Appendix F). This form is intended to be used by item reviewers and developers who have received training on the considerations. It consists of a list of the considerations, without the supporting text. Using this form, item reviewers and developers can go through items and flag for further discussion areas of concern or alteration. For item reviewers, there is an additional form on which comments may be recorded explaining why some aspect of an item was flagged (Appendix G).

Issues Related to Universal Design

One of the most important outcomes of this review process was the identification of issues that surround the development of universally designed assessments. These issues highlight the complexities of a process without easy answers. The issues discussed in this section are not meant to be an exhaustive list of the challenges related to the universal design of assessments, but instead provide some guidance about the challenges that might be encountered when using the considerations.

1. **Universal design is not a cure all**. Just because a test is universally designed, or has used the elements of universal design to guide its development, does not mean that a test is accessible to all students. The considerations recommended in this report are just that, considerations. They are meant to be used to guide test developers and reviewers in creating tests that are accessible to the greatest number of students possible. However, some changes to a test that might make it more accessible to one group of students, might actually make it less accessible to another group. For example, eliminating or altering an illustration accompanying an authentic reading text may clarify an item by removing a distraction for some students. On the other hand, eliminating it may remove or change some useful context for the passage. Issues of accessibility need to be carefully considered and discussed openly so that informed decisions can be made without hindering the construct being tested. Universal design can be a useful tool for developing better assessments, but it is not a tool that can magically make all tests accessible to all students.

2. **Universal design does not replace accommodations**. While universal design may remove some barriers for students with disabilities and English language learners, it in no way eliminates the need for testing accommodations. Some students may still need accommodations such as large print or assistive technology. A goal of universally designed assessments is to anticipate common accommodations and design tests that allow accommodations to be more easily integrated into the format of the test.

3. **Universal design does not replace good instruction**. The goal of universal design is to think about the full range of students taking an assessment so that they all can demonstrate what they have learned. A student who has not had an opportunity to learn the material tested will not be helped by a universally designed test.

4. **Universal design does not lower standards**. Some may perceive a universally designed assessment to be a "watered-down" or "easier" assessment. It is important to make clear the purpose of universal design is to make sure that the content being tested is more universally accessible to all of the students taking the test and thus a better measure of student learning.

5. **Technology use is challenging**. The quality of technology available across schools is an important issue when creating a computer-based assessment. It is difficult to anticipate what accessibility issues will arise when a test is administered on a variety of different systems with a variety of assistive technologies. Trying to anticipate these issues is important, however, when reviewing items.

Recommendations

These considerations can be used to make assessments more universally accessible to the entire population of test takers. Here are some specific recommendations for the use of the considerations of universal design at all stages of test development.

1. **Incorporate elements of universal design in the early stages of test development**. Universally designed assessments present an opportunity to bring more people to the table in the early stages of test development including experts in disability, language acquisition, and technology. These experts are able to give more structured input at different stages of the test development process if they understand universal design and have these considerations for item development and review at hand. It is more cost effective to consider universal design in the early stages of item development, rather than at the end when items have already been developed and field-tested.

2. **Include disability, technology, and language acquisition experts in item reviews.** Every effort should be made to involve experts in item review who can judge whether items meet all of the considerations.

3. **Provide professional development for item developers and reviewers on use of the considerations for universal design**. Explanation and discussion of each consideration will ensure use by item developers and reviewers.

4. **Present the items being reviewed in the format in which they will appear on the test**. When item reviewers examine items to be included in an assessment, it is important to format items as closely as possible to how they will appear on the test. Since many of the considerations have to do with format, it is not useful to look at items that are not in the font, size, or format in which they will appear in the actual test booklet.

5. **Include standards being tested with the items being reviewed**. Above all other considerations, the first consideration—does the item measure what it intends to measure—is of primary importance in constructing universally designed assessments. Consequently, item review teams using the considerations of universal design to guide their work must have the standard (grade level expectations) that each item is intended to test at hand. It is only by knowing what an item is intended to test that reviewers can judge whether an element of the item might interfere with student access. Each item needs to be presented with the corresponding standard being tested in that item.

6. **Try out items with students**. Some of the elements of an item that distract or confuse students are not easily recognizable by adults or native English speakers. For this reason, trying items out with students by conducting think-aloud studies can provide valuable information about whether an item is testing the content intended (Thompson, Johnstone, & Miller, in press).

7. **Field test items in accommodated formats.** In order to ensure that the content an item is intended to measure is not being changed when an accommodated format of a test is being used, include students using accommodated test formats in field tests. While this can add additional expense to the field test, there are ways of doing such studies that can progressively build a database. For example, a field test could focus on the use of certain accommodated formats one year and others the next, building up a database for the various forms of the test. Again, qualitative data from student interviews in this area can provide important information that can be used to improve items.

8. **Review computer-based items on computers.** To judge whether computer-based items are universally designed, item reviewers need to use the technology that will be

used to deliver the test. Using a paper print-out of an assessment does not allow a review team to meaningfully consider the format of the test.

Conclusion

We hope that the process detailed in this report has produced not only a better set of considerations of universally designed assessments for all students, but has also clarified some of the opportunities and challenges that universally designed assessments present. While using universal design does not guarantee the accessibility of any test to all students, using the considerations to openly discuss issues of test design throughout the test development process can make any assessment more inclusive. Making the process of test development more transparent, informed, and focused on the needs of the entire population of students will help ensure that the assessment results are more meaningful for the widest range of students.

References

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: interactions with student language background*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Adler, M., & Ziglio, E. (Eds.). (1996). *Gazing into the Oracle: the Delphi method and its application to social policy and public health*. London: Jessica Kingsley Publishers.

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, A.G. (1985). *Becoming a nation of readers*. Urbana, IL: University of Illinois, Center for the Study of Reading, National Institute of Education, National Academy of Education.

Arditi, A. (1999). *Making text legible.* New York: Lighthouse.

Assistive Technology Act of 2004 (Brief Title: ATA 2004). (P.L.108-364).

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32,* 323–340.

Brown, P.J. (1999). *Findings of the 1999 plain language field test*. Newark, DE: University of Delaware, Delaware Education Research and Development Center.

Calhoun, M.B., Fuchs, L., & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23,* 271–282.

Carter, R., Dey, B., & Meggs, P. (1985). *Typographic design: Form and communication.* New York: Van Norstrand Reinhold.

Cole, C., Tindal, G., & Glasgow, A. (2000). *Final report: Inclusive comprehensive assessment system research, Delaware large scale assessment program.* Eugene, OR: Educational Research Associates.

Council for Exceptional Children (1999). *Universal design: Research connections.* Retrieved September 3, 2004, from the World Wide Web: http://ericec.org/osep/recon5/rc5sec1.html

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67* (1), 67–92.

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats.* West Columbia, SC: Center for Rehabilitation Technology Services. (ERIC Document No. ED 405689)

Gregory, M., & Poulton, E.C. (1970). Even versus uneven right-hand margins and the rate of comprehension in reading. *Ergonomics, 13* (4), 427–434.

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76,* 35–40.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hanson, M.R. (1997). *Accessibility in large-scale testing: Identifying barriers to performance*. Delaware: Delaware Education Research and Development Center.

Hanson, M.R., Hayes, J.R., Schriver, K., LeMahieu, P.G., & Brown, P.J. (1998). *A plain language approach to the revision of test items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 16, 1998.

Harker, J.K., & Feldt, L.S. (1993). A comparison of achievement test performance of nondisabled students under silent reading plus listening modes of administration. *Applied Measurement, 6,* 307–320.

Hartley, J. (1985). *Designing instructional text* (2nd Edition). London: Kogan Page.

Heines. (1984). *An examination of the literature on criterion-referenced and computer-assisted testing*. ERIC Document Number 116633.

Hoener, A., Salend, S., & Kay, S.I. (1997). Creating readable handouts, worksheets, overheads, tests, review materials, study guides, and homework assessments through effective typographic design. *Teaching Exceptional Children, 29,* (3), 32–35.

Individuals with Disabilities Educational Improvement Act (Brief Title: IDEA 2004). (P.L. 108-446).

Johnstone, C.J., Miller, N.A., & Thompson, S.J. (in press). *Using the think aloud method (cognitive labs) to evaluate test design*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington DC: Council of Chief State School Officers.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: Center for Research on Standards and Student Testing.

Lau vs. Nichols, 414 U.S. 563, 94 S.Ct. 786 (1974).

MacArthur, C.A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: handwriting, word processing, and dictation. *Journal of Special Education, 21* (3), 22-42.

Mace, R. (1998). *A perspective on universal design.* An edited excerpt of a presentation at Designing for the 21st Century: An International Conference on Universal Design. Retrieved January, 2002, from the World Wide Web: www.adaptenv.org/examples/ronmaceplenary98.asp?f=4.

Menlove, M., & Hammond, M. (1998). Meeting the demands of ADA, IDEA, and other disability legislation in the design, development, and delivery of instruction. *Journal of Technology and Teacher Education. 6* (1), 75–85.

Muncer, S.J., Gorman, B.S., Gorman, S., & Bibel, D. (1986). Right is wrong: An examination of the effect of right justification on reading. *British Journal of Educational Technology, 1* (17), 5–10.

National Research Council. (1999). *High stakes: testing for tracking, promotion, and graduation.* In J. Heubert & R. Hauser (Eds.), Committee on Appropriate Test Use. Washington, DC: National Academy Press.

Osborne, H. (2001). In other words…communication across a life span…universal design in print and web-based communication. *On Call* (January). Retrieved January, 2002, from the World Wide Web: www.healthliteracy.com/oncalljan2001.html.

Popham, W.J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W.J., & Lindheim, E. (1980). The practical side of criterion-referenced test development. *NCME Measurement in Education, 10* (4), 1–8.

Rakow, S.J. & Gee, T.C. (1987). Test science, not reading. *Science Teacher, 54* (2), 28–31.

Schiffman, C.B. (1995). *Visually translating materials for ethnic populations*. Virginia: ERIC Document Number ED 391485.

Schriver, K.A. (1997). *Dynamics in document design.* New York: John Wiley & Sons, Inc.

Sharrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the National Curriculum Mathematics Tests at Key Stage 2. *Educational Research, 41* (2), 123–136.

Silver, A.A. (1994). Biology of specific (developmental) learning disabilities. In N.J. Ellsworth, C.N. Hedley, & A.N. Barratta, (Eds.), *Literacy: A redefinition*. New Jersey: Erlbaum Associates.

Smith, J.M., & McCombs, M.E. (1971). Research in brief: The graphics of prose. *Visible Language, 5* (4), 365–369.

Szabo, M., & Kanuka, H. (1998). Effects of violating screen design principles of balance, unity, and focus on recall learning, study time, and completion rates. *Journal of Educational Multimedia and Hypermedia, 8* (1), 23–42.

Thompson, D.R. (1991). *Reading print media: The effects of justification and column rule on memory.* Paper presented at the Southwest Symposium, Southwest Education Council for Journalism and Mass Communication, Corpus Christi, TX. (ERIC Document Number 337 749)

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., Johnstone, C.J., & Miller, N.A. (in press). *Universally designed assessments from the end user's perspective: Using a think aloud method* (Policy Directions). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study. *Exceptional Children, 64* (4), 439–450.

Tinker, M.A. (1963). *Legibility of print.* Ames, IA: Iowa State University Press.

Trotter, A. (2001). Testing computerized exams. *Education Week, 20* (37), 30–35.

West, T.G. (1997). *In the mind's eye: Visual thinkers, gifted people with dyslexia and other learning difficulties, computer images, and the ironies of creativity.* Amherst, NY: Prometheus Books.

Worden, E. (1991). *Ergonomics and literacy: More in common than you think.* Indiana. (ERIC Document Number 329 901)

Zachrisson, G. (1965). *Studies in the legibility of printed text*. Stockholm: Almqvist and Wiskell.

Understanding and Using Data

The true test of whether the time, effort, and money invested in testing paid off

Ultimately, the extent to which an assessment provides the needed information to make a decision or take some action determines its value. If educators follow informed assessment practices, the assessments they develop should provide the information they seek. The next essential steps are to understand what the test results are indicating and to determine what to do next.

Just as understanding a problem represents 90% of the solution, so is understanding assessment results crucial for their proper use. Unfortunately, this is the step where far too many educators and other stakeholders—from parents to policy makers—fall short. The reasons stem from many of the factors highlighted in the previous sections. In some cases, people don't understand some of the basic measurement principles and their implications, perhaps most commonly "comparability"—the extent to which the results from one test can be legitimately compared to results from another. In other cases, many don't fully understand how purpose drives test content, so that an assessment designed for one purpose will probably not be useful for another purpose. So, for example, a general achievement computer-adaptive test might efficiently produce a total test score indicating the level of overall student learning, but it can't produce accurate diagnostic insights. Addressing these misunderstandings simply requires some professional development, coaching, and other reinforcement.

Assuming users of assessment results understand what they mean, the right conditions must exist in a school or district to enable them to put that understanding to good use. Creating a data-informed culture takes commitment, expertise, and persistence; but ample evidence exists that it can make a tremendous difference in fostering student (and staff) success. This final step completes the cycle of Informed Assessment Practices, and the benefits will show up in increased staff effectiveness and efficiency and improved student outcomes.

The following examples illustrate common mistakes—opportunities where assessment literacy can help educators do more with less:

Vignette: A parent doesn't understand why her child shouldn't skip grades since the child scored two grades above grade level on a vertically equated NRT.

Vignette: A state board of education member doesn't understand why 70 percent correct isn't the appropriate cut score for proficiency in a statewide accountability test, since 70 percent was what the minimum passing grade was when he was in school.

Vignette: A university reading educator notices that a state's reading test and an off-the-shelf NRT produce very different percents of "proficient" students, and concludes the two tests are uncorrelated and something is wrong with the state test.

Vignette: A U.S. Congressional representative believes that state testing results (percent proficient students) that differ from NAEP state results are a serious problem and that it would be solved by the states using national content standards.

Vignette: Despite his state, with its large urban population, moving to number one on NAEP, a soon-to-be-elected governor indicated he wanted to discontinue the highly regarded state assessment program because it hadn't eliminated the achievement gap.

ISSUE PAPER

Measurement Error, Human Error, and Decisions Based on a Test

by Stuart R. Kahl, Ph.D.

measured progress

Number 2

October 2003

Issues associated with the use of test results are not new. However, the implementation of individual states' high-stakes testing programs and then the No Child Left Behind Act of 2001 (NCLB) have made the appropriate use of testing even more important. Media coverage over the past few years, most recently a Discovery Times documentary entitled, "Making the Grade" and an accompanying front-page piece in the *New York Times,* have highlighted human errors in statewide assessment programs. Yet, few, if any, articles in the mainstream media have been written about the impact of measurement error and its implications in high-stakes decision-making based upon test scores.

While "Making the Grade" provided a valuable overview of many issues facing states and assessment providers in the wake of NCLB, no single newspaper article or one-hour television program could adequately cover the nature and implications of error in large-scale assessment. So, I thought it was time for me to expand on this subject a bit further.

The nation's concern about testing errors is legitimate, and it is important that they be eliminated. Yet the almost myopic focus on the implications of human errors detracts attention from a far more fundamental problem—the inappropriate use of tests in making high-stakes decisions about students. Assessment experts, critics of testing, and educators have agreed for years on this matter. Indeed, testing manuals and long-established professional standards make it very clear that it is inappropriate to base important decisions about students or groups of students solely on the results of a single test. This admonition is important even in the absence of the human errors that attract so much attention for the following reasons: measurement error and validity issues.

Measurement Error

The following only scratches the surface of measurement error. For the curious and mathematically inclined reader, I recommend another paper, which explains more about the theory behind and the computation of the standard error of measurement and related issues. Entitled, "Standard Error of Measurement," it was written by Leo Harvill as an instructional module published by the National Council on Measurement in Education. By arrangement with the publisher, we are making this piece available.

A test score estimates something—a student's mathematical proficiency, perhaps. It is an estimate because it is based on a small sampling of the universe of items that could have been included on the test. Further, a test score is affected by factors other than the student's mathematical proficiency, such as: how well or motivated the student feels, whether there were distractions or interruptions during the testing session, and whether the student made good or bad guesses, to name a few. These factors, which can all be sources of measurement error, explain the difference between a student's calculated score on a particular test and that student's hypothetical "true" score. That true score, forever unknown, would reflect the student's real level of proficiency.

Furthermore, the magnitude and significance of measurement error is rarely a consideration of test-score consumers.

Measurement error and, more specifically, standard error of measurement are complex measurement concepts, which assessment providers tend to oversimplify in explanations to the public. To account for measurement error, the industry reports a student's test score with a band of plus or minus a few points and suggests that the "truth" regarding the student's proficiency almost certainly lies within that band. However, this is not really true. Furthermore, the magnitude and significance of measurement error is rarely a consideration of test-score consumers.

Score Distributions

Unless a test is exceptionally hard or easy, more students will fall in the middle range of the score scale, with fewer students near the ends. Measurement experts don't force this; it's reality. Key statistics used to describe the distribution of scores on a test include the mean (average) score and the variance of the scores, or the square root of the variance (standard deviation), which reflects how spread out the score distribution is.

For a variety of reasons, raw test scores (the number of items answered correctly or the number of points earned) are often transformed to a convenient scale. For example, some intelligence tests have a mean of 100 and a standard deviation of 15. Initially, the Scholastic Aptitude Test (SAT) had a mean of 500 and a standard deviation of 100, and the American College Test (ACT) had a mean of 25 and a standard deviation of 5. When those scales were established, approximately two-thirds of the students' scores fell within one standard deviation of either side of the mean and approximately 95 percent fell within two standard deviations. This means, for example, that two-thirds of the students who took the SAT scored between 400 and 600, and 95 percent scored between 300 and 700.

Reported Test Scores and Standard Error of Measurement

As noted earlier, reported student test results often include a specific score followed by "+/-" a certain number of points. That number is the standard error of measurement. For many tests with acceptable levels of reliability, the standard error of measurement is approximately 0.3 to 0.5 times the standard deviation of the test. The standard error of measurement for the SAT is more than 30 points; for the ACT, it would be approximately 2 points. *(The Harvill piece I cited earlier provides a useful explanation of the concepts and calculation of reliability and standard error of measurement.)*

It is misleading to suggest that a score reflecting the student's "true" level of proficiency almost certainly lies within the interval defined by the reported score plus or minus the standard error. On the other hand, if we placed that interval around the student's hypothetical "true" score, and if the student was tested many, many times (with equivalent test forms), then indeed two-thirds of his or her scores would fall within that interval. One-third of the scores would lie outside of the interval. *(Note that the standard deviation described earlier refers to the variability in test scores of different students. The standard error of measurement has two interpretations. It is the standard deviation of differences between students' hypothetical true scores and their obtained scores. But it also has the meaning expressed above related to the hypothetical situation of a student being tested many times. This concept is further explained in the Harvill piece.)*

BUT THE FACT IS WE DON'T KNOW ANY STUDENT'S TRUE SCORE. Thus, in student score reports, we put the interval around the student's obtained test score. In this case, the plus-or-minus-one-standard-error interval actually means, "If we were to test the student many, many times, and properly construct a standard-error interval around each obtained score, two-thirds of those intervals would contain the student's true score." Sorry, but that's the way it is. Two-thirds is not terribly impressive if you're trying to measure the student's "true" proficiency, and it is even less so the larger the standard error.

Is it any wonder that testing experts advise against basing important decisions about students on scores from a single test?

To get to 95 percent of the intervals capturing the hypothetical true score, we'd have to go with plus or minus two standard errors. Can you imagine reported SAT scores having +/– 60 points after them? The only ways to appreciably improve our estimates of a student's "true" proficiency score are to use many repeated measures or to give an *extremely* long test (one with impossibly high reliability). Is it any wonder that testing experts advise against basing important decisions about students on scores from a single test? More on this later.

Validity Issues

Besides measurement error, the other reason to avoid the misuse of single test scores involves test validity. In the preceding sections, we barely scratched the surface of measurement error. Test validity can be equally as complex, but our discussion of it here will be just enough to explain some key issues.

High reliability is a characteristic of a good, consistent test of something. Although there are many kinds of test validity, at the root of them all is just what that *something* is. Is the test a good measure of what it is supposed to measure? If not, then conclusions drawn from the test are not justifiable. As an example, the rapid turnaround of results required to meet NCLB accountability requirements has led some states to use only multiple-choice tests. Critics of these tests argue that such

measures are indirect, or surrogate, measures of the skills of interest and therefore, are not particularly valid.

Asking students to draw conclusions is quite different from asking them to pick a correct answer from a list. Solving an equation is a higher-level task than identifying the solution to an equation by simply substituting solution options into the equation. Going beyond the matter of item types, universal design experts point out that students demonstrate what they know and can do in different ways and that a traditional paper-and-pencil test is not the best way for some students. These are the kinds of validity issues that challenge the use of single-test results in making important decisions about students' achievement.

A "Single" Test Score

As obvious as it may seem on the surface, the meaning of a "single" test score is not so clear-cut. Programs requiring students to pass a test in order to graduate from high school often allow students to be retested several times between the spring of tenth grade and spring of the senior year. This practice clearly addresses the concern of a student scoring below the minimum passing score because of measurement error. The chances of erroneously failing several times are slim. However, for a student who fails a test several times before finally passing, that passing score may be in error. Nevertheless, the student is given the benefit of the doubt in such a situation. (I should note that some score-based decisions—such as those involving advancement to the next grade or required summer school—might not offer these retesting opportunities.)

The designers of such graduation testing programs argue that multiple measures are involved in this high-stakes decision. Critics, on the other hand, argue that equivalent forms of a test used for retesting do not constitute multiple measures. While retesting attempts to address measurement error, it ignores validity issues. The intent of requiring multiple measures is to use different types of tests to allow for student differences in the ways they best demonstrate their capabilities. If a testing program with high stakes for individual students has an appeals process, then it may be argued that alternative forms of evidence can be introduced into the decision process, even if there is no alternative form of *the* test.

The tests mandated by NCLB must employ multiple measures. However, that mandate has been interpreted as meaning different item types within a test (e.g., multiple-choice and constructed-response). Again, critics argue that even a test with multiple item types is still a traditional paper-and-pencil test and that multiple measures would more appropriately include performance demonstrations, portfolios, etc. (which represent truly different means of measuring student achievement). Actually, the U.S. Department of Education has not strictly enforced the requirement of multiple measures, even in terms of item types within a test.

> . . . a student scoring just above a cut score and a student scoring just below the same cut score are virtually indistinguishable from one another . . .

The document, "Standards for Educational and Psychological Testing" (1999), prepared jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), clearly distinguishes between multiple opportunities to succeed on equivalent test forms and construct-equivalent testing alternatives. Standard 13.6 calls for providing students with both types of opportunities.

The Realities of Cut Scores

A cut score is the minimum score a student must attain on a test to be classified in the upper of two performance groups (e.g., the score separating students who are "proficient" from those who are "partially proficient"). In the case of graduation, advancement, or summer-school decisions, performance relative to a cut score may mean the difference between passing or failing, the latter having significant consequences. *The harsh reality of cut scores, given what we know about measurement error, is that a student scoring just above a cut score and a student scoring just below the same cut score are virtually indistinguishable from one another in terms of their proficiency based on the evidence from the test.* In fact, there is close to a fifty-fifty chance the positions of their "true" scores

are reversed. Again, this is why more information is needed to make important decisions about these students.

Human Error

As has been pointed out in numerous news stories, documentaries, and studies, many types of human error can lead to erroneous test scores and wrong decisions about students. These errors range from faulty test questions and answer documents to myriad mistakes that can occur during the numerous steps preceding the reporting of test scores. One type of "error" that might not even be considered an error in some circumstances involves the points awarded by human readers to student responses to constructed-response questions. Errors in this scoring process do, in fact, contribute to measurement error and are taken into account in determining total test reliability.

Multiple-choice items have associated with them a form of scoring error, also contributing to measurement error. A student who guesses correctly on a multiple-choice question gets full credit for the response, the same as a student who really knew the answer. Indeed, we encourage students to use partial guessing on this type of question, telling them to eliminate obviously wrong answer options and guess an answer from the remaining ones. Students who guess correctly get one point, the same as students who knew the correct answer. Those who guess incorrectly get zero points, the same as students who knew less and were not able to identify obviously wrong answers.

Assuming the correct answer key is used by the computer to score the multiple-choice responses, nobody questions that scoring. Regarding the responses scored by humans, however, particularly in a high-stakes situation, a one-point scoring error not in favor of the student is not as acceptable, even if the bottom-line reliability of the constructed-response or mixed-format test is equal to or greater than that of an all-multiple-choice test.

Many of the contractor errors on large-scale, high-stakes assessments that have been publicized over the past couple of years were at the same time miniscule and colossal. For example, the thousands of New York City students erroneously sent to summer school initially received scores placing them at the fifteenth percentile, below the cut score. Their corrected scores placed them at the sixteenth percentile, above the cut score. Given what we know about measurement error, students with scores corresponding to the fifteenth and sixteenth percentiles are virtually indistinguishable, based on test results alone.

In fact, in many cases in which contractor errors led to serious consequences for students, schools, and the companies themselves, the impact of the errors on actual scores was extremely small. In several instances, miscalculations, equating errors, and the like resulted in errors in final test scores that equaled a fraction of the standard error of measurement. This in no way excuses the errors. These types of errors are unacceptable, and far greater effort must be made to avoid them. My message regarding these errors is simple: for errors of such small magnitude to lead to such severe consequences is ironic, at best, and tragic, at worst, given that measurement error, which is far greater, is virtually ignored in interpreting scores, as well as in setting and using passing scores. What the testing companies' human errors do is make the inappropriateness of basing high-stakes decisions on a single test score even more obvious. But in debates about the wisdom of using test scores in making decisions, measurement error is the real concern.

Generally, the key to avoiding the over-interpretation and overuse of a test score involves using additional evidence of student achievement.

Alternatives to the Use of Single Tests in High-Stakes Decisions

Generally, the key to avoiding the over-interpretation and overuse of a test score involves using additional evidence of student achievement. For example, if a test score places a student below a particular group of his or her peers, but school grades place that student above that same group in terms of achievement, consistently throughout the year and regardless of the scale of school grades, the latter placement is more justified. For selection or exclusion decisions, a test might be used as an initial screening device to over-identify

candidates for acceptance or rejection. Follow-up investigation of additional information, including teacher judgments, can shorten the lists. Standard 13.7 in the AERA/APA/NCME standards document strongly recommends this approach to improving the validity of such decisions. Clearly, opportunities for retesting, alternative testing approaches, and appeals processes are advisable.

I did not write this paper to argue against accountability, or even against high-stakes testing. Rather, I hope it represents a good argument in favor of using different kinds of evidence of student performance when we make high-stakes decisions. Additional information acknowledges the importance of measurement error that is inherent in any test result, reduces the impact of human error—which, despite the best of intentions and efforts, may occur—and leads to better, more valid decisions for our students.

Stuart Kahl is president and chief executive officer of Measured Progress. Founded twenty years ago as Advanced Systems in Measurement and Evaluation, Measured Progress provides highly customized, standards-based assessments, alternate assessments, and professional development tools and services to clients across the nation.

100 Education Way, Dover, NH 03820
800.431.8901
www.measuredprogress.org

Turning data into knowledge

Data-based decision-making is an important tool for educational improvement. Yet, WCER researchers Bill Clune and Norman Webb remind us that making effective use of data is one of the continuing challenges of building capacity in systemic reform. In order to use data as a guide for continuous improvement around coherent goals of student achievement, schools must develop their organizational, technical and analytical capacities.

Effective use of school data requires planning and persistence. Data development and use must become an active part of school planning and improvement processes, and it must become infused and accepted into the school culture and organization.

With funding from the Joyce Foundation, Clune and Webb are working with school and district staff in the Milwaukee Public Schools (MPS) to develop the capacity of staff to understand and apply data strategically, using QSP (Quality School Portfolio) software, an analytical and reporting tool developed at UCLA by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Once fully integrated into a school's systems, data can be transformed from mere numbers to useful information, and can then contribute to schools' and district knowledge in effective and meaningful ways.

Milwaukee educators involved in the project consistently mention the need to develop a process and the skills to analyze and use data for answering questions, problem-solving, monitoring and decision-making. During the last two years, WCER researchers and the school staff have collaborated in developing analytical models and decision-making processes, collecting and managing data, and analyzing, reporting and applying data.

Clune and Webb have found that using data to support inquiry, to inform schools' instructional mission and continuous improvement requires coordinated changes in many areas- school processes and culture, use of technological tools, data collection and management, and the analytical skills and abilities of school personnel.

The school teams have found that making the needed changes to implement data-based decision making poses difficult, but not insurmountable challenges. With hard work, each school has made progress toward meeting many of the challenges.

For example, one school analyzed multiple measures of individual student data generated at the school site to evaluate or monitor student performance changes in reading, programs, student placement, etc. After reviewing the information, the principal and teacher leadership team decided on a course of action. For the subsequent school year, they reallocated school resources

in reading, identified low performing students to receive additional reading resources, and hired two new reading specialists. Team members are tracking these interventions this year to see whether the students' reading performance improved.

Another school began to analyze what staff called "event-based" data-that is, data that refer to specific incidents or actions, rather than to test scores or student demographic variables.

Tracking the pattern of events such as discipline referrals and attendance infractions, the principal provided summary information to teacher teams. School staff used these summaries discuss student behavior and related teacher behavior management practices. The data were also used to support decisions about resource allocation. The school hired an additional counselor for the following school year to help students who have difficulties in their lives outside of school, which staff concluded affect student behavior.

Building capacity

Clune and Webb identified six challenges schools face as they build capacity for using data-based decision making:

• Cultivating the desire to transform data into knowledge

• Focusing on a process for planned data use

• Committing to the acquisition of data

• Organizing data management

• Developing analytical capacity

• Strategically applying information and results

Strong leaders who support the local use of data help establish a school culture that not only accepts the use of data but considers data as a source of information that contributes to problem solving and knowledge building. Whether key staff or the school principal provide the leadership, it is essential for a school to gather support, obtain commitment, allocate resources, and identify goals to ensure that its data efforts are a success.

It is important for schools to link their use of data to their school planning and decision-making processes. Such a focused approach saves time and effort and allows for more efficient use of limited data. An approach that aligns data inquiry to school planning and decision-making

processes right from the start is more likely to produce answers to specific questions, evidence to support school goals, and information that sheds light on identified problems. Planned and targeted data inquiry helps to keep data analysis on track, as well as ensure that information is fed back into the planning process and that key decision-makers get timely answers to their questions.

Data do not magically appear, ready-made, to provide evidence of success and solve all of a school's problems. School staffs struggled to build the internal will, capacity, and organization to make data work for them. They had to learn where to get data, how to manage data, how to ask good questions of the data, how to analyze the data accurately, and how to apply data results appropriately, and ethically.

The final challenge for schools is to learn how to appropriately apply results and make purposeful and ethical use of information for improving teaching and learning. Appropriate and ethical use of data implies that a school has taken the necessary precautions and steps to ensure that data is accurate, valid, and reliable and that the analytical process is complete, equitable, and fair. If schools have followed a continuous improvement process for planning and decision-making, the results will be easily linked back to specific questions, goals, and problems. By focusing the data analysis to target specific issues, schools are poised at the end of the analytical process to make sense of and draw meaning from results. The final step is to share the new information and results with staff to inform school planning and decisions. The results can be used in a variety of ways-to identify progress, explore problems, and target strategies for change, to mention a few. In this manner, schools successfully transform data into information and apply that information to school improvement.

Raising Questions or Providing Answers: Effective Use of Interim and Benchmark Assessments

Stuart Kahl, Ph.D.
Founding Principal

The new academic year finds education experts and other stakeholders still trying to sort out which tests serve what purpose. I'd like to enter the fray by expanding on the ideas I expressed here last year. Unfortunately, the formative versus summative distinction continues to be confusing. For all I've said about the mischaracterizations in this area, I'd like to clarify how tests that don't yield formative information about individual students can play a positive, formative role in program evaluation and improvement. But first, I'll take a step backward.

Accept for now that "formative assessment" refers to any of a number of assessment activities that teachers and students engage in day after day to gauge students' understanding of content while it is being taught and to help inform instructional decisions pertaining to that content.

Once a class moves on to new content, assessments of the previously taught material are no longer formative. At that time, they provide evidence of the extent of student learning about the specified content or skills. We're talking about end-of-unit,-marking period, and -semester tests, all of which typically contribute to students' grades. State assessments, off-the-shelf norm-referenced tests, and other general achievement measures also fit in the category of "summative assessment." Some of these assessments, administered throughout the year, are often called "interim" or "benchmark" assessments.

Whether interim or benchmark, summative assessments cover a whole year's content or some fraction thereof. They generally apply a shotgun approach to the coverage of the subject area domain or subdomain; that is, they provide breadth of coverage of a large body of content—not depth of coverage of any major concept or narrowly defined set of skills that would be part of a discrete lesson or unit of study. Thus, because of the timing of these tests and given that they are not diagnostic at the individual student level, they could only be used formatively to evaluate and improve instructional programs, most likely for the benefit of the next class of students to pass through a grade. (These tests may also provide an "early warning" to identify students at risk with respect to passing some later high-stakes test; however, further investigation would probably be necessary to pinpoint the students' specific areas of weakness. If these areas include content addressed by instruction earlier in the school year, one is in a remedial mode with respect to these students.)

Looking at a group of students, average total test scores, subtest scores, and subgroup scores that interim or benchmark assessments yield might help answer "how are we doing" questions. But these data probably do a better job of raising questions than answering them. Such questions might be "Why is our performance in this subdomain so much lower than expected?" or "Why did these two subgroups perform so differently?" I have seen such questions arise in a number of schools.

The staff at a middle school was surprised by a geometry subtest score that was relatively low compared to scores in other math strands. An examination of the test questions and curriculum materials, as well as a conversation with the math teacher, made it clear that the content had not yet been taught but was scheduled to be taught later in the school year. Conclusion: no problem.

An elementary school administrator wondered why reading comprehension levels for informational passages were so much lower than those for literary passages in a particular grade. The administrator found that one teacher relied almost exclusively on literary passages in her class. Furthermore, she was not familiar with how reading strategies and skills differ for the two types of material. Solution: professional development for the teacher, coupled with assurance that greater attention be devoted to reading informational texts.

Finally, the gender differences in math scores in another elementary school were much greater than the statewide gender differences, which had all but disappeared over the years. Conversations with students, parents, and teachers revealed that two teachers were communicating lower expectations to their female students and doing less to promote higher achievement for girls than for boys. Solution: providing research-based information on causes of gender differences in math to the teachers, coupled with conscientious efforts to change and monitor the teachers' behavior.

I fear that interim, benchmark, and other summative test results aren't used to raise such questions nearly enough. If we only use these tests for accountability purposes, a singular opportunity for genuine program improvement is being squandered, as is the chance to get more out of assessment dollars. Teachers and administrators should, through training and practice, learn to identify the important questions that subtest and subgroup results raise—and where to search for their answers.

What do you think?
Let us know at twocents@measuredprogress.org

Turning data into achievement

Although No Child Left Behind (NCLB) has its many faults, one important outgrowth of the law has been a focus on using student assessment data to drive instruction. This focus has continued under the Obama administration, which has doled out millions of dollars in federal funding to encourage states and school districts to adopt such practices.

Yet while most schools now measure students' progress more frequently throughout the school year, using this information to target instruction more effectively can still be a challenge. As recently as last year, an Education Department report noted that states and school systems were making significant progress in building educational data systems—but school leaders still were searching for examples of how best to connect student data to instructional practices.

In this report, we'll look at how a handful of K-12 schools and districts have taken this critical next step of turning data into achievement. The schools we surveyed for this report range from suburban Seattle, Wash., to rural Alabama—but all share some key characteristics that have contributed to their success, such as a recognition that turning data into achievement involves changing the entire school culture and can't be done without intensive training and support.

Another key to their success has been the use of a new breed of data platforms that tie together the entire instructional cycle.

School leaders have been asking for end-to-end solutions that bring together the tools used for instruction, assessment, data review, and professional development, with the ultimate goal of putting real-time data into the hands of teachers and administrators so they can make informed instructional decisions.

Schools have always had data, but typically this information has been all over the place, housed in disparate systems that don't necessarily talk to one another.

"Up until now, one of the challenges that customers have had is that they have all of these different assessment pieces in different locations, and they get the data … in different ways," said Deborah L. Smith, director of product and portfolio management for Riverside Publishing, a division of Houghton Mifflin Harcourt.

Education vendors are responding with solutions that tie together curriculum, instructional tools, professional development, and parent outreach—all in a seamless manner.

“We’re moving to an age now where we can do all types of assessment through one single system, and all that [information] comes together in one place,” Smith said. And the results can be immediate, giving teachers longitudinal data for each student in real time, at their fingertips.

The idea of using data to inform instruction is not new, but it is now coming to fruition, thanks in part to these more robust solutions. And school districts nationwide are reaping the benefits: from getting off school improvement lists, to increasing graduation rates and making teachers’ jobs easier.

The advantages are enormous. Educational processes are faster and more efficient, because administrators can see everything about a single child or classroom with one login. Not only are these systems pulling together the data in one spot; they also can prescribe remedial content based on assessment data, which leads to more personalized instruction.

This is a great new trend, said Ann Ware, project director for the Consortium for School Networking’s Data-Driven Decision Making initiative.

She cited the work of Project RED: Revolutionizing Education, which surveyed K-12 districts nationwide and found that one of the most effective uses of technology was in delivering targeted intervention to students. “Providing inventions and using technology to do so was the top factor in improving student performance,” Ware said.

Restructuring the school day makes a big difference

To improve student achievement, a Michigan school has restructured its school day to make time for teachers to analyze assessment data, discuss the data with their colleagues, and offer students remediation or enrichment accordingly.

“We use data whenever we can, and I think part of the reason we’ve been successful is because we found ways to build time in our school day to give staff an opportunity to look at data and to collaborate, but also time in the day for students to take enrichment classes,” said Douglas Langmeyer, principal of Ring Lardner Middle School, which serves roughly 550 students in grades seven and eight.

Two to three hours per month, grade-level teaching teams, called Data Teams, meet to discuss assessment results and share strategies. Meanwhile, students head to the computer lab for enrichment classes. They complete self-paced instruction that is prescribed based on recent assessment results, either for remediation or acceleration.

“Building those things into the day has been a huge part of what we’ve been able to accomplish. It’s required us to do some creative scheduling, but doing these things during the day instead of having staff stay late in the afternoon or asking kids to come in after school so we can get to it—that’s part of the secret to our success,” Langmeyer said.

During Data Team meetings, staff look at pre- and post-test data from regular classroom instruction. They are collaborating, asking: What do students already know? What areas did they miss? How should we design our instruction for the following week? What works? What doesn’t?

“The nice thing about the Data Team process is that it’s very immediate. If I’ve given a post-test and I see that, ‘Wow, I’ve really missed the mark in this area,’ I can immediately go back and readdress that,” Langmeyer said.

He added: “Sometimes with standardized test data … it’s almost like it’s an autopsy. It tells you what you did wrong or what isn’t working, [but] it’s so far down the road that you don’t really have an opportunity to reteach.”

Teachers at Ring Lardner also give common assessments quarterly, to see what kids know from nine weeks’ worth of instruction. They use DataDirector and other tools from Riverside Publishing to give these tests and analyze the results.

“That’s obviously very helpful,” Langmeyer said. “If we taught this particular concept three weeks into the marking period, [and] I find out [that students] still don’t have that knowledge, I can put them in this enrichment program that I spoke of.”

Currently, Ring Lardner is using DataDirector only for its common assessments, but school officials hope to move all their assessments to the system eventually.

DataDirector captures student assessment data within all levels of instruction. All of the information resides in one place, including state-administered assessments, district benchmarks, and even classroom tests. It also has test-creation capabilities, so teachers can create an online test or an answer sheet for a paper test that is scanned back into the system. Teachers can use their own questions or ones from Riverside Publishing’s Assess2Know item bank.

DataDirector also includes a suite of pre-built reports “that are immediately available to report out student performance [in comparison] to state or Common Core standards … where those alignments exist,” said Karen Burkhart, director of product management for Riverside Publishing’s formative assessment business unit.

The software lets users look at the data in many ways, depending on the questions they might have. “You can break it down by demographics. Any information that is tied to a student ID can be pulled into DataDirector. That enables users to have a really rich data set they can work with,” Burkhart said.

What’s more, DataDirector now has the ability to link directly to the instructional resources a school already subscribes to, such as netTrekker or Destination Reading, to offer the remediation students need based on their test results.

“We are really trying to bridge the gap between assessment and instruction,” Burkhart said. “Making those links available directly within the assessment report is a great step in making it easier for teachers and administrators to quickly identify the standards that need more instruction, and what resources are available to them within the instructional resource program they subscribe to.”

Using assessment data to plan instruction does work. After seeing low writing scores, Ring Lardner officials devised a building-wide initiative to stress writing skills. They trained staff, adopted a structured writing piece, and had kids writing across all curriculum areas, including physical education.

“We’ve seen some improvements in writing. We are still not where we want to be. We have some areas we will continue to focus on, but there’s been some improvement,” Langmeyer said.

In math, on the other hand, the school’s average is 10 percentage points above the state average for proficiency on the MEAP, Michigan’s state test. “We have a high at-risk population, and when you consider our demographics…we’ve seen some real improvements. It’s been continual growth over years—not one year up, one year down,” Langmeyer said.

Ring Lardner has achieved these gains despite a student population in which nearly 60 percent qualify for free or reduced-price lunches.

“It’s definitely the use of data. It’s collaboration. It’s also the fact that we’ve got a strong teaching staff. We’ve got good people in place,” Langmeyer said. “Parental involvement, socioeconomic factors—those things all take a backseat to having wonderful teachers in the classroom.”

A web-based, end-to-end system

Using GlobalScholar’s Pinnacle Suite as its platform, Washington’s Federal Way Public Schools is creating its own web-based, end-to-end system that integrates student information, curriculum,

policy, assessment, and professional development. And the real-time assessment data this system makes possible is enabling teachers to differentiate instruction for each student.

“Part of the challenge is that we have a diverse student population with wide-ranging needs, and if we are going to be serious about educating all children, then we have to have … the infrastructure to generate the information that’s going to inform daily decisions on instruction,” said Robert Neu, superintendent of Federal Way Public Schools, which is 22 miles south of Seattle.

That’s where Pinnacle comes in. The software lets teachers extract information and build profiles on individual students, so that teachers can retrieve historical as well as current information about each child. “I can’t tell you how important that is. Without it, you can’t take differentiated instruction to the level that we need to,” Neu said.

Additionally, the district has developed a series of common formative and summative assessments. After teachers give a test, teacher teams get together to discuss the results, share ideas, and form strategies to meet individual students’ needs.

“Without that infrastructure, … we weren’t able to meet students’ needs in real time; we weren’t able to have those best-practice conversations in real time, and then it serves no point because the data [are] basically too late,” Neu said.

The district is also redesigning its standards system. Starting in fall 2011, all K-12 courses will be aligned, vertically and horizontally, by power standards. There will be no more than 15 power standards in each class and two to three learning targets per power standard, Neu said.

A power standard is the essential component of the curriculum that each child has to master to pass the class and move on to the next level of instruction, Neu said. Each power standard has to be something that will endure beyond the test; it’s got to carry on to the next level of instruction in that content area, and it has to transfer into other areas of the curriculum.

“To take it to the next level—and this is the part that gets really exciting for us—students have to perform mastery on these learning targets to pass a power standard, and they have to do it twice,” Neu said.

This new system will necessitate individualized instruction. “Some students take longer to … master the standards, [and] those who don’t [will] continue to work on it,” Neu said. “We’ve got to be able to collect that [information] on each child to inform daily instruction.”

Federal Way has outperformed state and national averages, despite its student diversity—making it one of the top 10 school districts in the state of Washington, Neu said. He expects to see even further student growth as the district realigns its instruction around the power standards.

Pinnacle's ease of use has played a role in the district's success. When it's easy to enter and retrieve data, teachers can make decisions that have more impact, Neu said.

Kal Raman, CEO of GlobalScholar, says the integration of data is a real problem for America's school districts.

"It's [often] too much effort to convert [data] into actionable information," Raman said. "Either the data get collected all throughout the day and people see [the information] after it's too late, or the data get collected without proper statistical modelling or applying some collaborative filtering. It's not actionable even when it's real-time."

Individual coaching pays dividends

An Alabama school district got one of its schools off the school improvement list and increased its graduation rate to 96 percent by using assessment data to inform instruction.

With data analysis and coaching provided by Software Technology Inc. (STI), the school turned itself around within two years, said Dave Sewell, technology specialist for Houston County Schools in southeastern Alabama.

"The school gives [STI] credit, but actually, it was the hard work of everyone put together. STIAchievement data showed where they needed to focus their efforts," Sewell said.

Now, all nine schools in Houston County use STI services to drive instruction. "They saw the success at the first school, and they were eager to copy that success," Sewell said of the other school leaders. The rural district, in which 76 percent of students qualify for free or reduced-price lunches, has seen improvement in all of its schools that have been working with STI for more than a year, he added.

STIAssessment lets educators do benchmark testing throughout the year, in four subject areas from third grade through 12th grade. The tests can be taken online for immediate results, or the answer sheets can be scanned back into the system for marking. STIAcheivement lets users pull a myriad of reports based on different criteria.

STI pairs its suite of software tools with in-person professional development and coaching. The company, which started in student information systems 26 years ago, has professional coaches—

mostly former school improvement specialists—who travel on-site to a school or district to design a custom program.

First, the coaches analyze data collected through STI's formative assessment program, benchmark testing, and even high-stakes testing.

"We normally ask for three to four years' worth of data from the client before we get started, and throughout the engagement we'll [hold] updated progress meetings," said Jenna Wood, marketing director for STI.

In these data meetings, the coaches explain the results in terms of strengths, weakness, and goals.

"During the data meetings, we will break it down for each teacher, for each grade level, and for each subject area, for the teacher to see first-hand," Wood said. "It's kind of different hearing it verses actually seeing the results. We work with the teachers individually and [explain] what their actual data show."

After the progress meetings, the coaches work with teachers to make changes that would help strengthen students' knowledge. "We can nail down what the issues are and how to resolve those and improve test scores," Wood said.

The coaches help develop pacing guides, change instructional strategies, change school culture, or develop individual student lesson plans—whatever is needed based on the data analysis.

"We will build a plan for them based on what we know from the data three to four years ago and where we stand today. And we will work together in partnership to raise those weaknesses and maintain those strengths," Wood said.

She continued: "Ours is not a cookie cutter-type service, where one model fits all. We actually customize every plan based on each individual school. And then we work our plan around that, and who our coaches need to be, and we determine how many days we would suggest for this type of service."

Not only do the coaches analyze data and prescribe remediation; they also focus on creating a partnership, getting teacher buy-in, and raising moral, Wood said, adding: "They become part of the district. They get down and dirty with the teachers and assist them in any way possible."

Benchmarking once again is key

Like the other districts we talked to, North Carolina's Granville County Schools attributes its success in turning data into achievement to teacher training and support, frequent benchmarking, and regular meetings to discuss the data. Granville County also uses longitudinal data to predict a student's future performance on state exams.

"Our high schools have noticed a huge jump in scores, and our elementary schools, grades three to eight, have had a 14-percent increase," said Superintendent Timothy J. Farley.

The district started using its data analysis tools in earnest three years ago. For data analysis, Granville County uses Education Value-Added Assessment System (EVAAS) from SAS Institute. The system takes multiple years of data and predicts how a child will perform in the next grade level based on that prior information, and it also generates other data reports.

For benchmarking, the district uses the ClassScape Assessment System from North Carolina State University. "Not only do we know where we start from using EVAAS, we check along the way to make sure our teaching and learning are aligned," Farley said.

One of the biggest challenges was training principals and teachers to read, interpret, and use the data correctly. Now, teachers meet weekly to discuss the data and plan lessons and assessments. They also align the curriculum and assessments "to make sure what is taught is tested and what is tested is taught," Farley said.

Using data has been key in eliminating any bias a teacher might have in recommending a student for more advanced coursework. "In particular, we've used it to place more students in Algebra 1," Farley explained.

Not only does the district use data to inform instruction, but officials have done audits in every department—resulting in huge cost savings.

"We've managed to save a huge amount of money in using data to change our operations," Farley said. "We saved well over $1 million on our maintenance, transportation, and technology side of the house."

Bibliography

Assessment Literacy

Kahl, Stuart R., Ph.D. "What teachers as assessors must know and be able to do." AdvancED. Fall 2010.

Kahl, Stuart R., Ph.D. "Innovative assessment: Effective and ineffective assessment reform." 16th Annual Education and Law Conference, Portland, ME. 21 July. 2009.

Darling-Hammond, Linda. "A vision built on educational research and successful practices." *Performance Counts: Assessment systems that support high-quality learning*. Council of Chief State School Officers (CCSSO). 2010. All rights reserved.

Kahl, Stuart R., Ph.D. "A balanced assessment system: A different perspective." November. 2006.

Perie, Marianne; Marion, Scott; Gong, Brian. "Moving toward a comprehensive assessment system: A framework for considering interim assessments." *Educational Measurement: Issues and Practice*. National Center for the Improvement of Educational Assessment. Fall. 2009. Vol. 28, No. 3, pp. 5-13.

Kahl, Stuart R., Ph. D. "The assessment-literate school administrator." December. 2007.

Kahl, Stuart R., Ph.D. "You can't squeeze blood out of a turnip: What diagnostic testing is—and isn't." 2011.

Kahl, Stuart R., Ph.D. "Helping teachers make the connection between assessment and instruction." January 2006.

Kahl, Stuart R., Ph.D.; Sweeney, Kevin P., Ph.D. "Large-scale assessment: Choices and challenges." March. 2004.

Formative Assessment

Black, Paul; Wiliam, Dylan. "Inside the black box: Raising standards through classroom assessment." Copyright: Phi Delta Kappa International. October. 1998.

Kahl, Stuart R., Ph.D. "Where in the world are formative tests? Right under your nose!" [Revised] June. 2005.

Kahl, Stuart R., Ph.D. “Formative assessment and professional development: A questions of mind (set) over subject (matter).” December. 2008.

McManus, Sarah. “Attributes of effective formative assessment.” Copyright: Council of Chief State School Officers (CCSSO). 2008.

Kahl, Stuart R., Ph.D. “Are good grading practices like putting your thumb in your navel?” June. 2008.

Heritage, Margaret. “Formative assessment: What do teachers need to know and do?” Copyright: Phi Delta Kappa International. October. 2007.

“Technology takes formative assessment to a whole new level.” *eSchool Media, Inc*. 4 August. 2010

Alignment of Assessments

“Aligning assessments and standards.” Wisconsin Center for Education Research at the School of Education, University of Wisconsin-Madison. September. 2007.

“Maine leads once again with Common Core pilot.” *eSchool Media, Inc*. 18 April. 2011.

Hess, Karin K.; Carlock, Dennis; Jones, Ben; Walkup, John R. “What exactly do ‘fewer, clearer, and higher standards really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and implement assessments.” White paper available [online] http://www.nciea.org/publications/cognitiverigorpaper_KH11.pdf. 2009.

Hess, Karin K. Cognitive Rigor matrix for ELA and Math-Science. In *Local Assessment Toolkit: Exploring Cognitive Rigor***.** Available [online] http://www.nciea.org/cgi-bin/pubspage.cgi?sortby=pub_date. 2009.

Performance Assessment

Kahl, Stuart R., Ph.D. “Performance assessment: An idea whose time has come (again).” May. 2010.

Wood, George; Darling-Hammond, Linda; Neill, Monty; Roschewski, Pat. “Refocusing Accountability: Using local performance assessments to enhance teaching and learning for higher order skills.” *Briefing Paper Prepared for Members of the Congress of the United States*. 16 May. 2007.

Prabhu, Maya T., former assistant editor. "New test measures students' digital literacy." *eSchool Media, Inc*. 2 April. 2010.

Accessible Testing

"Feds to schools: Make sure ed-tech programs are accessible." *eSchool Media, Inc*. 26 May. 2011.

Russell, Michael. "Digital test delivery: Empowering accessible test design to increase test validity for all students." Arabella Advisors. © 2011 Bill & Melinda Gates Foundation.

"Using systematic item selection methods to improve universal design of assessments." University of Minnesota. National Center on Educational Outcomes, Minneapolis, MN. September. 2006.

Thompson, S.J.; Johnstone, C.J.; Anderson, M.E.; Miller, N.A. "Considerations for the development and review of universally designed assessments." *Technical Report 42*. University of Minnesota. National Center on Educational Outcomes, Minneapolis, MN. 2005.

Understanding and Using Data

Kahl, Stuart R., Ph.D. "Measurement error, human error, and decisions based on a test." October. 2003.

"Turning data into knowledge." Wisconsin Center for Education Research at the School of Education, University of Wisconsin, Madison. March. 2002.

Kahl, Stuart R., Ph.D. "Raising questions or providing answers: Effective use of interim and benchmark assessments." September. 2006.

Erenben, Cara, contributing editor. "eSN Special Report: Turning data into achievement." *eSchool Media, Inc*. 2 June. 2011.

About

STAR, School Technology Action Report, is a roundup of current news and information on a specific topic provided by the resources of ***eSchool Media, Inc.***, case studies, white papers, and industry reports and surveys.

eSchool Media, Inc. is a news and information organization delivering print, web, eMail, and video communications to nearly one million K-20 decision makers throughout North America and around the world. Its information networks provide education-technology content and services for leaders in schools and colleges and help educators successfully use technology and the internet to transform education and achieve their educational goals.

eSchool News is the flagship publication of ***eSchool Media, Inc.***, which also includes *eCampus News*, serving higher education exclusively, and *eClassroom News*, an electronic resource for teachers and other classroom practitioners.

Measured Progress is a not-for-profit company that has led the industry for almost 30 years in developing rich K-12 assessments designed to improve teaching and learning. It has provided customized and customizable products and services to the federal government, state and/or local education agencies in 46 states, foundations, and others, addressing the needs of general and special student populations. Currently working on cutting-edge education reforms, it is helping clients meet today's challenges—and tomorrow's.

eSchool Media, Inc.

7920 Norfolk Avenue, Suite 900
Bethesda, MD 20814
800-394-0115
301-913-0119 (fax)