Listening. Learning. Leading.®

R&D Connections

No. 12 • December 2009

Leveling the Field on Math and Science Tests for Students with Learning Disabilities

By Elizabeth Stone and Linda Cook

If scores on a state math or science assessment have been evaluated using data only from test takers without learning disabilities, can we assume the inferences made based on the test's scores will be fair and valid for all students?

What is a *nonstandard* test administration? Does administering the test under such conditions affect the fairness of the test or the validity of the inferences made from the test's scores?

And how do we know whether test scores reflect the same skills and knowledge for different groups of students — specifically for students with disabilities?

For anyone trying to understand the meaning of scores on state standards-based achievement tests, these are important questions to ask. Over the past few years, ETS has carried out studies of state standards-based achievement tests that were administered to students with disabilities. Specifically, the studies sought evidence about the fairness and validity of the inferences made from scores on these tests. These studies have considered those students with disabilities who took tests under standard conditions and those who took the tests under nonstandard conditions.

In the context of educational measurement, the terms we use in this article have meanings that may differ slightly

Editor's note: Elizabeth Stone and Linda Cook are, respectively, a senior research associate and principal research scientist in the Foundational and Validity Research area of ETS's Research & Development division.

Key Concepts

Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) provides these definitions for three concepts that are central to the topic of testing students with disabilities:

- *Validity* — "The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184).
- *Construct* — "The concept or the characteristic that a test is designed to measure" (p. 173).
- *Fairness* — "In testing, the principle that every test taker should be assessed in an equitable way" (p. 175).

from the ways in which they are used in general conversation, so it is important to clarify:

Validity is defined in *Standards for Educational and Psychological Testing* as the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p.9).

The same publication (AERA, APA, & NCME, 1999) gives this general definition of *fairness*: "In testing, the principle that every test taker should be assessed in an equitable way" (p. 175).

By *state assessments* or *state standards-based achievement tests*, we mean the tests that states require in order to determine whether students and districts are meeting the educational achievement goals established by state and federal mandates.

The validity-related literature that we reference gives great importance to the term *construct* and uses it frequently. *Standards for Educational and Psychological Testing* defines construct as the "concept or the characteristic that a test is designed to measure" (AERA, APA, & NCME, 1999, p.173).

The term *accommodation* as we use it in this article is reserved for test changes that a state believes do not alter the underlying construct measured by the test. The term *modification* refers to test changes that a state believes *may* alter the underlying construct that the test measures.

Why is validity such an important concept in testing? If test scores are to be used to make decisions that affect the direction of a test taker's life — for example, to make decisions about admission to college or graduate school, to allow a student to graduate, or to allow an examinee to receive certification — it is imperative that the test scores provide information that allows those interpretations to be made appropriately.

While there is a justifiable concern about the way test score interpretations affect individuals, consider also that test scores are used in the K-12 setting for accountability purposes that lead to evaluations and comparisons at the school and state levels. Inferences made at these levels can affect funding, staffing, and even curriculum development, and thus it is important to realize that test score validity can have more far-reaching consequences than is immediately apparent.

When students with learning disabilities take achievement tests in math and science, do their scores reflect their knowledge and skills, or do they reflect the presence of other unrelated factors?

In the case of the way these issues are discussed in this article, validity means this: When K–12 students with learning disabilities take state standards-based achievement tests in math and science, do their scores reflect their actual knowledge and skills in these subjects, or do they reflect the presence of some other factors that are unrelated to math and science?

More Inclusion

Research studies have shown that a smaller percentage of students with learning disabilities participate in state assessments than do their peers without learning disabilities. Furthermore, there is almost always a perfor-

mance gap between these groups of students on these assessments.

It is important to evaluate whether a performance gap on a state test is truly due to differences in proficiency or whether there are obstacles irrelevant to what the test is supposed to measure that are preventing students with disabilities from demonstrating the full extent of their knowledge.

In this article, we discuss some research that has taken place involving the issues of participation and proficiency. The article also provides examples of some of the work that ETS has done to examine validity and fairness issues related to scores on state standards-based math and science tests administered to students with learning disabilities.

For a long time, students with learning disabilities were generally instructed and tested separately from students without learning disabilities. The No Child Left Behind Act (NCLB) was one step that helped to change this situation. NCLB requires schools to demonstrate not only the proficiency of their total student population, but also the performance of different demographic groups within that population. Students with disabilities make up one such *accountability subgroup*.

The mandatory inclusion of students with disabilities in calculations of *adequate yearly progress*—which K–12 educators often refer to as AYP—is an important, but also challenging, requirement for schools to meet.

The number of students with disabilities in U.S. public schools is quite large. According to findings based on the National Center for Educational Statistics (NCES) Common Core of Data, in the 2003-2004 school year, more than 6 million students with disabilities—approximately 14% of all students—attended U.S. public schools (Cortiella, 2007).

Of these students with disabilities, about 46% were classified as having specific *learning* disabilities, which in this case refers to "a disorder in 1 or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in the imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations" (Individuals with Disabilities Education Act [IDEA], 1997, 2004).

The mandatory inclusion of students with disabilities in measures of adequate yearly progress is an important, but also challenging, requirement for schools to meet.

This definition is part of a body of U.S. legal code that has been developed over the years to address the educational needs of students with disabilities. One of the most widely cited parts of the legal code related to students with disabilities, the IDEA (1997, 2004) requires states to provide a means for participation in assessments.

One way to include students with learning disabilities is to allow them to take tests under nonstandard conditions, using various types of *testing modifications* or *testing accommodations*. Generally, when states feel that nonstandard testing conditions *do not* change the test's underlying construct — the knowledge, skills, or abilities it targets — they include the results of the assessments when they calculate their AYP. When states feel that the nonstandard testing conditions *do* change what is being tested, they *do not* include that student's scores in calculation of AYP.

Categories of Test Changes

Cortiella (2005) describes examples of four kinds of accommodations for students with learning disabilities:

Presentation

- Large print or braille test forms
- Magnification devices
- Calculators
- Arithmetic tables
- Audio accommodations

Response

- Marking answers in the test booklet rather than on a separate answer sheet
- Having another individual record answers for the student

Timing and scheduling

- Allowing the student extended time to complete the test
- Splitting the test session up into multiple sessions
- Allowing the student to take breaks more frequently than are usually provided

Test setting

- Testing somewhere separate from other students — for example, at home or in a quiet corner

By way of illustration, providing extended time is often considered to be a minor change in testing conditions that does not change what is being tested. On the other hand, the use of a calculator on a mathematics assessment is frequently considered to be a change that *does* affect what is being tested.

It is important to recognize that the definition of a reasonable accommodation may vary by state and by assessment. Cortiella (2005) discusses four types of accommodations, which we will refer to later.

Finding Performance Evidence

It is often difficult to get a good sense of how students with disabilities perform on state math and science assessments when compared with students without disabilities. The data necessary to evaluate performance on state assessments by any one demographic group — which measurement experts often refer to as *subgroups* — are not always readily available.

Even when subgroup results are available, states often report the results of students with learning disabilities as part of the larger, more generally-defined group of students with disabilities. This can complicate the task of making inferences about the meaning of scores because students with different disabilities can have very different experiences when taking the same test.

Studies that have looked at this sparse evidence have suggested that students with disabilities do not perform as well on state mathematics and science assessments — even when they take versions of the tests with accommodations (VanGetson & Thurlow, 2007; National Center for Educational Statistics, 2006).

Furthermore, VanGetson and Thurlow (2007) found that this *achievement gap* grows as students get older, so that the gap in performance between students with and without disabilities is larger in high school than it is in elementary school.

It is important to ask: If accommodations are administered to "level the playing field" (Tindal & Fuchs, 1999), do the achievement gaps observed by VanGetson and

Thurlow, and others, reflect true differences in achievement between students with and without disabilities? Or, do the observed differences in performance reflect *other* aspects of students' disabilities that the accommodations have not accounted for and that are not relevant to the construct (or characteristic) that the test is intended to measure?

Research on Accommodations

The complicated set of factors involved in testing students with various disabilities has led to much research into how these students perform on assessments, the effects of accommodations on their performance, and the validity of inferences people make about their test scores.

There are many types of accommodations that can be used on math and science assessments. Cortiella (2005) provides the four categories often used: presentation, response, timing and scheduling, and setting.

A goal of allowing accommodations on assessments is to level the playing field for students with learning disabilities.

Presentation accommodations include large print or braille test forms, magnification devices, calculators, arithmetic tables, and audio accommodations such as having a human read the test aloud or having an audio-recorded reading of the test on tape or CD. The latter type is also commonly referred to as a read-aloud, audio, or oral accommodation. Presentation accommodations benefit students who have trouble reading test questions or options.

Response accommodations include marking answers in the test booklet rather than on a separate answer sheet or having another individual record answers for the student. Students who are not able to go back and forth between pages or who have memory or writing difficulties may benefit from this type of accommodation.

Timing and scheduling accommodations include allowing the student extended time to complete the test, splitting the test session up into multiple sessions, or allowing the student to take breaks more frequently than they are usually provided. A student who has specific types of physical conditions or a student who has trouble concentrating for a prolonged period of time may benefit from this type of accommodation.

Finally, *test setting* accommodations involve testing somewhere separate from other students, for example at home or in a quiet corner. A student who has trouble concentrating in large groups may benefit from these types of accommodations.

A goal of using accommodations on assessments is to remove obstacles for students with disabilities. Much of the research into assessments for students with disabilities focuses on how various accommodations affect performance or how they change what the test measures.

For instance, students with disabilities may need more time to complete a task than students without disabilities need. If the test is not intended to measure speed in completing a task, then allowing extra time should not affect the scores on the test (Sireci, Scarpati, & Li, 2005).

Time limits usually have a more practical function. Studies have shown that, on most tests, giving students extra time usually does not give them an unfair advantage (Bridgeman, McBride, & Monaghan, 2004). Thus, allowing extra time should not affect the meaning of test scores.

Validating Tests

Professionals validate tests against various criteria:

- Does the test cover the right content and does it cover enough of it?
- Is there a link between the scores and external measures of the same or a similar construct?
- When answering test questions, do test takers use the skills being tested?
- Are the claims made about the test borne out in the consequences of using the scores?
- Do statistical analyses indicate that performance may be related to membership in a particular demographic group?

A similar argument might be made for allowing students to mark answers directly in the test booklet or having someone help them to record the responses. If the test is not intended to measure the ability to transcribe, allowing this type of accommodation should not change the meaning of scores (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998).

A more controversial accommodation is a read-aloud accommodation (c.f., Bolt & Thurlow, 2006; Bolt & Ysseldyke, 2006; Huynh, Meyer, & Gallant, 2004). Most of the debate surrounding this accommodation centers on whether to allow it for a test of reading skills, but researchers have also explored the use of read-aloud as an appropriate accommodation for tests that are not intended to measure any aspect of reading, such as a math or science test.

On such tests, where students' scores are intended to reflect their knowledge of math or science, it may make sense to consider allowing students with reading-based disabilities to use a read-aloud accommodation in order to more readily display their skills in math or science.

Studying Validity and Fairness

Testing professionals work to ensure that the test scores can be used as intended by engaging in validation practices that usually involve evaluating the test against various criteria. *Standards for Educational and Psychological Testing* calls upon test developers and researchers to examine the following sources for evidence of validity (AERA, APA, & NCME, 1999, pp. 11-15):

- *Test content* — Does the test cover the right content and does it cover enough of it? To establish this, content or subject matter experts develop and review the test questions.
- *Relationship of test scores with other variables* — Is there a link between the scores on an assessment and external measures of the same or similar construct? For example, if students who perform well on a college admissions test also perform well in college, this may lend credibility to the predictive claims of the admissions test.

- *Evidence of student response processes* — In attempting to answer a test question, do test takers use the skills being tested? This kind of validity research can be performed, for example, through the use of *think aloud* or *cognitive lab* procedures that involve listening to test takers talk about what they are thinking while they ponder test questions.
- *The consequences of assessments* — Are the claims made about the test borne out in the consequences of using the test scores in decision making? For example, if the test is supposed to be able to distinguish between those who will and those who will not be successful in a particular job, the subsequent job performance of candidates who were hired may provide validity evidence.
- *Investigation of internal structure of assessments* — Do statistical analyses of the test results indicate that test takers' performance on the test may be related to their membership in a particular demographic group (e.g., *students with disabilities*)? Such investigations may be performed using statistical procedures known as *factor analysis* and *differential item functioning analysis*.

This last bullet describes an important part of ETS's work in this area.

Powerful Tools

So what is factor analysis? And what is differential item functioning analysis (also known as *DIF analysis*)?

Factor analysis allows researchers to identify underlying *dimensions,* or factors, that a test measures. The general question asked in the ETS studies that use factor analysis is this: Do the test's questions measure only one dimension, such as reading ability, or more than one dimension, such as understanding vocabulary and reading comprehension?

Factor analysis also can be used to investigate whether a test has the same underlying set of dimensions when it is given to different groups of students. For further details on factor analysis and how it is used in validity research, readers may consult Kane (2006).

Differential item functioning (DIF) analysis helps to identify test questions that may be functioning differently for groups of test-takers who have the same level of proficiency. Most DIF analyses compare individuals from different gender, race, or ethnic groups, but they can also be used to compare groups with different disability status. The assumption is that test takers of equal ability should have the same chance of correctly answering a test question regardless of the group they belong to.

An item, or test question, is said to *show DIF* if, once groups are matched on a measure of ability, one group gets a question right significantly more often than the other group does. This is just one simplified way to describe a common method, though the underlying idea holds for most DIF procedures.

If a question appears to show at least a moderate level of DIF, the question comes under review by specially trained subject matter experts. Depending on the result of that review, the question may be left as is, revised, or deleted from the pool of possible test questions. Testing officials may also decide not to include the question when determining test scores. Interested readers may consider Holland and Wainer (1993) for more information on DIF.

Taken together, the procedures of factor analysis and DIF are powerful analytical tools

to help determine if a test is measuring the same thing for all students — both those with and without learning disabilities. This question is particularly important for anyone interested in interpreting scores from tests that states use for accountability purposes.

Comparing Groups

States rightly do not want to combine, for the purpose of accountability, the scores from different groups if the test is measuring different knowledge or skills for different groups. Doing so would provide an inaccurate picture of the combined group's proficiency.

Researchers at ETS have carried out studies that used DIF analysis and factor analysis to examine the fairness and validity of fifth grade math and science tests for students with learning disabilities. The goal of both the factor analysis and the DIF studies was to explore whether or not the assessments were measuring the same underlying skills for the groups defined as follows:

- Students without disabilities taking the test under standard conditions
- Students with learning disabilities taking the test under standard conditions — i.e., without *accommodations* or *modifications* (see the earlier discussion of the distinction between accommodations and modifications and what testing changes states typically consider to be in each category)
- Students with learning disabilities taking the test with *accommodations*
- Students with learning disabilities taking as the test with *modifications*

The results of the factor analysis of the math test indicated that the test measured three related factors that appeared to be common for all of the groups (students without learning disabilities and students with learning disabilities testing under various conditions) listed above. The results of the factor analysis of the science test showed that the test measured a single factor that appeared to be common for all of the same groups.

The DIF analyses of the science test yielded no evidence that the items were functioning differently for the groups described earlier. Some DIF was detected on the math test in the comparison involving students with disabilities using math modifications. This result may be of interest for more focused research.

However, neither test showed large amounts of DIF for any of the groups studied. The interested reader can find details of the factor analyses for the math and science tests in Cook, Eignor, Steinberg, and Sawaki (2008), and Steinberg, Cline, and Sawaki (2008), respectively. The results of the DIF analysis for the math and science tests are presented in Cline, Cook, and Stone (2008).

Summary

The goal of presenting a level playing field for standardized testing is one of great consequence. Information gained from research studies such as those mentioned here may be used during test development to design tests that are more accessible for all students.

In addition, knowledge obtained through this research may lead to new and more effective accommodations that will allow students with disabilities to demonstrate their proficiency without undue impediment. At ETS, these studies form part of a comprehensive program of research that is integral to the organization's mission to provide assessments that are of high quality and are fair for all students.

References

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

Bolt, S. E., & Thurlow, M. L. (2006). *Item-level effects of the read-aloud accommodation for students with reading disabilities* (Synthesis Report 65). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19,* 329-355.

Bridgeman, B., McBride, A., & Monaghan, W. (2004). *R&D Connections — Testing and time limits* (Report No. RDC-01). Princeton, NJ: Educational Testing Service.

Cline, F., Cook, L. L., & Stone, E. (2008, March). *An examination of differential item functioning on grade 5 math and science assessments for students with disabilities*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cook, L. L., Eignor, D. R., Steinberg, Y., & Sawaki, Y. (2008, March). *Using factor analysis to compare the internal structure of a state standards-based math assessment*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cortiella, C. (2005). *No Child Left Behind: Determining appropriate assessment accommodations for students with disabilities*. Retrieved December 8, 2008 from http://www.readingrockets.org/article/10938

Cortiella, C. (2007). *Rewards and roadblocks: How students with disabilities are faring under No Child Left Behind.* Retrieved July 28, 2009 from http://www.ncld.org/on-capitol-hill/policy-related-publications/rewards-a-roadblocks

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education, 17*, 38-57.

Individuals with Disabilities Education Act of 1997, 20 U.S.C. 1412(a) (17) (A). (1997).

Individuals with Disabilities Education Act of 2004, 20 U.S.C. S 1400 et seq. (2004).

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 18-64). Washington, DC: American Council on Education/Praeger.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

National Center for Educational Statistics. (2006). *The nation's report card: National and state reports in science now available.* Retrieved April 2, 2009, from http://nationsreportcard.gov/science_2005/

No Child Left Behind Act of 2001, 20 U.S.C. 6301 *et seq.* (2001) (PL 107-110).

Sireci, S. G., Scarpati, S.E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75,* 457-490.

Steinberg, J., Cline, F., & Sawaki, Y. (2008, March). *Examining the factor structure of a state standards-based assessment of science for students with disabilities*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities in large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

VanGetson, G. R., & Thurlow, M. L. (2007). *Nearing the target in disaggregated subgroup reporting to the public on 2004-2005 assessment results* (Technical Report 46). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 1, 2008 from http://education.umn.edu/NCEO/OnlinePubs/Tech46/

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
e-mail: **RDWeb@ets.org**

Editor: Jeff Johnson

Visit ETS Research & Development on the Web at **www.ets.org/research**