At an orchestra concert, a music critic evaluates the performance on the basis of how well the musicians and the conductor interpret the notes and phrases the composer put to paper.

At an employee’s performance review, a supervisor evaluates the employee’s performance by how well the employee measured up to the expectations outlined in the job description.

In a classroom in a standards-based education system, the teacher evaluates student performance using assessments that align well with the established standards. Standards and assessments work together to guide the system toward students learning what they are expected to know and do.

But correspondence between state-level standards and assessments tends to be only moderate, particularly in terms of depth of knowledge and range of knowledge, according to a recent study by WCER researcher Norman Webb and colleagues at WCER’s National Institute for Science Education (NISE).
Webb and colleagues analyzed the alignment of assessments and standards in mathematics and science in grades 3 through 10 and in four states. Members of Webb’s 17-person team are affiliated with national-level organizations including the Council of Chief State Officers, the National Research Council, and the National Center for Improving Science Education. The team’s June 1999 report, jointly published by NISE and the Council of Chief State School Officers, identifies the match between standards and assessments using the following four criteria:

**Categorical concurrence**: To what degree do standards and assessments address the same content categories? This criterion is met if the same or consistent categories of content appear in both documents.

**Depth-of-knowledge consistency**: What degree of depth or complexity of knowledge do standards and assessments require? This criterion is met if the assessment is as demanding cognitively as the expectations standards set for students. Webb and colleagues judged depth of knowledge at four levels:
1. Recall of a fact, information, or procedure.
2. Skill in using information, conceptual knowledge, procedures, two or more steps.
3. Strategic thinking, requiring reasoning, developing a plan or sequence of steps, involving some complexity, having more than one possible answer, generally taking less than 10 minutes to do.
4. Extended thinking, requiring an investigation, time to think and process multiple conditions of the problem or task, and requiring more than 10 minutes to do nonroutine manipulations.

**Range-of-knowledge correspondence**: Does the span of knowledge a standard expects of students correspond to the span of knowledge that students need to correctly answer the assessment items or activities?

**Balance of representation**: This criterion indicates the extent to which assessment items are evenly distributed across learning objectives within a standard.

How the alignments rated
Alignment between assessments and standards varied, without any discernable pattern, across grade levels, content areas, and states. Assessments and standards of two of the four states satisfied the categorical concurrence criterion—that is, if an assessment had at least six items measuring content from a standard. (The number six is based on esti-
Percent of Multiple-Choice Items of Total Assessment by State, Content Area, and Grade

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Grade</th>
<th>Total Number of Assessment Items N</th>
<th>Percent of Multiple-Choice Items %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>3</td>
<td>44</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>70</td>
<td>86</td>
</tr>
<tr>
<td>Math</td>
<td>3</td>
<td>50</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>61</td>
<td>83</td>
</tr>
<tr>
<td>State B</td>
<td>Math</td>
<td>4</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>86</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>State C</td>
<td>Math</td>
<td>4</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>74</td>
<td>84</td>
</tr>
<tr>
<td>State D</td>
<td>Science</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>Math</td>
<td>4</td>
<td>54</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>51</td>
<td>61</td>
</tr>
</tbody>
</table>

State D had fewer items in the same grade level and content areas as other states, but it incorporated more constructed-response items.

The number of items that could produce a reasonably reliable scale for estimating students' mastery of content on that scale.) Two of the four states lacked a sufficient number of assessment items measuring content knowledge for more than one-quarter of the standards. Even when using a high number of assessment items at a grade level, some states have distributed their items unevenly so that one-fourth or more of the standards had fewer than six items measuring knowledge related to each of these standards.

Alignment was weak on depth-of-knowledge consistency. For this analysis, at least 50 percent of the assessment items corresponding to a learning objective had to be at or above the level of knowledge of the learning objective. A high percentage of the state assessments used items that were less demanding than those of the corresponding objectives.

The lowest degree of alignment was found on range-of-knowledge correspondence. At least 50 percent of the objectives within a standard had to have a related assessment item or activity to be judged acceptable. This cutoff for acceptance is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. Only one of the four state standards evaluated attained a high degree of range-of-knowledge correspondence for at least two of the analyses completed across all grades and content areas—grade 8 mathematics in that state had 86 percent of the standards meet this criterion and grade 10 mathematics had all of its standards meet this criterion. For the other states, items were generally clustered among a few of the objectives rather than covering the full range of objectives within a standard. As a consequence, many of the tests measured students' knowledge of only a small proportion of the full domain of content knowledge specified by the standards.

Most of the assessments and standards analyzed were aligned according to the balance of representation. Webb offers the explanation that the items were generally distributed among the corresponding objectives without a disproportionate number measuring any one objective.

Conclusion

A major goal of Webb’s study was to develop a valid and reliable process for analyzing the alignment between standards and assessments. The process distinguishes among the different attributes of alignment and detects specific ways that alignment can be improved.

Webb’s study refined the procedures for determining degrees of alignment, making them more standardized and useful, so that states and districts can better use them to understand the agreement between their own standards and assessments.

For more information, see Norman L. Webb, “Alignment of Science and Mathematics Standards and Assessments in Four States,” published jointly by the National Institute for Science Education and the Council of Chief State School Officers, June 30, 1999 (NISE Monograph No. 18).

TERMS USED

Standards refers to the most general expectations for a grade and content area. Goal refers to the next level of specificity of expectations. Objectives further delineate expectations stated as a goal.
Testing accommodations can help students

With the passage of the 1997 Individuals with Disabilities Education Act, using testing accommodations has become an important way of encouraging all students with disabilities to participate in statewide and other large-scale assessment programs. Testing accommodations are commonly defined as a change in the way that a test is administered or responded to by the student tested and are intended to “correct” for distortions in scores caused by a disability. Some examples of accommodations are providing students extra time to complete the assessment, giving a test in an isolated room or providing accessibility furniture, allowing oral responses to replace written responses, and providing assessment in large print or changing the language or mode of presentation. Often these accommodations are provided in various combinations, sometimes called “packages.”

Accommodations are meant to reduce the impediments facing students with disabilities so that when they take assessments they may fully show what they know and can do. Using appropriate testing accommodations also can increase the validity of the inferences one makes from the test scores of students with disabilities.

UW–Madison professors Stephen N. Elliott and Thomas R. Kratochwill say that testing accommodations can play an important role in increasing the validity of the resulting scores if the purpose of the test is understood, yet few data-based reports of this relationship have been published. Meanwhile, educators on the front lines who are administering tests to all students must select and implement “valid and reasonable” testing accommodations. This gap between conceptual issues of test validity and testing practices must be bridged.

The ramp metaphor

One metaphor for testing accommodations is an access ramp. Historically, access ramps were part of a package of testing accommodations for students with significant physical impairments. If students with disabilities can’t get to the testing room, they certainly can’t demonstrate what they know or can do. Testing accommodations facilitate access to a test for students with a wide range of disabilities, just like a ramp facilitates access to a building for students with physical disabilities. The tests that students are required to take are designed to measure some specific target cognitive skills or abilities, such as mathematical reasoning and computations, but almost always assume that students have the skills to access the test, such as attending to instructions, reading story problems, and writing responses. Some students, in particular many students with disabilities, have difficulty with some of the cognitive access skills needed to “get into” the test.

Valid testing accommodations, just like an access ramp, should be designed to reduce problems of access to a test and enable students to demonstrate what they know and can do with regard to the skills or abilities the test is targeting. Educators then have a more accurate or valid picture of student abilities in the area tested.

Most states now have guidelines that define testing accommodations and that identify acceptable accommodation practices for students with disabilities. But because of a lack of experimental studies, little has been known about the effects of accommodations on test scores.

To address that need, Elliott and Kratochwill and several graduate students are studying the use and effectiveness of testing accommodations packages being provided to a sample of 100 fourth-graders. Of this sample, 59 are disabled and 41 are identified as students with a variety of disabilities (i.e., learning disabilities, emotional disturbances, cognitive disabilities, speech or language disabilities, autism, speech difficulties, or hearing impairment).

Midway through a three-year study funded by the U.S. Department of Education’s Office of Special Education and Rehabilitative Services, Elliott
and Kratochwill have determined that more than 75 percent of the testing accommodations packages they examined with these students have a moderate to large effect on test scores earned on complex mathematics and science performance tasks. For a small percentage of students (less than 20 percent), the effects of the individualized accommodations recommended by their teachers have not been positive. On average, the resulting effect size, when comparing accommodated scores to unaccommodated scores for the same type of student, was .83. It should be noted, however, that decisions about the validity of these accommodations are professional judgments and involve several variables, including the purpose of the test, the type of score interpretation method (i.e., norm-referenced or criterion-referenced), and the nature of a student’s disability.

Elliott and Kratochwill share these additional observations:

- Researchers examining the effects of only one accommodation in isolation are studying a condition that rarely, if ever, exists. Testing accommodations are “packages” of changes to the testing event and conditions, as noted above.

- Use of several control or comparison conditions are needed to advance understanding of the relative effects of testing accommodations on students’ scores and to facilitate validity inferences.

- Testing accommodations also can improve the test scores of students without disabilities. The finding is not unexpected and deserves further study, say Elliott and Kratochwill. Perhaps some of the students in the nondisabled sample actually had significant academic disabilities or difficulties, but had not been officially identified as having a disability, or perhaps the performance tasks simply need more refinement to be easily comprehended by most students. Or perhaps the accommodations are not construct-independent and are yielding invalid scores. Elliott and Kratochwill encourage researchers to provide detailed accounts of the accommodations provided to students and to go beyond dichotomous group comparisons, such as students with and without disabilities. Elliott and Kratochwill believe that future work on testing accommodations will be more effective if the cognitive ability of student participants is controlled or, better yet, considered a separate variable.

**Making best use of scores**

Given the increasingly higher stakes that are placed on state and districtwide test scores, attention must be given to individuals’ perceptions and judgments about the usefulness and fairness of accommodations for students with disabilities. Although federal law now requires educators to provide accommodations to students with identified disabilities, researchers have yet to find out exactly what types of practices are used in the schools to make accommodations for students on statewide and districtwide tests, how accommodated students’ scores are reported and, ultimately, how their scores are used to make decisions at various educational levels (e.g., student, school, district, or state).

Studies are needed to obtain evidence of the consequences accommodations may have by surveying students, teachers, and parents about their perceptions of the validity and fairness of testing accommodations. In particular, feedback from teachers is needed to learn how they use test results obtained from students receiving accommodations and to procure their professional judgment about the degree to which they believe accommodations interfered with the purposes of the test.

For more information on testing accommodations and test validity visit the WCER web site (http://www.wcer.wisc.edu/Projects.htm) and explore “School Psychology and Special Education” projects or see Elliott’s new book, *Educational Assessment and Accountability for All Students: Facilitating the Meaningful Participation of Students with Disabilities in District and Statewide Assessment Programs*, available through the Wisconsin Department of Public Instruction (800) 243–8782, pubsales@dpi.state.wi.us.
Assessment drives student learning

College students get a pretty good idea of what the instructor considers important on the first day of class when the syllabus is handed out. Along with the syllabus, students are cued to “what’s important” when testing time comes around. Students tend to learn what their teachers guide them to learn through the assessments they use.

Yet traditional testing methods have been limited measures of student learning and have had limited value for guiding student learning. The methods are often inconsistent with the increasing emphasis being placed on students' ability to
1. think analytically
2. understand and communicate at both detailed and big-picture levels, and
3. acquire lifelong skills that permit continuous adaptation to workplaces that are in constant flux.

Because assessment is in many respects the glue that links the components of a course—its content, instructional methods, and skills development—changes in these components require coordinated changes in assessment.

Field-tested Learning Assessment Guide

The College Level One (CL–1) project of WCER's National Institute for Science Education (NISE) has established a web site (http://www.wcer.wisc.edu/nise/cl1/flag/) to provide resources that enable college-level science, mathematics, engineering, and technology (SMET) instructors to design more effective assessments of student learning.

The Field-tested Learning Assessment Guide (FLAG) offers a readily accessible, up-to-date resource of classroom-tested assessment tools for instructors who wish to share and implement new approaches to evaluating student learning, attitudes, and performance.

Each of the techniques and tools in this web-based guide has been developed, tested, and refined in classrooms and teaching laboratories at colleges and universities throughout the country. Each has been reviewed by experienced instructors in the appropriate discipline and by an editorial board of recognized experts in testing and measurement.

The site aims to help educators match their course goals with appropriate classroom assessment techniques (CATs)—self-contained, self-instructional, web-based modules that introduce useful techniques for assessing conceptual, attitudinal and performance-based course goals in SMET disciplines. Each CAT is linked to a set of tools or instruments that may be implemented directly in SMET courses with only a modicum of additional effort. The “tool chest” may be searched by discipline (e.g., agriculture, chemistry, engineering) or by technique (e.g., attitude survey, concept mapping, portfolio).

Another section of the site provides supplementary information on assessment practices in college and university SMET disciplines, including links to related web sites, an annotated bibliography, and a list of SMET instructors and assessment specialists who have expertise in one or more of the CATs.

Assessment moves with the target

To many, the word “assessment” simply means the process by which students are assigned grades. Assessment is much more than this, however, says Arthur B. Ellis, UW–Madison Meloche-Bascom Professor of Chemistry. “Assessment is a way to provide instructors with data for improving their teaching methods and for guiding and motivating students to be actively involved in their own learning,” he says. As such, assessment provides important feedback to both instructors and students.

Assessment gives educators essential information about what students are learning and about the extent to which educators are meeting teaching goals. But the true power of assessment comes in also using it to give feedback to students. Improving the quality of learning in college-level SMET courses involves not just determining how well students have mastered content at the end of the course. Improving the quality of learning also involves determining how well students master content throughout the course.

continued on next page
Classroom conversations socialize students

There’s a lot going on below the surface of classroom conversations.

No matter what the subject—biology, chemistry, U.S. history—classroom conversations are the medium of student learning. Through classroom conversation, students and teachers build, share, and reflect on knowledge. Classroom conversations equip students to join larger conversations that transcend a given class, a particular school year, and a specific subject matter area.

WCER’s Project on Academic Language Socialization (PALS) is identifying the processes through which students and teachers form coherent conversations.

WCER researcher and UW–Madison English Professor Jane Zuengler explains that her research addresses students’ language socialization across the curriculum with a focus on linguistic diversity. Some students in the classrooms she is studying are native speakers of Spanish, Hmong, Lao, or Thai.

Zuengler emphasizes the importance of language as embedded in learning across the curriculum. Learning in a subject area such as math, science, or social studies means both learning through language and learning to use language. Teacher preparation for all subject matters should recognize the integral role of language in the learning process. Zuengler and colleagues focus on the class as a dynamic community, a social system in development. Their research emphasizes process and development.

“The processes of teaching, learning, and co-constructing the discourse modes within a subject matter class must be understood so that teachers across the curriculum can foster the kinds of classroom interactions that lead to subject matter learning,” Zuengler says. By looking closely at the details of classroom talk, the study should shed light on significant areas of educational activity that have not yet been fully studied.

Coherent or cohesive conversations

As part of their socialization process, students learn the values, beliefs, and social practices of their classroom, school, and wider communities.

Just as important, students are socialized “to language.” They are required to demonstrate their knowledge in particular ways, appropriate to the conventions and discourse rules of the different communities in which they participate and to which they belong. They participate in specific conversations, with highly conventionalized patterns of interaction, including turn-taking, topic nomination, and argument structure. The language socialization framework PALS uses allows researchers to examine and discuss both coherent conversations and cohesive conversations as they occur in classrooms. Coherence refers to the relationships of shared knowledge between speakers, while cohesiveness refers to grammatical relationships among “text” elements identifying and developing new tools that can be used to assess student learning and to share their ideas and experiences.

For example, linguistic elements, such as the use of certain pronouns, can reveal important aspects of the construction of roles in a particular subject matter classroom and can help explain why some students become active participants in ongoing interactions while others remain on the margins.

In their choices of even small grammatical aspects of conversations, teachers and students demonstrate and build ongoing patterns of subject matter learning. This analysis contributes to both the quantitative and qualitative components of the study of language development over time among both native and nonnative speakers of English.

Assessment

Assessment is undergoing exciting changes in college SMET courses,” says Ellis. “The intent of the FLAG web site is to capture the vitality of assessment.” The CL–1 Team views assessment as a moving target and this web site as a living product, providing both a mechanism for rapid dissemination of assessment-related developments and a forum for their discussion.

NISE’s College Level One Team invites college SMET instructors to join their colleagues who are identifying and developing new tools that can be used to assess student learning and to share their ideas and experiences.

For more information about the National Institute for Science Education (NISE), see www.wcer.wisc.edu/nise/.

For more about the NISE College Level One project, see www.wcer.wisc.edu/nise/cl1/.
or “discourse” elements. This rather technical-sounding distinction nevertheless has an important bearing on how students’ classroom talk works toward or against effective learning. The focus on exchanges and unfolding classroom talk allows researchers to see the ways in which “cohesive” conversations are constructed. Exchanges are identified, in part, through the presence of linguistic features that mark each participant’s contribution as tied to the one before.

The coherence of these conversations, however, can only be determined by the participants themselves and in the interaction that’s evident when students treat one another’s contributions as “coherent” (meaningful) or not. This research looks for conversational patterns leading toward coherence as well as away from it. And, importantly, coherence in and of itself is not necessarily educationally beneficial. PALS researchers have observed a number of exchanges in a biology class that are coherent (in that students and teacher work toward a shared orientation), but that emphasize grade-getting and appropriateness of form, rather than foster creative thinking about how, for example, genetic inheritance figures in people’s lives. PALS research also is discovering the ways subject matter becomes coherent for the students over the course of the school year. Examining videotaped interactions from early in the year through the end of the year, PALS researchers observe how bodies of knowledge are shaped (or not) by teachers and students toward subject matter goals as well as toward larger educational goals. Since PALS researchers follow some of the same students as they proceed in science and social studies from one year to the next, the project is also determining what knowledge and understanding the students carry with them into their conversations from year to year. These building blocks may lead students toward participating in the larger conversations about science and social studies that transcend biology and chemistry, for example. What’s being said in classrooms shapes student development on many levels. In the near future, this research will provide a much clearer picture of just how this development works. For more information, contact Zuengler at zuengler@facstaff.wisc.edu, or visit the CELA web site at http://www.wcer.wisc.edu/CELA.