Strategies to Assess the Core Academic Knowledge of English Language Learners

Stanley Rabinowitz, Sri Ananda, & Andrew Bell
WestEd

With the goal of eliminating the achievement gap between advantaged and disadvantaged students, the No Child Left Behind Act of 2001 (NCLB) requires that **all** students achieve proficiency in English language arts and mathematics by 2014. Even subgroups considered at risk must demonstrate continuous progress towards proficiency in the core academic areas of English language arts and mathematics, as measured by their performance on state assessments. Failure to do so results in serious consequences for schools, districts and states.

As a result of NCLB, many states are scrambling to increase the accessibility of their core academic assessment program to the broadest range of students. Prior to NCLB, most states did not have sufficient data to make sound judgments about the academic progress of specific subgroups of students, especially those with special needs (e.g., students with disabilities and English language learners [ELLs]). The availability of better assessment data can help drive instructional improvements and increase student learning (Jamentz, 2001).

Measurement of academic progress of ELL students, as required by NCLB, poses specific challenges for states. Historically, ELL student performance on state tests has been low and improvement has been slow, with non school-related factors (e.g., parent education and socioeconomic status) substantially correlated with student learning (Abedi, 2004; Abedi & Dietel, 2004). The ever-changing composition of the ELL subgroup—with high-performing ELLs being redesignated as Fluent English Proficient and the continuous influx of new, often low-achieving ELL students—is another formidable challenge that recent changes in NCLB have attempted to address for accountability purposes. Finally, valid assessment of ELL performance is often elusive because ELL students' limited language ability (often both in English and their native language) makes it difficult to measure accurately their content knowledge.

This paper focuses on this last assessment issue: How do you increase the validity of assessments of ELL student performance on core academic content?[1] We begin by exploring NCLB expectations for ELL assessments and an increasingly popular approach to meeting these requirements proposed by some states—translation of assessments into students' native languages. Then, we present key research findings on attempts to increase access to and validity of assessment for ELLs. We conclude by proposing a comprehensive strategy for the assessment of ELL students' performance in core academic content.

[1] This paper focuses on NCLB's Title 1 core academic assessment requirements for ELL students, not on Title III's English proficiency testing requirements.

The Challenges of Test Translation

NCLB provisions for assessing the core academic knowledge of ELLs include the following:

- Assess in a valid and reliable manner;
- Make reasonable accommodations;
- Use language and form "most likely to yield accurate and reliable information (to the extent practicable);" and
- Make every effort to develop linguistically accessible academic assessment measures.

Some states have interpreted these provisions as *requiring* the development of primary language (i.e., translated) assessments for ELL students that are aligned to the state English language arts and mathematics content standards. This is not a correct interpretation. NCLB only requires adaptations in language and form "to the extent practicable." However, if a state were to consider assessing reading and mathematics in students' native language, several significant challenges would have to be overcome:

- *Determining which languages and dialects*: In the past, states could target its translation efforts on a few languages, primarily Spanish. But Spanish includes several dialects. Deciding if and when to include dialect variants of Spanish words in translation of test forms is a complicating factor. Beyond Spanish-speaking groups, states are now facing large numbers of growing student populations with various native languages. For example, California may have more than 70 distinct language groups in its K – 12 schools (FUSD, 2004). How will such states determine which languages (or dialects) to target for translation? Some states have turned to a numeric cutoff point (e.g., minimum of 1,000 students) to determine which language minority subgroups of students will be assessed in their native language and which will not. While this "minimum N" approach may be desirable from a cost/benefit perspective, it may be difficult to justify on the basis of assessment validity or fairness considerations.

- *Ensuring consistency of meaning across original and translated assessments*: Experienced content and assessment experts recognize that the art of translation requires more than just a word-for-word adoption from English to other languages. Experts use the term "adaptation" to describe the transformation process of a test from English to a second language (Stansfield, 2003). This process requires trained and experienced translators who are fluent in both languages and the specific content. Such translators must understand nuances in both language and the content standards in order for the assessments to be valid for all targeted purposes and populations.

- *Developing various support documents*: Translating or adapting the assessment is just one step in the process. All support materials (e.g., manuals, answer documents) need to be translated and tried out as well.

- *Acquiring native language proctors*: It may be necessary to use native language proctors to oversee the actual test administration. Such individuals may be difficult to locate and train, reflecting yet another "hidden" cost for the translation process.

- *Scoring*: Translating or adapting assessments that incorporate constructed-response items poses specific challenges. Qualified native-language speakers to score such items must be located and trained by either the state or its testing contractor. States must demonstrate not just that these native language scorers can score reliably, but to the same standards as those scoring in English.

- *Demonstrating Comparability*: Once the adapted assessments are developed, the state must demonstrate that they meet the same technical standards as the original, English version. Reliability and validity studies are expensive and may be difficult to implement with small numbers of ELL students. At a minimum, states must show that the adapted assessments are at least as reliable and valid for the subpopulation of students as the English assessments are for native English speakers. Moreover, evidence should be obtained that the adapted assessments (and any use of targeted accommodations) are sufficiently *more* valid for the target ELL population than the original English version of the assessments (e.g., ELL scores on the adapted assessment are higher than on the original version; indicators of bias on the adapted version are lower for ELLs than on the original version[2]). What constitutes a sufficient increase in validity that justifies the expense of translating and adapting assessments is subject to debate. Nonetheless, unless a sufficient increase in validity can be demonstrated, states would be hard pressed to justify costs of assessment translation to constituents.

- *Native language illiteracy*: The assumption driving the translation effort is that ELL students will better demonstrate their mastery of content standards using a native-language assessment. However, many ELL students may be less literate in their native language than they are in English. Many have received no content instruction in their native language. An important principle for the assessment of exceptional students states that the testing context should be as close as possible to how instruction is received. In cases where ELL students are less literate in their native language, the English assessment may represent the most valid assessment of their reading and mathematics achievement, especially for those most at risk of failure.

[2] While these examples alone do not guarantee increased validity for the accommodated assessment, taken together they represent cumulative evidence of greater validity.

What Does Research Say about Current Practices in Assessment of ELLs on Core Academic Content?

As noted in the discussion above, simple translation of assessments into various native languages creates several challenges for state efforts to meet NCLB requirements for the assessment of ELL students. In effect, it represents a rather naïve approach to the problem of increasing reliability, validity and accessibility of core academic assessments to students not yet fully fluent in English. Fortunately, a growing body of research is available to inform ELL assessment practice. Advances may benefit testing programs interested solely in the translation approach as well as those adopting a more comprehensive strategy that may involve, in part, some degree of translation or adaptation along with the use of accommodations and test design principles that have shown some promise. In particular, the work of Jamal Abedi at CRESST and Ron Hambleton and Steve Sireci at the University of Massachusetts provide excellent guidance. Various international studies of academic achievement are also useful resources for states and districts as they design appropriate practices for ELL assessment. International studies, such as the ones for Trends in International Mathematics and Science Study (TIMSS) address comparability issues across various language versions of a test on a routine basis.

Research on Accommodations

Accommodations refer to changes in test administration procedures or formats designed to increase accessibility of the assessments to various student populations. With the notable exception of translation, many of the accommodations available to ELL students were originally developed for students with disabilities (Liu, Anderson, & Thurlow, 1999). Prior to NCLB, ELL students did not experience the same legislative support for accommodations as students with disabilities. Nor were their results so prominently a part of formal accountability systems. Current assessment practices of ELL students include different kinds of accommodations: providing students with extra time, bilingual dictionaries, glossaries, or translators; modifying the language of the test; or translating the test from English into another language.

Research indicates that certain types of accommodations improve the performance of ELLs. Studies done by CRESST researchers demonstrated that ELL students improve their assessment scores when provided extra time and a glossary of key terms (unrelated to content) or when provided with an assessment with simplified language (Abedi, 1999). While such accommodations may lead to performance improvements, care must be taken to ensure that they allow for fair and accurate measurement of targeted skills and not overcompensate for lack of English proficiency. That is, do scores mean the same thing (i.e., measure the same constructs) for non-ELL students who take a given assessment without accommodations as for ELL students who take that same assessment with accommodations?

As with special education students, research on the use of accommodations for ELL student populations needs to focus on the concept of construct invariance, i.e., whether the content being assessed (e.g., English/Language Arts, mathematics) remains unchanged when assessed under altered circumstances. The accommodation is designed to level the playing field, not to favor the target audience(s). Such methods as differential item functioning (DIF) and factor analysis can be important tools in ensuring the validity and comparability of accommodated assessments.

Research on Translation and Test Adaptation

As indicated above, many states are considering translation as an accommodation for ELLs, even though less than 10 states have translated tests from English into another targeted language. Test translation is fraught with difficulties. Translating tests from one language to another language often introduces inaccuracies and biases. A translated test is often qualitatively different from the original, such that test scores are not comparable across the two language forms of the test (Sireci, 1997; Kester & Pena, 2002).

Studies of state and international assessments show that rigorous statistical checks are needed to ensure the quality of item translations (van der Linden, 1998). Translated items may function differently from the original versions even with careful adherence to translation guidelines and methodical review by content and linguistic experts. Psychometricians use DIF and other techniques to examine the empirical response distributions of items in populations with different languages. Items that perform differently across language populations are modified or dropped from the assessments. DIF analyses may be used both to measure bias across subpopulations taking the same English form of the test as well as investigating bias across different language versions of the assessment.

Many researchers believe that word-for-word translations are insufficient. They prefer the broader concept of *test adaptation*, an extension of test translation that makes explicit adjustments to test items for cultural experience, content and wording (Stansfield, 2003). Adaptation of assessments is appropriate when the new target population differs substantially from the original population with which the measure is used in terms of culture, language or country (Geisinger, 1994).

To assist test developers and practitioners, the International Testing Commission (ITC) developed a set of Guidelines to standardize translation practices and make cross-cultural psychometric tests comparable, equitable and valid (Hambleton, 1996; Hambleton & Kanjee, 1997). Twenty-two guidelines arranged into four general categories—Context; Instrument Development and Adaptation; Administration; Documentation—were created. See the box below for the full set of guidelines.

[insert translation guidelines box]

Research is underway on use of *dual language test booklets* as an alternative to basic test adaptation. Use of dual language test booklets strives to balance the often-competing

goals of test fairness and score comparability. It involves formatting of test booklets such that the original English assessment items are placed on one side of the booklet and the corresponding translated items are placed onto the facing pages. In one study, Duncan, Parent, Chen, Ferrara and Johnson (2002) studied the effects of using a dual language test booklet in which test items were presented in both English and Spanish. Using National Assessment of Educational Progress (NAEP) grade 8 math items, the study compared student performance on an English-only form with the dual language Spanish-English form. It used *forward translation* of the original English booklet (i.e., translation from English to Spanish) and *back translation* of the Spanish version (i.e., translation from Spanish to English) as a check to ensure the accuracy and appropriateness of the Spanish translation.

The results showed that students with higher levels of English proficiency scored significantly lower on the dual language form as compared to the English-only form. No differences in native Spanish speakers' test performance across forms were found after accounting for both English proficiency and the language students used to answer the test questions (i.e., whether students actually used the Spanish-language accommodation). This finding suggests that the dual language and English-only test booklets are psychometrically equivalent. Interestingly, all students in the study reported that the dual language test booklet was beneficial.

Citing the previous work by researchers that suggested psychometric equivalence of dual language test booklets with English-only booklets, Sireci and Khaliq (2002) conducted several different statistical analyses to examine the structural comparability of an English and English-Spanish version of a statewide mathematics test. Having found slight differences in the structure of the test across the two versions, they concluded that more research on dual language test forms is needed to better gauge the practical benefits and limitations of this test accommodation option.

Solano-Flores and Trumbull (2003) expands the notion of dual language test development by proposing a *concurrent assessment development* model that entails simultaneous development of a test in two languages. They advocate a new paradigm to guide ELL research and testing that more systematically addresses the complex nature of language and its relationship to culture. In their approach, two teams of test developers work interactively and in concert to produce two language versions of the same assessment. The forms evolve together and undergo the same number of item reviews and tryouts. Another unique feature of their approach is the use of generalizability theory (i.e., a psychometric theory of measurement error—see Brennan [2001]) to compare ELL student performance on the same items in both English and their native languages. Their results demonstrated that ELL student performance varies considerably across the different items and languages, such that some ELLs perform better on some items in one language and better on other items in the other language. As such, they conclude that valid assessment of ELL students requires a more nuanced understanding of the interactions among first and second language proficiency as well as the linguistic and content demands of test items.

A Comprehensive, Multi-Faceted Solution to the Problem

In assessing ELL student proficiency on core academic content, states should pursue a comprehensive solution that entails more than individual decisions about specific accommodations and test translations. Such a solution involves an iterative series of decision points with careful attention to technical, political, financial and practical considerations. Summarized below is an overall strategy to assessment development, administration and research that is designed to maximize the reliability, validity and accessibility of core academic assessments for ELL students.

- *Determine all relevant factors as early as possible.* The earlier in the test development cycle that content-based, structural and logistical factors related to assessment of ELL students are addressed, the more options are available to ensure reliability, validity and access. For example, attention is being drawn to the principles of *Universal Design* (Johnstone, 2003; Thompson, Johnstone & Thurlow, 2002; Thompson, Thurlow & Maloof, 2004) as a tool to improve access to a broader range of students. With Universal Design, test developers include items and develop procedures upfront during test development that reduce the need for post-hoc accommodations or adaptations. For example, a careful review of the vocabulary included on mathematics exams' word problems may identify instances where simple substitution of words used in the problem (e.g., car for automobile) may be an effective tool for increasing access for many students, including ELLs. Certain page layouts and computer-assisted test administration formats and procedures can also improve access for the full range of students. Research on Universal Design options is underway across the nation. Other upfront considerations that are likely to benefit at-risk students include ensuring that assessed content focuses on the most important components of state content standards, and developing item and test review procedures (statistical and human) that examine differences across student groups. Differences in performance should reflect actual achievement gap, not bias of or access to the means of assessment.

- *Fully identify needs and available resources.* The key to identifying and implementing practical solutions is to understand the extent of the problem. If a state has one primary second-language group, then the option for translation or adaptation may be more realistic as compared to a state in which there are many different groups of language-minority students. Similarly, understanding the full budget needed for adaptation, including activities beyond the mere translation process, will focus attention on areas of greatest need and payoff. For example, states may decide to use a strategy of test adaptation for its primary language group and accommodations for all other students.

- *Start with the most efficient options and expand, only if necessary and practical.* States should begin by applying Universal Design principles and expanding to accommodations and adaptations only if such extra efforts would significantly

increase reliability, validity and accessibility of ELL students to the assessment. In many cases, a strategy focusing on Universal Design and specific accommodations may be successful in ameliorating a substantial degree of the challenges of assessing the core academic knowledge of ELL students. States should allow accommodations already available for special education students to be made available to a broader range of students. The application of Universal Design principles and incorporation of expanded accommodation options may be sufficient to address the assessment needs of ELL students. All test development and administration activities should be reviewed regularly to ensure they provide maximal accessibility for ELL and other student groups.

- *Prepare a research agenda to study the reliability and validity of proposed options and alternatives.* While there is much to learn from the experiences of other states and assessment programs, each situation has idiosyncratic components that must be addressed. Thus, all testing programs, including local, state or national have a responsibility to demonstrate the technical adequacy of all intended assessment uses (AERA/APA/NCME, 1999). States and assessment developers must design and implement comprehensive studies that demonstrate that their approach for assessing ELL students can stand the same technical scrutiny as their program for all other students, whether the model selected includes accommodation, translation, adaptation or combinations of each.

- *Follow all standards and guidelines*. Should a state or testing program make the decision to move ahead with the translation or adaptation process, it must be prepared to do this right. Careful adherence to the AERA/APA/NCME standards and the ITC guidelines is essential if the program is to be able to defend its model for assessing different student populations, and more importantly, to increase the reliability, validity and accessibility of these important content-based assessments for ELL and other at-risk students.

Conclusion

NCLB has raised the stakes of assessments for **all** students, most notably those who have traditionally failed to meet standards or have been excluded from the lens of accountability. Instruction that is guided by student-centered data that include performance on content-aligned assessments is key to ending the achievement gap that prompted the enactment and implementation of NCLB. Too often, assessment data for at-risk students is not sufficiently reliable and valid for the complex purposes for which we rely on them. Thoughtful awareness of the challenges and careful adherence to the comprehensive approach detailed in this paper can help ensure that the data used to drive instructional decisions reflect the core academic achievement of ELL and other student populations.

Guidelines for Adapting Educational and Psychological Tests
International Testing Commission

Context

C.1: Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.
C.2: The amount of overlap in the constructs in the populations of interest should be assessed.

Instrument Development and Adaptation

D.1: Instrument developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended.
D.2: Instrument developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended.
D.3: Instrument developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.
D.4: Instrument developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.
D.5: Instrument developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.
D.6: Instrument developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument.
D.7: Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations.
D.8: Instrument developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.
D.9: Instrument developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.
D.10: Non-equivalent questions between versions intended for different populations should <u>not</u> be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately.

Administration

A.1: Instrument developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.
A.2: Instrument administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.
A.3: Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended.
A.4: Instrument administration instructions should be in the source and target languages to minimize the influence of the unwanted sources of variation across populations.
A.5: The instrument manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the instrument in a new cultural context.
A.6: The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for the instrument should be followed.

Documentation/Score Interpretations

I.1: When an instrument is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.
I.2: Score differences among samples of populations administered the instrument should not be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.
I.3: Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.
I.4: The instrument developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the instrument, and should suggest procedures to account for these effects in the interpretation of results.

References

Abedi, J., (1999). *Examining the Effectiveness of Accommodation on Math Performance of English Language Learners.* Paper presented at the annual Meeting of the National Council on Measurement in Education, Montreal Canada.

Abedi, J., (2004). The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues. *Educational Researcher*, 33(1), 4-14.

Abedi, J., & Dietel, R. (2004, June). Challenges in the No Child Left Behind Act for English Language Learners. *Phi Delta Kappan, Volume 85* (Number 10), 782-785.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological testing.* Washington, D.C., AERA.

Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag.

Duncan, T., Parent, L., Chen, L., Ferrara, S., & Johnson, E. (2002). *Study of a Dual Language Test Booklet in 8th Grade Mathematics.* Presented at a paper session, *Validity of Assessments for Linguistic Minorities and Students with Disabilities,* at the annual meeting of the American Educational Research Association, New Orleans, LA.

FUSD. (2004). *Fresno Unified School District 2002-03 Facts and Figure*. http://www.fresno.k12.ca.us/facts.html

Geisinger, K. (1994). Cross-cultural Normative Assessment: Translation and Adaptation Issues Influencing the Normative Interpretation of Assessment Instruments. *Psychological Assessment, Volume 6* (Number 4), 304-312.

Hambleton, R. (1996). *Guidelines for Adapting Educational and Psychological Tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Hambleton, R., & Kanjee, A. (1997). Enhancing the Validity of Cross-cultural Studies: Improvements in Instrument Translation Methods. In T. Husen & T.N. Postlewaite (eds.), International Encyclopedia of Education (2nd edition). Oxford: Pergamon Press

Jamentz, K., (2004). *Accountability Dialogues: School Communities Creating Demand from Within.* San Francisco: WestEd.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Kester, E., & Peña, E. (2002, July). Limitations of Current Language Testing Practices for Bilinguals. *ERIC/AE Digest Series EDO-TM-02-03,* 3-4.

Liu, K., Anderson, M., Swierzbin, B., & Thurlow, M. (1999). *Bilingual Accommodations for Limited English Proficient Students on Statewide Reading Tests: Phase 1*, State Assessment Series, Minnesota Report 20.

Sireci, S. (1997). Problems and Issues in Linking Assessments Across Languages. *Educational Measurement: Issues and Practice, Volume 16*(1), 12-19, 29.

Sireci, S., & Khaliq, S. (2002). *An Analysis of the Psychometric Properties of Dual Language Test Forms.* Presentation at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Solano-Flores, G., & Trumbull, E. (2003, March). Examining Language in Context: The Need for New Research and Practice Paradigms in the Testing of English-Language Learners. *Educational Researcher, Volume 32* (Number 2), 3-13.

Stansfield, C. (2003). Test Translation and Adaptation in Public Education in the USA. *Language Testing*, v.20, #2, pp. 189-207.

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. http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.ht

Thompson, S.J., Thurlow, M.L., & Maloof, D.B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, downloaded October 10, 2004 from http://www.testpublishers.org/atp_journal.htm.

van der Linden, (1998). A discussion of Some Methodological Issues in International Assessments. *International Journal of Educational Research, Volume 29,* 569-577.