Listening. Learning. Leading.®

R & D Connections

No. 8 • December 2008

Ensuring Valid Content Tests for English Language Learners

By John W. Young

Who is an ELL?

According to the definitions in the No Child Left Behind Act, an English Language Learner:

- Is between the ages of 3 and 21
- Is enrolled or preparing to enroll in grades pre-kindergarten (PK) to 12
- Was not born in the United States or has a native language other than English
- Has difficulties in English that are sufficient to deny the student the ability to meet his or her state's proficiency levels of achievement

As students in school, all of us have taken tests that assessed our knowledge or skills in specific subject areas, such as mathematics.

These tests, which are commonly called content assessments, are important components of many large-scale assessment programs. With passage of the No Child Left Behind (NCLB) legislation, schools are now being held accountable based on their students' performance on required annual content assessments.

Most people are aware that NCLB has had a huge impact on education in the United States. Fewer people are aware, however, of how much NCLB has changed the education of students who are English language learners (ELLs).

NCLB requires ELLs to be included in content assessments, such as in mathematics, reading, and science. This raises an important challenge for ETS and other test publishers: How do we ensure that our tests are valid and fair for all groups of students, including those who may still be learning English?

Although non-English speaking students have always been enrolled in American schools, our greater understanding of their instructional needs coupled with the increased emphasis on testing for accountability purposes means that any content assessment we use must accurately measure their content knowledge rather than their proficiency in English.

ELLs are among the numerous demographic groups that are included under the NCLB accountability system. According to NCLB, an ELL is defined by the following four characteristics: (a) between the ages of 3 and 21; (b) enrolled or preparing to enroll in grades pre-kindergarten (PK) to 12; (c) not born in the United States or has a native language other than English; and (d) has difficulties in English that are sufficient to deny the student the ability to meet his or her state's proficiency levels of achievement.

According to the National Clearinghouse for English Language Acquisition (NCELA), ELLs in grades PK–12 are one of the fastest growing subpopulations of students in U.S. schools (NCELA, 2007).

Editor's note: John W. Young is a research director in the Foundational and Validity Research area of ETS's Research & Development division.

"By the year 2025, English language learners are projected to comprise 1 in 4 students in the United States."

During the 2005–06 school year, ELLs numbered over 5 million nationwide and represented 1 in 9 students in U.S. classrooms. By the year 2025, ELLs are projected to comprise 1 in 4 students in the United States. Although approximately 80% of ELLs are native Spanish speakers, ELLs collectively speak nearly 400 different home languages.

Students who take content assessments vary according to their native language and current level of English language proficiency. Most school districts and states use a classification system similar to the following one:

- **Native English speaker**—A student who speaks only English.
- **English language learner**—A student whose lack of English proficiency may affect his or her school performance.
- **Initially fluent English proficient**—A student who is bilingual or multilingual but is proficient in English and does not require special language services or accommodations in school.
- **Redesignated fluent English proficient**— A student formerly classified as an ELL who has acquired English proficiency and can function in a mainstream classroom without special language services or accommodations in school.

Individuals in the last three categories are sometimes referred to collectively as language-minority students.

Validity of Content Assessments

It is important to distinguish between content assessments and assessments of English language proficiency. Content assessments focus on subject matter knowledge and skills, while assessments of English language proficiency focus on the ability to communicate in English in one or more modalities (listening, reading, speaking, and writing).

Regardless of their level of English proficiency, all students should have an opportunity to demonstrate their knowledge or skills in the content being assessed. Although students may have different English proficiency classifications, the meaning of their scores on content assessments should be comparable.

Comparability of scores is necessary to ensure that an assessment is fair for all examinee groups. Before we can discuss what is necessary to ensure this, we must understand several key psychometric concepts:

Construct—This refers to the knowledge, skills, or proficiency an assessment targets.

Construct validity—This refers to the appropriateness of inferences or decisions made based upon a set of test scores. An assessment, in and of itself, is neither valid nor invalid. Rather, validity refers to whether we have the theoretical and empirical evidence to support the interpretations we attach to test scores.

What is construct-irrelevant variance?

These types of language may cause content assessment items to test something other than what they are intended to test:

- Unfamiliar vocabulary that is not related to the construct
- Cultural references or idiomatic expressions (such as "being on the ball") that are not equally familiar to all students
- Syntax that may be confusing or ambiguous (such as negatives or double negatives)
- Low-frequency, long, or morphologically complex words and long sentences
- Sentence structure that may be confusing or difficult to follow (such as passive voice or sentences with multiple clauses)
- Syntax or vocabulary that is above the test's target grade level

Construct-irrelevant variance—Factors that influence students' test scores but are not directly related to the construct may introduce what we call construct-irrelevant variance into the scores. This is undesirable because it means that test performance may be partly due to factors that differ from the construct the test is designed to measure.

An analogy from sports may help to illustrate construct-irrelevant variance: A sprinter who runs with a strong wind at his back during the entire race (such as in the 100 meter dash) will often have a faster time than he would have if he ran the same distance with little wind or if he ran into the wind. His wind-aided time would thus be affected by construct-irrelevant variance. In fact, in certain track events, there is a limit on wind speed for the purpose of certifying records.

For most content assessments, the target construct is the student's knowledge or proficiency in the subject matter. Construct definitions can be broad, such as *knowledge of elementary algebra*, or more narrow, such as *being able to solve equations with one unknown*.

Since the main purpose of most content tests is to assess a student's subject matter knowledge, the test questions should not require a level of English proficiency that is so high as to pose difficulty in understanding the task presented in the question. This is a concern for all students, including proficient native speakers, but it is especially a concern for students who may not be fully proficient in English, such as is the case with ELLs.

If some examinees have difficulty in comprehending the test questions on a content assessment, this potentially creates a situation in which construct-irrelevant variance (in this case, as an outcome of English proficiency level) may play a role in some students' scores. For example, imagine a fourth-grade student who recently moved to the United States from another country; this student has strong mathematics skills but is still learning English. As this student is an ELL, his score on any mathematics assessment he takes should reflect his knowledge of mathematics, rather than his proficiency in English.

For assessments of language and literacy skills learned in English language arts courses, the distinction between English language proficiency and content knowledge is not always clear. By explicitly defining the skills and knowledge that are developed in English language arts, we can better define the construct that should be assessed.

Ensuring Valid Assessments

To ensure that content assessments are valid for all students, including ELLs, test developers should minimize construct-irrelevant variance as much as possible. In fact, avoiding construct-irrelevant variance is the main guiding principle that all test developers attempt to adhere to in creating any assessment.

In the case of ELLs, one major cause of construct-irrelevant variance in content assessments is the use of language that is not entirely accessible. This can occur in a number of different ways, including through the following usages:

- Unfamiliar vocabulary that is not related to the construct (August, Carlo, Dressler, & Snow, 2005; Bailey, Huang, Shin, Farnsworth, & Butler, 2007)
- Cultural references or idiomatic expressions (such as "being on the ball") that are not equally familiar to all students (Bernhardt, 2005)
- Syntax that may be confusing or ambiguous (such as negatives or double negatives) (Abedi, 2006; Cummins, Kintsch, Reusser, & Weimer, 1988)
- Low-frequency, long, or morphologically complex words and long sentences (Abedi, 2006; Abedi, Lord & Plummer, 1995)
- Sentence structure that may be confusing or difficult to follow (such as passive voice or sentences with multiple clauses) (Abedi & Lord, 2001; Forster & Olbrei, 1973; Schachter, 1983)
- Syntax or vocabulary that is above the test's target grade level (Borgioli, 2008)

Test developers use various strategies in their efforts to ensure valid content assessments for ELLs, including:

- Using linguistically modified or simplified language
- Minimizing unnecessary language or information
- Including examples with familiar contexts (such as school settings)
- Including examples using objects familiar to all students (such as school materials, like pencils, rather than home language such as chores or cooking)

These test development strategies improve the quality of assessments not only for ELLs, but also potentially for other groups of examinees, including students who read below grade level or have reading or other disabilities.

Testing Accommodations

ELLs may also be eligible for testing accommodations when taking content assessments. The main purpose for providing accommodations to examinees is to promote equity in assessment. For ELLs, testing accommodations are a means to ensure that they have the same opportunity as other students to demonstrate their knowledge, skills, or proficiency.

Note that accommodations provided for testing may differ from those provided during the course of classroom instruction. Examples of testing accommodations that are currently allowed for ELLs in one or more states include:

- Test versions translated into a student's native language
- Access to bilingual glossaries or word lists
- Testing in smaller, rather than larger, groups

"Testing accommodations vary widely from state to state and research is not yet conclusive on the appropriateness and effectiveness of different accommodations for ELLs."

- Extended testing time
- Extra breaks during testing

Testing accommodations vary widely from state to state and research is not yet conclusive on the appropriateness and effectiveness of different accommodations for ELLs (Rivera, Collum, Willner, & Sia, 2006).

One research finding that appears to be promising, however, is that testing accommodations that provide direct linguistic support (such as access to bilingual glossaries) are generally more effective than accommodations classified as providing indirect linguistic support (such as extended testing time).

However, we should interpret such findings cautiously. For example, access to bilingual glossaries is not uniformly effective for all ELLs, as this accommodation does not help those with low levels of English proficiency as much as it helps students with greater English proficiency (Rivera et al., 2006).

Validity Research for Content Assessments

After an assessment has been administered, it is generally useful to conduct research studies on the test results in order to understand whether the assessment functioned as expected. The general topic of examining differences in test validity for different examinee groups is known as differential validity.

When examinees with different levels of English proficiency take the same content assessment, we can perform certain types of statistical analyses to determine whether it is valid to compare their scores. Two of these analyses provide information on statistical indicators known as *internal consistency reliability* and *internal factor structure*.

Internal consistency reliability is a measure of how similarly the items in an assessment relate to each other. How internal consistency reliability is calculated is beyond the scope of this article, but essentially it is an index of the degree of similarity of responses to all items in a test. In other words, internal consistency reliability assesses the degree of homogeneity among the test items.

The specifics of using factor analysis would likewise take too long to explain in this article. In general, though, when psychometricians conduct factor analysis, they are interested in the number and significance of the constructs that account for how students answer test items correctly. One measure of a test's validity for different groups of examinees is whether the constructs and their statistical relationships with the test items are similar across groups.

Another type of analysis that is useful for investigating examinee group differences is *differential item functioning* (DIF). DIF is a method for checking statistically whether a test item is fair for all test takers regardless of their membership in some defined subgroup—say, *females*, *African Americans*, or *English language learners*. When they conduct DIF analysis on a test item, psychometricians compare people in the subgroup with people of *similar ability* in a comparison group.

Here is a simplified way of understanding DIF: Psychometricians might compare males who scored 87 on the test as a whole with females who also scored 87. If

"Differential item functioning is a method for checking statistically whether a test item is fair for all test takers regardless of their membership in some defined subgroup. We can use this procedure to compare ELLs with native English speakers."

there is an item on which the females at this ability level performed significantly worse than their male counterparts, the item is said to "show DIF."

In reality, we carry out a DIF analysis across the entire range of scores on the test. We can use the same procedure to compare English language learners with native English speakers.

In comparing groups of examinees with different levels of English proficiency, results from analyses of internal consistency reliability, internal factor structure, and differential item functioning, as well as other statistical information, are useful for identifying items on which groups perform differently. These differences may be worthy of further investigation since they may indicate that different groups of examinees are responding differently to particular items. If so, it may be necessary to revise these items.

The causes of these differences are harder to uncover since they may be due to many factors. Some of these factors may be related to the assessment, such as the presence of items having certain characteristics that make them easier or harder for some groups but not others, or they may be related to external influences such as differences in opportunity to learn the content being assessed. Qualitative or ethnographic studies, using samples of students similar to those who originally took the assessments, may be necessary in order to further disentangle these factors. This information can be especially useful to researchers and test developers in refining and improving any assessment.

Summary

ELLs are a rapidly growing and increasingly visible segment of the U.S. student population. Because they, along with other students, are required to demonstrate proficiency in school subjects, the validity of the content assessments that are used is of primary concern. Ensuring that content assessments are valid for all examinees, including ELLs, is a complex undertaking.

It is critical that test developers have an understanding and clear definition of the construct being assessed as well as knowledge of the language skills of examinees with different levels of English proficiency. Applying the strategies mentioned earlier for minimizing construct-irrelevant variance can increase the likelihood that the resulting content assessments will be valid for all students, including ELLs.

Although research on ELLs taking large-scale content assessments has a relatively recent history, findings are beginning to emerge that provide some guidance on best practices in developing and administering assessments to ELLs.

References

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Abedi, J., Lord, C., & Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disability Research and Practice*, *20*, 50–57.

Bailey, A. L., Huang, B. H., Shin, H W., Farnsworth, T., & Butler, F. A. (2007). *Developing academic English language proficiency prototypes for 5th grade reading: Psychometric and linguistic profiles of tasks* (CSE Technical Report 727). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133–150.

Borgioli, G. (2008). Equity for English language learners in mathematics classrooms. *Teaching Children Mathematics*, *15*, 185–191.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405–438.

Forster, K. I. & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition*, *2*, 319–347.

National Clearinghouse for English Language Acquisition. (2007). The *growing numbers of limited English proficient students*: 1995-96 – 2005-06. Retrieved December 2, 2008, from: **http://www.ncela.gwu.edu/stats/2_nation.htm**.

Rivera, C., Collum, E., Willner, L. S., & Sia, J. K. (2006). An analysis of state assessment policies regarding the accommodation of English language learners. In Rivera, C. & Collum, E. (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1–174). Mahwah, NJ: Lawrence Erlbaum Associates.

Schachter, P. (1983). *On syntactic categories*. Bloomington, IN: Indiana University Linguistics Club.

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

Send comments about this publication to the above address or via the Web at **www.ets.org/research**