A Practical Guide to Standards-Based Assessment in the Native Language

by

Melissa Bowles
University of Illinois at Urbana-Champaign

and

Charles W. Stansfield
Second Language Testing, Inc.

January 2008

TABLE OF CONTENTS

Acknowledgements ____ *iv*

Preface ____ *v*

1. Introduction: Why create a native language version of a test? ____ *1*

2. What are the different kinds of assessment in the native language? ____ *2*

3. What are the advantages and disadvantages of translation, adaptation, and parallel development? ____ *3*

4. What states have the most experience in using assessments in the native language? ____ *4*

5. In what languages have states provided (written) assessments in the native language? ____ *6*

6. What has been learned from the experience of states using assessments in the native language? ____ *6*

7. In what circumstances is the use of written assessments in the native language appropriate? ____ *7*

8. In what circumstances is the use of written assessments in the native language not appropriate? ____ *7*

9. What other options exist? (Oral options for ANL) ____ *8*

9.1 Recorded Audio Translations (AKA Scripted Oral Translations) ____ 8

9.2 Sight Translation ____ 9

10. What additional costs are involved for the state in using assessments in the native language? ____ *11*

11. When do the numbers justify the cost? ____ *14*

12. How does one deal with within-language differences? ____ *17*

13. How do states contract to carry out the translation or transadaptation of assessments? ____ *18*

14. Who can score the native language version? ____ *19*

15. For translation or adaptation, what are the advantages and disadvantages of bilingual test booklets versus a separate monolingual test booklet? ____ *19*

16. Which content areas are most amenable to assessment in the native language? For which content areas have states provided (written) assessments in non-English languages? ____ *20*

17. What effect can assessment in the native language have on reliability, validity, and score comparability? ____ *21*

17.1 Effect on Reliability ____ 22

17.2 Effect on Validity ____ 23

17.3 Effect on Score Comparability ____ 24

18. What effect does a decision to create a native language version have on the test development process? ____ *24*

19. What are possible political issues associated with assessment in the native language? ____ *24*

20. What quantitative evidence could a state provide to demonstrate that a translated assessment is comparable to the English version? _____ *26*

21. What qualitative evidence could a state provide to demonstrate that a translated assessment is comparable to the English version? _____ *27*

References _____ *29*

Acknowledgements

The authors wish to express appreciation to the following individuals for their help with this document.

Ms. Kathleen Leos, director of the Office of English Language Acquisition of the US Department of Education, was supportive of the development of such a document. Dr. Ed Roeber of Michigan State University made helpful comments on the outline of the document and reviewed the final version. Dr. Lynn Shafer Willner of George Washington University provided updated information for Tables 1 and 2 in this document. Dr. Ray Fenton, a former district testing coordinator, Ms. Gabriela Sweet and Dr. Hossein Farhady of Second Language Testing, Inc., Dr. Ana Maria Velasco, professor of Translation at the Monterey Institute for International Studies, Ms. Marijke van der Heide, former director of the Federal Court Interpreter Certification Program, Dr. Frances Butler, a language education researcher formerly of UCLA/CRESST, and Dr. Alexis Lopez of the University of Los Andes in Bogota Colombia all reviewed the entire document and made many helpful comments. Ms. Kathryn Doherty, coordinator of the LEP Partnership of the US Department of Education and Dr. Stanley Rabinowitz of the Assessment and Accountability Comprehensive Center at West-Ed reviewed the final version.

In spite of the contributions of the above individuals any errors of fact or opinion are our own.

Preface

Because there are many issues associated with the assessment of English language learners (ELLs), the US Department of Education has recognized that many states have a number of questions concerning the testing of ELLs, which is required under the *No Child Left Behind Act*. As a result, it established the LEP Partnership. The purpose of the LEP Partnership is to provide additional technical assistance to states to help them address these issues. The LEP Partnership brings together state department of education professionals (from their Title I, Title III, and assessment offices) several times per year to hear presentations by national authorities concerning testing issues related to ELL inclusion in the assessment program, and to provide an opportunity for informal discussions, interactions, and exchange of information.

In addition, the US Department of Education has funded the development of a series of monographs related to the improvement of ELL assessments and assessment policies. These monographs are funded through ED's Assessment and Accountability Comprehensive Center, which is operated by West-Ed, a research, development, and technical assistance provider located in San Francisco, California. West-Ed, in turn, has contracted with relevant organizations and individuals to write these monographs.

This guide is intended to raise the awareness of native language assessment (NLA) by those involved in state assessment programs and to lay out the pros and cons of the various kinds of native language assessments so that each state can make appropriate decisions for its populations of ELLs. It is presented in an easy-to-read frequently asked questions (FAQ) format so that states and districts can scan through the document and quickly find answers to the questions they face regarding assessment in the native language. The intended audience is educators, particularly those involved in decisions at the state level regarding appropriate accommodations for ELLs taking state-wide assessments. The document is not directed to translators or translation managers, although it does contain a good deal of information that would be of benefit to a translator or translation manager who is about the take on the task of preparing a non-English version of a state assessment. The document is designed to be practical; it is not a review of the literature on test translation. It is based on the authors' experience monitoring what takes place in the field.

This document is organized around 20 questions, which are followed by answers of varying lengths. The document discusses the reasons for using a written test in the student's native language to assess his or her content knowledge. Then, it discusses different approaches to native language assessment and the pros and cons of each. It then reviews what states have done in this regard over the past decade and what has been learned from their experiences. Subsequently, the document treats the effects of native language assessment on reliability and validity. Returning to practical matters, we cover the effect of a decision to create a non-English version on the test development process. Since states are often asked to justify that any special accommodations on standardized tests do not affect the comparability of scores across groups, the document discusses the kinds of evidence that can be presented.

The authors are solely responsible for the information in this document. The document does not constitute US Department of Education policy or opinion, nor does it constitute the opinions of

the Assessment and Accountability Comprehensive Center. No endorsement by either organization of any statement in this document should be implied.

1. Introduction: Why create a native language version of a test?

To a far greater degree than their native English–speaking peers, English Language Learners (ELLs) must process the language of tests and negotiate the cultural expectations embedded in assessments. For ELLs, every test becomes (at least in part) a test of language proficiency. The *Standards for Educational and Psychological Testing* [*Standards*] indicate that any test that employs language is, at least in part, a measure of language skills (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), (1999). Therefore, ELLs' "test results may not reflect accurately the qualities and competencies intended to be measured" (AERA et al., 1999, p. 91).

The *No Child Left Behind* (NCLB) *Act* of 2001 requires that all schools and all students be assessed annually in grades 3-8 and high school to demonstrate adequate yearly progress. One of the main requirements of NCLB is to include all students, even those who have been exempted from statewide assessments in the past, such as students with special needs and students with limited English proficiency. NCLB provides clear instructions regarding those populations in Section 1111(3)(C), requiring states to test limited English proficient students with "assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English language proficiency." The use of tests in a language other than English is permitted for a period of three years, but the NCLB allows local schools to extend the testing in a non-English language for an additional two years if the student's lack of English proficiency would impede the tests in English from yielding valid and reliable results (US Congress, *No Child Left Behind Act*, 2001).

Moreover, a summary guidance document for states recently issued by the US Department of Education specifies: "If native language assessment [testing in the native language (NLA)] is practicable, and if it is the form of assessment most likely to yield valid results, then a State must utilize such assessments." That is, a student's limited proficiency in English should not be a barrier to assessing his or her skills and abilities. If the student is unable to demonstrate his or her abilities because of a lack of proficiency in English, the assessment should be given in the student's native language, if possible.

Legislation is not the only reason states are interested in creating assessments in students' native languages, however. The availability of valid, reliable assessments of a student's knowledge and skills is also a psychometric issue. In the psychometric community, the false inflation or deflation of test scores due to measurement error (construct-irrelevant variance) is widely discussed. Specifically, attempts are routinely made to reduce construct-irrelevant variance stemming from the type of language used in assessments. The *Standards* further specify, "Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences" (AERA et al., 1999, p. 97).

Furthermore, since assessments written in English are unlikely to adequately test what ELLs know and can do, test scores may not be useful for course placement and instructional purposes. This lack of information about ELLs' knowledge and skills therefore causes pedagogical problems as well, since teachers do not have enough information to inform their classroom

practices and address students' needs. Offering versions of a test in the students' native language, then, is one way states can more accurately assess students' content knowledge, separate from their English language proficiency (Stansfield & Bowles, 2006).

With increasing enrollments of ELLs in US schools, teachers, administrators, and legislators alike are becoming interested in assessment in the native language. Indeed, native language accommodations are appropriate for ELLs in many cases, and many states have begun using either written or oral translations of assessments in at least one language other than English.

However, the literature on assessment in the native language (NLA) is scant, especially in the context of K-12 state assessment. This leaves many states not knowing where to turn for guidance about how to get started using assessments in the native language. Among the many issues states must consider are what languages they should offer; which of their assessments should (or can) be validly translated; whether a given population of ELLs will be more appropriately accommodated using written or oral translations; and how to best perform translations and ensure their comparability to the English versions. This guide addresses these and other related issues.

2. What are the different kinds of assessment in the native language?

There are three different kinds of assessment in the native language: translation, adaptation, and parallel development.

A *translated test* is one in which written test content originally in English is rendered into a written non-English language. The original English version of the test and the translated test differ only in the language, not in the content or the constructs intended to be measured.

Most translated tests normally require minor adjustments or adaptations to accommodate the language of the non-English version. For example, a math or science test translated to Spanish will reverse the use of commas and periods in the translated version (the number 10,215.64 is written 10.215,64 in Spanish). Such minor changes make the translated version more linguistically and culturally authentic without affecting score comparability. Without such minor adaptations, the language or other features of the translated version would not reflect natural usage or expression. Since these changes constitute a minor form of adaptation, a growing number of some people use the term *transadaptation* to refer to the standard direct translation of an assessment.

An *adaptation*, however, involves more substantial changes, such as the replacement of a number of items with others that are more appropriate for either the culture or the language of the new test (Stansfield, 2003). One can view adaptation as *modification* of the test. For example, in translating a multiple-choice test of English grammar to Spanish, Auchter and Stansfield (2001) found that approximately 50% of the items required either some modification, substantial modification, or a completely new item. Because this level of adaptation affects the ability to compare scores on the standard and adapted versions, such tests are most accurately characterized as adaptations rather than translations. The change in test content raises validity

concerns, especially if a substantial number of items are changed. As a result, it becomes necessary to demonstrate the equivalence of the constructs measured by the standard and adapted instruments. Because this process is long and expensive, adaptation is rarely used in state assessments in the US. Instead, tests in which validity and comparability may change if translated or adapted are normally not translated or adapted at all (Stansfield, 2003).

Parallel development is the least commonly used of the three kinds of NLA and usually involves a native language version of a test being developed concurrently with the English language version. The test content and specifications are similar, because they are based on the same content standards, but all items are developed separately in each language. For example, a parallel development of a math assessment would produce something approximating parallel forms in two different languages. While each item would be unique and original, the tests should exhibit similar validities, since they measure similar content.

Related terms. A variant of parallel development has been called *concurrent development* (Solano-Flores et al. 2002). This variant employs a specific framework to keep the tasks represented by the items on the two versions of the assessment as similar as possible. Tanzer (2005) has used the term *simultaneous development* to refer to a situation where bilingual item writers create items in one language and then immediately translate them to the other language. After comparing the two versions, adjustments are made in both to make them as equivalent as possible in the two languages. As used by Tanzer, simultaneous development is similar to direct translation or transadaptation, except that under simultaneous development each item is translated as the test is developed, instead of after the original test is completed, and items in each language can be modified to make them more equivalent in expression.

3. What are the advantages and disadvantages of translation, adaptation, and parallel development?

The advantage of translation (or transadaptation) is that it is the easiest and least expensive of the three forms of assessment in the native language; however, translation into many different languages can still be expensive. Also, as stated above, the content of the test and the construct measured do not change. Therefore, the interpretation of the test score is less threatened by translation than by other forms of NLA.

The advantage of adaptation depends on the need for it. When adaptation is needed, this technique can produce a more valid test. Adaptation usually begins with an examination of how the construct differs in the new language. For example, when adapting a test of automobile mechanics for use in France, it would be more appropriate to include more items related to knowledge of automobiles made specifically in France and more generally in Europe, and fewer items that deal with American automobiles.

Adaptation typically involves a situation where the knowledge that it is appropriate to assess is different for different populations, as illustrated in the example above. With the exception of a language arts test, this situation does not apply to a standards-based achievement test. When tests are based on a specific set of content standards, the interpretation of the score applies only to

those standards. Thus, there is no claim on the part of the test maker that the test score is defined broadly and without reference to a specific curriculum or set of content goals.

The disadvantage of adaptation is that the resulting test must be treated like a new test. It becomes necessary to field test new or revised items, to link the new test to the existing scale through the unchanged (merely translated) items, and to demonstrate the validity and comparability of the modified test. In some languages, however, there may not be a large enough sample of available field-test examinees to conduct this process.

The advantages of parallel development are that the test consists of entirely new items, originally written in the non-English language. Advocates for parallel development believe that this provides greater validity and that the resulting test has greater credibility. In parallel development, the non-English assessment is usually based on identical or similar content standards and test specifications. Otherwise, it is difficult to claim that the test is parallel.

The disadvantages of parallel development are closely linked to the advantages. The process of parallel development is as involved as the process of test development for assessments in English. Therefore, it is necessary to assemble a new test development committee, examine the test specifications and modify them as necessary, draft and review items, field test on the non-English population, and set cut scores that relate the test to performance or achievement levels. Like in the case of adaptation, there may only be sufficient numbers of available test subjects to permit field testing and statistical linking in a few languages. Linking studies are sometimes conducted to link the parallel non-English version to the English version, and in some cases new validity evidence is collected as well. There is also the need for new alignment studies, and perhaps score comparability and score interpretation studies. All of this involves cost and human resources. For these reasons, parallel development is almost always the most costly of the three options for native language assessment.

4. What states have the most experience in using assessments in the native language?

In school year (SY) 2006-2007, a review of state assessment practices revealed that twelve states offered written native language versions of their statewide assessments. As shown in the map on the following page, Delaware, Kansas, Massachusetts, Minnesota, Nebraska, New Mexico, New York, Ohio, Oregon, Rhode Island, Texas, and Wisconsin all offered written translations or adaptations of their assessments. Additionally, Texas and New Mexico offered parallel Spanish language versions of some statewide assessments.

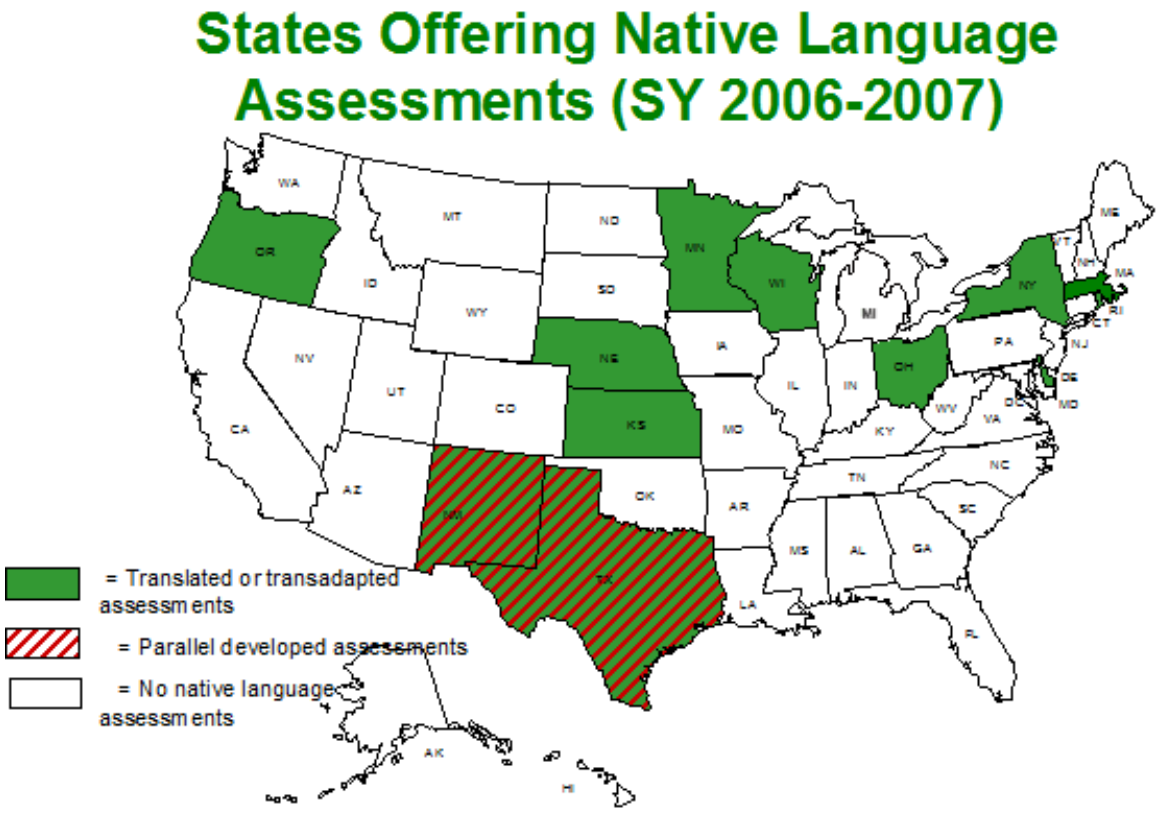

Figure 1. States Offering Native Language Assessments (SY 2006-2007)

NB. In Minnesota, native language versions are available for the high school graduation test only.

5. In what languages have states provided (written) assessments in the native language?

As Table 1 below shows, in SY 2006-2007, states provided written assessments in a total of eight non-English languages. The most common language of translation was Spanish; in fact, all twelve of the states that offered written translations of their assessments reported having Spanish language versions. Russian was the next most frequent language of written translation, available in both New York and Oregon, where there are large populations of Russian-speaking ELLs. There were written translations of statewide assessments in just one state for each of the remaining languages – Chinese (using the traditional characters employed in Taiwan and Hong Kong), Haitian, Hmong, Korean, Somali, and Vietnamese.

Table 1.
Languages for Which Written Translations of Statewide Assessments Were Provided (SY 2006-2007)

	Chinese	**Haitian**	**Hmong**	**Korean**	**Russian**	**Somali**	**Spanish**	**Vietnamese**
DE							✓	
KS							✓	
MA							✓	
MN			✓			✓	✓	✓
NE							✓	
NM							✓	
NY	✓	✓		✓	✓		✓	
OH							✓	
OR					✓		✓	
RI							✓	
TX							✓	
WI			✓				✓	
Total	**1**	**1**	**2**	**1**	**2**	**1**	**12**	**1**

6. What has been learned from the experience of states using assessment in the native language?

NLA is popular with teachers and students. No state that has started providing native language versions has stopped doing so because of complaints from teachers, students, or the parents of the students who take them. On the contrary, NLA allows the knowledge and skills of students who otherwise might have been exempted from statewide assessments to be measured. These results can be used to inform placement decisions and influence instructional practices.

7. In what circumstances is the use of written assessments in the native language appropriate?

Written translations are most appropriate for students who are (a) literate in the native language, (b) have had formal education in the home country/language, and/or (c) have been educated bilingually in American schools through a bilingual education program, but whose English language skills are not yet sufficient for testing in English. For these populations of ELLs, a written translation may be the best accommodation because it would be most similar to what those students would have received in their school in the home country.

8. In what circumstances is the use of written assessments in the native language not appropriate?

For students who are illiterate in their native language or who have not received formal education in their home country, written translations are not appropriate and will not yield reliable information about their knowledge and skills. For this reason, states must be careful not to make assumptions about ELLs' language backgrounds or formal education prior to immigrating to the United States. Rather than assuming that students will benefit from tests printed in their native language, states considering this option should use home language and academic background surveys to determine the students' level of literacy and educational experience.

An early experience of the Rhode Island Department of Education (RIDE) in offering written translations of statewide assessments provides a compelling example of problems that can occur when incorrect assumptions are made about ELLs' backgrounds. In 1996 Rhode Island began offering written translations of state assessments. RIDE identified the most common language backgrounds of ELLs in the state (Spanish, Portuguese, Khmer, and Lao) and chose those as the target languages of the translations. After administering tests in those languages the first year, the state discovered that the vast majority of the students from Cambodia and Laos were not literate in their native languages. Hence, it was determined only after the fact that written translation was almost never an appropriate accommodation for these students.

Literacy and previous education are not the only considerations states should keep in mind when discussing the possibility of written translation. Written translation is a cost effective accommodation if it can be provided to a large number of ELLs, so it is most appropriate in cases where there are a large number of ELLs from the same language background. If a state has small numbers of ELLs from many different language backgrounds, it would *not* be cost effective to provide written translations of the statewide assessments in each native language represented in the state.

Those who call for assessments to be translated to all languages or none at all are raising a barrier that can never be surmounted. Sometimes, those who make such calls may be opposed to non-English assessments altogether. There is no requirement that a state provide a written translation of its assessments to students. However, to the extent that a state does, it is demonstrating a commitment to linguistically fair testing for those ELLs who can benefit from it.

9. What other options exist? (Oral options for NLA)

A state may decide that a written translation (whether a translation, adaptation, or parallel developed assessment) is not appropriate for the reasons discussed above. However, the state can still provide another type of native language accommodation to those students. That is, they can assess the students by administering the assessments to them orally in their native language. These orally-presented assessments can be audio-recorded for standardized, on-demand administration, or they can be sight translated on a case-by-case basis. Fourteen states provided either audio-recorded or sight translated versions of statewide assessments in SY 2006-2007, as shown on the map on page 14. Many more states' policies provided for the translation of the directions into languages other than English, but only those that translated test items into another language are included in the count.

Both audio-recorded translations and sight translations have pros and cons, which are discussed below.

9.1 Recorded Audio Translations (AKA Scripted Oral Translations)

During SY 2006-07, just two states, Ohio and Michigan, opted to provide recorded audio translations (also knows as scripted oral translation) of assessments to ELLs. The Ohio Graduation Test (OGT) was provided in Spanish, Japanese, Somali, Korean, and Mandarin. For the assessments in grades 3-8, Ohio offered recorded translations in nine languages: Albanian, French, Japanese, Korean, Mandarin, Somali, Spanish, Ukrainian, and Vietnamese.

Michigan created DVDs with scripted oral translations of its statewide assessments in Spanish and Arabic, the most common languages of ELLs in the state.

Having the test recorded ensures standardization and eliminates variations between speakers, pauses, timing, and other extraneous factors that accompany a spontaneous sight translation. The main advantage of a recorded audio translation, therefore, is the guarantee of comparability from administration to administration across a district or state.

States or districts interested in producing recorded audio versions of assessments should follow a series of recommendations to ensure that the standardized oral administration is of the highest quality. First, the state or district should provide the script of the English audio version of the test to a professional translator to be translated. It is crucial that the translator receive the script of the English audio version (rather than the print version) because this script includes prompts and instructions that specifically tailor the test for recorded oral administration. This translation must be performed carefully and reviewed in the same way that a written translation would be. Once the translation has been finalized, it should be read aloud by a native speaker test moderator and then professionally recorded. When the recording has been completed, it should be compared with the script by an independent translator to verify that no item, option, or other material was inadvertently left out, and that all words are pronounced intelligibly.

Since producing a recorded audio translation of an assessment involves an iterative process of review and revision by professional translators, as well as labor costs for voice actors and sound engineers, its main detractor is cost. While ideally all oral test translations would be recorded, due to cost, audio translations are typically only provided in languages with substantial populations of speakers.

9.2 Sight Translation

Another oral translation option available to states is sight translation of the assessment. In sight translation, a translator/interpreter sits with the student who is taking the test and, looking at the English test, reads the test stimuli and items aloud in the non-English language. That is, the translator/interpreter must perform a simultaneous or on-the-spot translation for the student while s/he is taking the test. For a more detailed description of sight translation, including the challenges presented by this accommodation and recommendations for its implementation, readers are referred to *A Guide to Sight Translation*, published under separate cover.

Sight translation is an option that a state might choose if there were small numbers of speakers of a given language and creating a scripted oral translation would not be cost effective. Clearly, sight translation does not provide a standardized administration and there is variation inherent in the procedure. For this reason, it is not a preferred accommodation for ELLs (Stansfield & Bowles, 2006). However, if a state chooses to use this accommodation, a number of precautions, mentioned briefly below, and discussed more in depth in the separate guidance document, can help to ensure a high degree of accuracy for this accommodation.

Because it is difficult to provide an accurate, complete rendition of the test material on the spot in another language, the person selected to perform the sight translation should be an experienced translator or interpreter. Also, the person should be familiar with the test content and test administration procedures. Whenever possible, the sight translator should be allowed to study the test at least a day ahead of time in order to prepare for the sight translation.

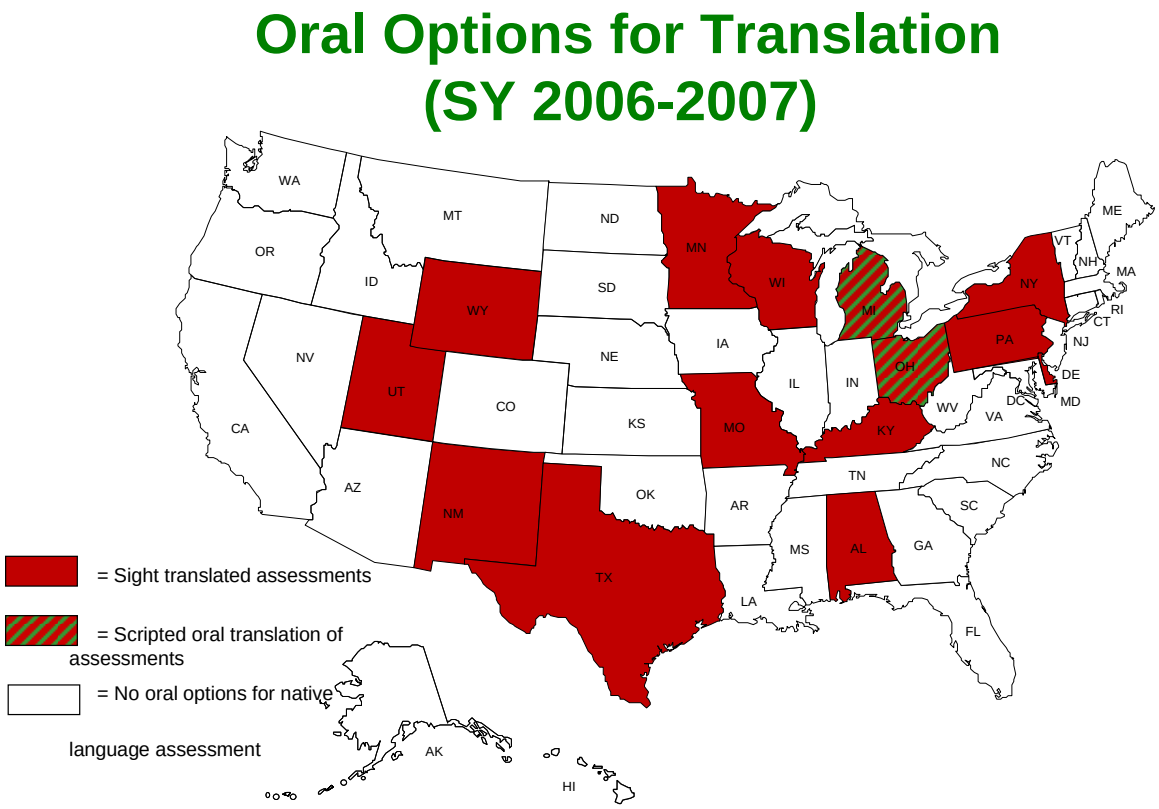

Figure 2. Oral Options for Native Language Assessments Used by States (SY 2006-2007)

10. What additional costs are involved for the state in using assessments in the native language?

It is difficult to accurately convey to the reader the costs associated with native language versions of state assessments. Considerable misunderstanding may exist concerning the cost, particularly in states that do not currently create translated versions of their assessments. Often, it is assumed that the cost is prohibitive. Claims are sometimes made that the cost is very high. In reality, test translation is far less expensive than is sometimes assumed, and claims that the cost is prohibitive are often made by those who are skeptical of such non-English accommodations for other reasons. On the other hand, the development and provision of a written translation does involve additional costs. However, when judging costs it should be remembered there is also a cost associated with provided a test booklet in English or parent guide in English. Thus, the costs of printing and scoring in non-English languages are partly counterbalanced by savings in the number of tests or test program-related documents printed or scored in English. The additional costs that may be incurred with using assessments in the native language are enumerated below.

When a state produces assessments in the native languages of the students, the cost of the translation or parallel development is not the only cost that will be incurred. States must also bear in mind the costs associated with printing, administering, and scoring the assessments in the native language.

If a state makes the decision to provide written translations of its assessments, it is making a substantial investment in terms of time and money as well as introducing a substantial change in its assessment system. Ancillary materials, such as test administration manuals, answer sheets, and score reports, which were developed for the English versions of the tests, almost certainly do not exist in the non-English languages for which the state is providing written translations. In the majority of states, a decision is made to translate some or all of these documents too.

In addition, if there are oral test administration instructions that must be read to the students, unless these have been audio recorded, they will need to be read by a test administrator who speaks the language in question.

Seven states—Arizona, Colorado, Delaware, Massachusetts, New York, Oregon, and Texas—translated at least some ancillary materials during SY 2000-2001. Some of those materials are available online through the state education agencies' websites, and links are provided in Figure 3.

Stansfield and Bowles (2006) advise that, at a minimum, states that provide a written translation of their assessments, should also translate ancillary materials that directly affect test administration. First and foremost, this means translating the standardized test administration procedures developed and carefully prescribed for the English test to ensure consistent administration across languages. Only the read-aloud portions of the directions for administration need to be translated. Secondarily, this means translating other documents students must use as

they complete the test, such as answer sheets or reference sheets (commonly used to provide formulas for mathematics tests).

By providing these documents in the non-English languages, states can help ensure that all students are receiving the same test under the same conditions, thereby helping to ensure score comparability on the original and translated versions of the assessments.

Test score reports and/or parent guides may be produced in non-English languages. Hawaii, Massachusetts and other states increasingly offer interpretive score report guides to parents in a range of non-English languages, beyond those in which translated assessments are provided. Providing a score report and/or an interpretive guide to parents is a positive assist to parents, and one that is envisioned by NCLB, which has provided the impetus for expanded and informative score reports for parents. Often parents exhibit less English language proficiency than their children, which makes a score report or parent guide in the native language especially helpful in conveying their children's progress in school. Some states, including those that do not provide translated versions of the assessments, translate the score reports to a substantial number of languages. For example, Massachusetts translates its parent guide to 10 languages and Hawaii translates it to 13 languages. There is a cost for translation, desktop publishing, and printing associated with doing this in each language.

Where there are constructed response items, hand scoring by native speakers of the language will be necessary, as will training for the scorers to ensure that they are fair and consistent in the way they assign points to the student response. The training may involve the identification of example responses in the non-English language. If native speaker scorers are not available, then student responses will have to be translated to English, and guidance must be created to ensure that the English translation is commensurate with the non-English response by the student. This all implies additional costs associated with non-English language versions of the assessment, apart from the costs of the actual translation of the assessment itself.

Once an assessment in a non-English language is administered, the state must/should examine the statistical quality of the test, just as it does with its English language versions. Therefore, states will incur additional expense by performing item analyses on the native language versions of the tests and assembling other qualitative and quantitative evidence that permits one to judge the comparability of the English and non-English versions.

FIGURE 3. Ancillary Materials in Languages Other than English Available Online

COLORADO
Parent guide in Spanish, *Una guía para los padres de familia*
http://www.cde.state.co.us/cdeassess/documents/parents/CSAP_Span.pdf

MASSACHUSETTS
Guide to the 2006 MCAS for Parents/Guardians
http://www.doe.mass.edu/mcas/pgguide.html
PDFs available in English, Cape Verdean Creole, Chinese, Haitian Creole, Khmer, Portuguese, Russian, Spanish, and Vietnamese

NEW YORK
Sample math tests in Chinese, Haitian Creole, Korean, Russian, and Spanish
http://www.emsc.nysed.gov/3-8/math-sample/home.htm

OREGON
Content standards
- Russian http://www.ode.state.or.us/search/page/?id=794
- Spanish http://www.ode.state.or.us/search/page/?id=791

Sample tests in Russian and Spanish
- Mathematics - http://www.ode.state.or.us/search/page/?id=441
- Science - http://www.ode.state.or.us/search/page/?=444
- Social sciences - http://www.ode.state.or.us/search/page/?=445

Glossaries of mathematics test item terms
- English/Russian
  http://www.ode.state.or.us/teachlearn/subjects/mathematics/assessment/knowledgeandskills/translatedterms_rus0607.pdf
- English/Spanish
  http://www.ode.state.or.us/teachlearn/subjects/mathematics/assessment/knowledgeandskills/translatedterms_spn0607.pdf

Scoring guides in Russian and Spanish
http://www.ode.state.or.us/search/page/?id=32

TEXAS
Information for parents about the statewide assessments in Spanish
http://www.tea.state.tx.us/student.assessment/resources/guides/parents/index.html

Explanations of score reports in Spanish
- Grades 3-6 Test
  http://www.tea.state.tx.us/student.assessment/resources/guides/parents/span3thru6.pdf
- Grades 3-8 Test
  http://www.tea.state.tx.us/student.assessment/resources/guides/parents/span3thru8.pdf
- Exit Level Test
  http://www.tea.state.tx.us/student.assessment/resources/guides/parents/spexitlevel.pdf

11. When do the numbers justify the cost?

Providing native language accommodations (written translations, audio-taped translations, or sight translations) costs states and districts additional expense. As previously discussed, the iterative process of review and revision and the costs of printing or audio recording the test in a non-English language make written translation and audio-recorded translations costly. Because of the cost, most states offered written or audio-taped translations for just the most populous non-English language backgrounds, and most often only in Spanish.

However, the decision about whether the numbers justify the cost of providing translated versions of assessments is an individual one. Population is an important factor that states consider. However, some states with small ELL populations provide NLA whereas some states with larger ELL populations do not. The state survey revealed that written translations were provided by states ranging drastically in population, from Delaware, with the smallest number of ELLs (just 5,094 ELLs statewide in SY 2004-2005, accounting for less than 4.2% of the school population), to Texas, SY 2004-2005 with over 684,007 ELLs, accounting for about 16% of the school population.

For the twelve states that offered written translations of their statewide assessments in 2006-2007, Table 2 lists total K-12 enrollment in the state and total ELL enrollment as a frequency and as a percentage of the total enrollment. Note that these demographic data are from SY 2004-2005, the most recently published information available from NCELA at the time of writing. In addition, the table provides the languages of translation and the top 5 non-English languages spoken in the state (based on the most recent available NCELA data, from SY 2001-2002). The Wisconsin data is from 2006-07 and includes only the top 3 languages.

Table 2. Written Translation of Statewide Assessments and Number of ELLs (SY2006-07)

State	Total enrollment	ELL enrollment	% ELL	Top 5 languages spoken by ELLs	%	Languages of translation
DE	119,038	5,094	4.3	Spanish	77	Spanish
				Haitian Creole	4	
				Chinese (includes Cantonese and Mandarin)	2	
				Korean	2	
				Arabic	1	
KS	445,941	23,512	5.2	Spanish (00-01 data for the 5 languages)	81	Spanish
				Vietnamese	4	
				Lao	2	
				Chinese Cantonese	1	
				Korean	1	
MA	975,574	49,923	5.1	Spanish	69.4	Assessments – Spanish only *Ancillary materials in top 10 most common languages
				Portuguese	10.0	
				Khmer	5.1	
				Vietnamese	4.9	
				Haitian Creole	3.0	
MN	838,503	56,829	6.8	Hmong	34.1	Hmong Spanish Somali Vietnamese
				Spanish	28.3	
				Somali	6.6	
				Vietnamese	4.4	
				Lao	3.6	
NE	285,761	16,124	5.6	Spanish	89	Spanish
				Vietnamese	4	
				Nuer	3	
				Arabic	3	
				Kurdish	1	
NM	317,00	70.926	22	Spanish	78.8	Spanish
				Navajo	14.6	
				Vietnamese	0.5	
				Arabic	0.1	
				Russian	0.1	

State	Total enrollment	ELL enrollment	% ELL	Top 5 languages spoken by ELLs	%	Languages of translation
NY	2,858,500	203,583	7.1	Spanish	62.2	Spanish Russian Chinese Haitian Creole Korean
				Cantonese	5.2	
				Russian	3.0	
				Chinese (unspec)	2.7	
				Urdu	2.7	
OH	1,847,116	25,518	1	Spanish (00-01)	39.2	Spanish
				Arabic (00-01)	8.2	
				Somali (00-01)	8.0	
				PA Dutch (00-01)	5.4	
				Japanese (00-01)	4.9	
OR	552,342	59,908	10.8	Spanish	69.8	Spanish Russian
				Portuguese	6.7	
				Cape Verdean	4.9	
				Khmer	2.5	
				Lao	1.8	
RI	156,498	10,921	7.0	Spanish	69.8	Spanish
				Portuguese	6.7	
				Cape Verdean	4.9	
				Khmer	2.5	
				Lao	1.8	
TX	4,405,215	684,007	16	Spanish	93.4	Spanish
				Vietnamese	1.9	
				Cantonese	0.7	
				Urdu	0.5	
				Korean	0.4	
WI	874,098	39,678	5	Spanish	48.2	Spanish Hmong
				Hmong	24.1	
				Other	27.7	

Clearly, the decision to offer NLA is based on multiple factors, including not only the density of each language group but also the groups' perceived literacy, educational background prior to enrolling in US schools, and the availability of bilingual education programs for speakers of the non-English language. In addition, legislation regulating access to language services in the state also play a role. For instance, despite its relatively high population of ELLs (15.4% as of SY2000-2001), Arizona no longer offers translated versions of its statewide assessments due to Proposition 203 (English Language Education for Children in Public Schools), which was adopted in November 2000 and took effect in SY 2001-2002. The bill established English as the official language of Arizona and required that all public school education in the state be conducted in English. It provides that students with limited English proficiency be placed in a structured English immersion program for not more than one year, at which point they are mainstreamed into English classrooms. The bill also mandates that all students take statewide

achievement tests in English, regardless of their English proficiency. Similarly, in 2002, Massachusetts voters approved the *English for the Children* Initiative. This initiative replaced a long-standing state law that allowed ELLs to be placed in transitional bilingual education in public schools. Under *English for the Children*, all public school children must be taught all subjects in English and must be placed in English language classrooms as soon as they enroll in Massachusetts schools, regardless of their language proficiency.

Texas law, on the other hand, provides for bilingual education (English-Spanish) in grades 3-6 and mandates that statewide assessments be available in Spanish at those grade levels.

On the matter of translating score reports and parent guides, in addition to considering numbers, states should consider the likelihood that parents will be literate in their native language. When making such decisions, consulting with the state ELL specialist and community representatives can be helpful. It may not be cost effective to translate a parent guide if parents are unlikely to be literate in their native language.

12. How does one deal with within-language differences?

It is sometimes alleged that people from different parts of a country or a linguistic region of the world speak so differently that good translation to their language is impossible. While within-language differences are a reality, it is also true that speakers of the same language have a great deal in common in the way they express themselves. If they did not, they would not be speaking the same language.

For most languages the written version follows fairly fixed conventions. Written language can be so different from speech that linguists make a distinction between speech and written language, sometimes referring to them as oracy and literacy. The conventions of literacy apply to written language, except when one wants to write in a way that imitates oral language. Written language is normally more formal than speech. Linguists employ the term *register* to refer to the degree of formality that is used in oral or written expression by a language user in a specific context. Register varies from highly formal to highly informal. Because a test puts an examinee in a formal situation, tests almost always employ formal register. The language of a test in English is straightforward and polished. The language of the native language version of the test should be similarly straightforward and polished. It must be carefully constructed to match the register and readability level of the English version, while conveying exactly the same meaning as the English version. Accomplishing this means that one has to use very standard language. The use of non-standard language (slang, argot, and idioms typical of low register) is to be avoided on a test, unless these are used in the English version. The fact that the reader may speak in low register is irrelevant to the translation of a document. The goal is to create a parallel document in the native language. The use of formal register in an assessment means that regional usage will be bypassed for standard usage. Thus, because of the formal nature of test language, a single translation using standard formal language is quite feasible, particularly in subjects like math and science.

It is sometimes alleged that because Spanish is the official language of 20 different countries and the second or third most widely spoken language in the world, a good Spanish translation is not possible. This is a gross exaggeration and an incorrect conclusion. Literature, magazines, textbooks, and newspapers produced in all these countries are can be read with ease across countries. In cases where different words are preferred to express a concept in one or more countries, there is always a multinational or neutral word or phrase that can be used to express the concept or idea.

The linguistic situation of spoken Arabic, which is the official language of 22 countries, is much more complex and varied, when compared to that of spoken Spanish. Still, the Arab world shares a common literature and common conventions in the use of the written language. Thus, successful written communication across dialects is the norm.

Every language has a diverse group of speakers, and translation poses different problems for each language. It is the responsibility of the translator to be sensitive to these differences so that communication with any competent reader will be successful. While all translators are not equal in their skills, it must be recognized that very good translators produce very good translations every day. Thus, the successful translation of a test is quite feasible and can be expected of a translation contractor. Within language differences are almost never a valid reason for not producing a version of the test in a student's native language, so long as the student is literate in the language.

13. How do states contract to carry out the translation or transadaptation of assessments?

States frequently contract with the test publisher that provides their standard (English-language) assessments for the non-English versions. However, since test publishers do not have well-developed item banks in languages other than English, they are not well equipped to provide assessments in languages other than English. For this reason, most test publishers subcontract with translation companies or other organizations to ensure that the translated or adapted assessments are of similar quality to the English assessments. Sometimes states contract for the non-English versions directly, and manage the project from the student assessment office.

Whether or not states decide to contract directly with their main testing contractor, the translation or adaptation should be carried out by people with expertise in translation, testing and item writing due to the precise nature of the wording needed in test items. Furthermore, states should insist that the translation or adaptation be carried out by individuals who are *educated native speakers* of the target language and have relevant content area expertise.

After the translated versions of the assessments have been produced, states should request that at least one independent review of the translation be carried out. At a minimum, this independent review should be done by the organization doing the translation before it is submitted to the state. In addition, the state may request another independent review by a separate organization, or it may assemble a committee of teachers or community members to review the translation and provide feedback for revision and finalization of the translation. Such independent reviews are

quality control measures by which the original and translated assessments are compared by translators and/or content area experts who have a strong command of English and the target language. The individuals doing the review identify any test items that are mistranslated or that could have a different interpretation in the target language than in the original, and propose revisions to the wording of items to make them comparable to the English versions.

14. Who can score the native language version?

In most cases, the same testing company that provides the scoring service for the regular English version will provide a scoring service for the non-English version, if that version contains student responses to constructed response items. For major languages like Spanish, the test scoring contractor is able to do this by recruiting and training bilingual scorers. These same scorers also score English versions of a test, such as Mathematics or Science, when not scoring a non-English version. This is probably the best situation for scoring that can be obtained, as the scorers used are subject to the same selection criteria (such as a teaching credential or other knowledge of the subject area) and undergo the same training as all other scorers. Thus, the ELLs' native language answer document is scored by raters who are just as qualified as the scorers who score the English version.

When qualified bilingual scorers are not available to score the student's responses, the test scoring contractor may pair a qualified scorer with a bilingual individual and together they will score the response. In this case the bilingual undergoes similar scorer training as the qualified scorer and orally renders the student response into English. Thus, each student's responses are scored by a qualified and trained scorer.

Another approach sometimes used is for a bilingual teacher, aide, or test administrator to translate the student's written response to English in the answer document following the administration of the test. Then the answer document is scored by one of the regular monolingual English speaking raters in the rater pool. This situation is somewhat less desirable, since the person doing the translation is not a professional translator and may not have appropriate proficiency in one or both languages. Also, the person may take liberties with the translation, and it may not be entirely faithful to the kind of response that the student made. In such situations, guidance in the form of policies should be developed at the state level, and then training should be provided at the district level on how to approach the translation of student responses.

An alternative to the written translation of student responses by building personnel is to have the test scoring contractor hires a translator to translate the students' written responses to English and then submit the answer books to the regular scoring process.

15. For translation or adaptation, what are the advantages and disadvantages of bilingual test booklets versus a separate monolingual test booklet?

Once a state has decided to offer a written translation of a statewide assessment, the state must decide on the format of the printed test booklet that will be given to the ELLs. There are two basic options: a monolingual test booklet, in which test stimuli and items are presented exclusively in the non-English language, and a bilingual (dual-language) test booklet, in which stimuli and items are presented in both English and the non-English language, side-by-side or in facing columns in the test booklet. Based on a number of studies that have been conducted to examine the effect of presentation format on ELLs (e.g., Garcia et al., 2002; Liu et al., 1999; Stansfield & Kahl, 1998), the bilingual test booklet format may be beneficial to many ELLs, who may wish to rely on the contextualizing information surrounding the items in their native language while referring to the English for terminology they may have learned in English in school. These studies have also found that even for ELLs who rely on just one language (either English or the native language) the bilingual format is not a hindrance. Also, the provision of a bilingual test booklet may reduce the onus on states to demonstrate the comparability of a translated assessment, since examinees with partial English language proficiency may glean some additional information from the English version of the test. For this reason, it is advisable for states to present written translations in a bilingual test booklet format.

A survey of states that provided written translations of statewide assessments in SY 2000-2001 (Stansfield & Bowles, 2006) revealed that four states used the bilingual test booklet format during that time period—Delaware, Massachusetts, Montana, and Oregon. Based on their experiences and on the research findings cited above, we recommend that other states follow their lead and begin to implement bilingual test booklets.

16. Which content areas are most amenable to assessment in the native language? For which content areas have states provided (written) assessments in non-English languages?

Some content areas are more amenable to translation than others. Certainly, concepts in areas such as mathematics, science, and social studies can be tested in translation without affecting the construct. Tests in these areas can be translated fairly straightforwardly. On the other hand, when tests of reading and writing are translated, the construct is altered (i.e., the test is no longer a test of reading or English Language Arts but rather a test of reading or grammar in the target language). As a result, adaptation is necessary, and often new items must be created for the non-English version of the test. This can represent a significant cost to the state.

Findings from a survey of states that provided written translations of statewide assessments in SY 2006-2007 indicate that a variety of content areas were offered in non-English languages. As Table 3 below shows, mathematics was the most commonly translated content area, with 11/12 states providing translations of a mathematics test. In addition, science (6/12) and social studies (5/12) were frequently translated. Surprisingly, a number of states also provided written assessments of reading and writing in non-English languages, although it is important to note that usually these were not direct translations but rather adaptations of the English test to the non-English language.

Table 3.
Content Areas for Which Written Translations Were Provided, By State (SY 2006-2007)

	Reading	Writing	Language Arts	Math	Science	Social Studies
DE				✓	✓	✓
KS				✓		
MA				✓		
MN				✓		
NE		✓				
NM[a]	✓	✓	✓	✓	✓	✓
NY[b]				✓	✓	✓
OH	✓			✓	✓	✓
OR				✓	✓	✓
RI				✓		
TX[c]	✓	✓		✓	✓	
WI				✓		
Total	**2**	**3**	**1**	**11**	**6**	**5**

[a] The New Mexico High School Competency Exam consisted of six subtests – reading, language arts, math, science, social studies, and writing – which were translated into Spanish.
[b] NY translated social studies and science for grades 5 and 8, in addition to US History & Government, Global History & Geography, Living Environment, Earth Science, and mathematics.
[c] In TX, the reading and math tests were translations, but the writing test was separately developed for Spanish.

Because social studies assessments tend to have US-centric content, they may be less amenable to translation and less useful than translated assessments of other content areas. Students educated in other countries frequently are unfamiliar with US history, civics, and government and may lack some of the cultural knowledge associated with particular events. On the other hand, this may not matter greatly in a standards-based assessment system. In such a system, agreement has been reached on content standards and these standards have been adopted by the state board of education or some other appropriate authority identified and codified in state law.

In addition, there may be issues with translating terminology that exists in English but not in the target language as fixed phrases. This problem can be ameliorated if the State Department of Education publishes English > target language glossaries for social studies terms used on assessments. These glossaries not only ensure that terminology is consistently translated across forms or years of an assessment but also can be used by curriculum writers and teachers in bilingual education programs. Both New York and Oregon (listed in Table 3 above) have published glossaries of content area terms in non-English languages.

17. What effect can assessment in the native language have on reliability, validity and score comparability?

Abedi (2002) has noted that when standardized tests are analyzed using ELL-only data, the tests often show lower degrees of reliability, which in turn makes them less valid, which in turn threatens score comparability. He attributes this to the role of English language proficiency in test performance and to other factors. Other factors include prior educational background, parent education and involvement in the school, the educational orientation of peers, self-confidence in relation to learning, access to highly trained teachers, etc. The factors other than English proficiency also apply to a standards-based achievement test in the student's native language. These factors combine to produce similar challenges for the reliability, validity, and score comparability of standardized tests in the student's native language.

This phenomenon is strongly related to two sampling concepts - restriction of range and restriction of variance. When a sample of students is more homogeneous than the total population of students, then the sample is biased. That is, it is not representative of the total population. One of the main ways that it is not representative is that the distribution of scores is more compressed (restricted) than the distribution of scores for the total population. Restriction of range means that the score distribution does not extend fully and normally across all possible scores. Restriction of variance means that the scores tend to be clustered together rather than spread out.

17.1 Effect on Reliability

This situation directly affects the statistical reliability of a test. If a test is designed to measure a broad range of achievement (knowledge, skills and abilities), and the sample that takes the test does not exhibit a normal distribution of scores (one that covers the full range of the test and one that approximates a bell-shaped curve), then the reliability of the test for the sample will be lower than the reliability of the test for the total population (the group that does not exhibit the restriction of range or variance).

Thus, it is important to recognize that although reliability is commonly stated to be a property of a test, in reality it is a property of the data produced by the sample of examinees who take the test. Thus, if the scores produced by the sample nicely reflect the spread of levels of knowledge, skills, and abilities tested by the assessment, then the fit between the test and the sample will be high. In this case, the reliability will be high. However, if the same test is given to a group that does not match the spread of levels of knowledge, skills and abilities measured by the test (for example, nearly all students do well or nearly all do poorly), then the reliability will be much lower.

Because translated assessments are normally administered to a subgroup of the total population of students, they normally show lower reliability than the English versions of the assessments. While this can be a source of concern, since we want psychometrically equitable assessments for ELLs, one must remember that for ELLs the translated version is likely to produce a more accurate score than the English version, so long as the student is literate in the language of the translation and is accustomed to reading content area material in the native language. So, while the original English version of the assessment may exhibit higher statistical reliability than the

translated version, it would be erroneous to claim that the English version is a more accurate measure for a non-English speaker or an ELL.

When judging the reliability of a translated test, it is important to consider the effect of restriction of range and variance on the reliability coefficient that is obtained from the data set. Standard corrections for this can be used to give a more realistic picture of the reliability of the translated version of the test when a normal distribution of test scores is obtained.

17.2 Effect on Validity

The reduced reliability that is often found in data produced by samples of examinees who take translated tests, can also affect the results of validity studies on the translated measures. That is, since reliable measurement is a prerequisite for correlation between two measures, a reduction in reliability will lower the correlation between the translated measure and any other measure or variable. Again, this is to be expected with a sample of examinees that exhibits restriction of range and variation in test scores. However, there are common statistical corrections for this, such as the correction for attenuation or unreliability in test scores. By applying these corrections, a more realistic appraisal of the evidence for the validity of the translated assessment can be achieved.

In judging the validity of a standards-based assessment, one must remember that the assessment is supposed to measure learning of the knowledge, skills, and abilities defined in the content standards. Thus, a standards-based assessment is essentially an achievement test, a test of achievement in learning the content delineated in the standards. Therefore, content validity is the essential starting point in considering the validity of a standards-based assessment. Content validity is routinely judged by evidence of alignment between test items and test standards. Alignment studies typically employ a framework for classifying the complexity of standards and items. Examples are Webb's Depth of Knowledge framework and Bloom's taxonomy of cognitive tasks. When an alignment study shows high alignment between the standards and the assessment, there is strong evidence for content validity.

But what happens when the assessment is translated to another language? Again, if the translation is accurately done, then each item tests the same content in both languages. Thus, the translated version retains the same content validity as the original English version. Whether or not the translation has been accurately done can be evaluated through a quality assurance procedure referred to as a translation verification study. In a translation verification study, the content standards, the original English version and the translated versions of the assessment are reviewed by bilingual subject matter experts. The reviewers are asked to determine whether each translated or adapted item is aligned to the same content standard, maintains the intended reading level of the original item, maintains the intended difficulty level of the original item by ensuring that the item was not simplified or clarified in the process of being translated, and maintains the essential meaning and style of the original item. (For more on translation verification, see the discussion of qualitative evidence below.)

When a translation verification study shows that all items have been properly translated, the translated test retains the same content validity as the original version.

17.3 Effect on Score Comparability

Score comparability is the degree to which the scores obtained on the original and translated versions of the test have the same meaning or demonstrate the same level of mastery of the content standards. In judging score comparability, it is common to compare the original and translated versions on the criteria of reliability and validity. However, as shown above, a test's reliability and validity are influenced by a restriction of range and variation in test scores which are to be expected. To evaluate score comparability, it is important to examine all evidence available, both quantitative and qualitative, and to interpret this evidence in a discerning way. In the case of quantitative evidence, that means considering the effects of restriction of range and variation in test scores on reliability and validity coefficients. Perhaps that one area of test analysis that is not affected by the sample that takes the test is content validity. Thus, one should look at evidence of content validity and the results of translation verification studies to determine whether content validity is equal for the original and translated assessment. If it is, then there is unbiased evidence for score comparability.

18. What effect does a decision to create a native language version have on the test development process?

Besides adding expense to the state's assessment budget, the decision to provide native language versions of assessments has other effects on the test development process, whether the assessment is provided in written or in audio format, and whether it is a translation or a separately developed instrument. The most significant effect on the test development process is that additional time is needed to prepare a native language version of an assessment above and beyond what is normally needed to develop and print the English language version. Even if the state chooses to translate or adapt an English language assessment (the fastest option), a minimum of several weeks to one month is needed to provide a high quality translation of the content and to ensure that the same constructs are being measured in both English and native language versions. If the state chooses to develop separate native language versions of assessments (parallel development), the test contractor will need to go through the same steps to create the non-English version as the English version, meaning that the two versions should be created in tandem if testing deadlines are to be met. This involves establishing another test development committee (or committees if more than one subject is involved), which means additional funding is required.

19. What are possible political issues associated with assessment in the native language?

Often the decision of whether to offer assessments in the native language is influenced and shaped by political factors in the state or district. Favorable laws toward the language and toward linguistic diversity typically support the provision of assessments in the native language,

whereas more restrictive laws that heavily favor the use of English typically serve to deter states and districts from offering assessments in non-English languages.

Two neighboring southwestern states, New Mexico and Arizona, provide examples of how political pressures can have a dramatic impact on language assessment policy. In New Mexico, native Spanish speakers comprise approximately 79% of the ELL population, making Spanish the largest minority language group in the state. Several laws have historically supported English-Spanish bilingual education in the state, beginning with the state's first constitution in 1911, which required that both English and Spanish be used in all government publications. Bilingualism has continued to be integral to the state's identity, and in 1973, the New Mexico Legislature passed the Bilingual Multicultural Education law, which is still in effect today to provide school funding for bilingual education in the state. New Mexico's education policies reflect the tradition of this legislation supporting public instruction in English and Spanish, making the provision of statewide assessments in Spanish mostly uncontroversial. In fact, the high school exit exam, the New Mexico High School Competency Exam (NMHSCE), has been available in both English and Spanish since the late 1980s, making New Mexico the first state to offer a statewide assessment in a non-English language.

In Arizona, New Mexico's neighbor to the west, slightly more than 15% of the K-12 student population in 2000-2001 was classified as ELL, and, as in New Mexico, a large percentage of those ELLs are native Spanish speakers. (According to NCELA data from SY 2000-2001, 86% of the ELLs in Arizona were Spanish speakers.) During the 2000-2001 school year, all three components of Arizona's Instrument to Measure the Standards (AIMS) test (reading, writing, and mathematics) were translated into Spanish and administered at the four tested grade levels (3, 5, 8, and 10). However, in November 2000, Arizona voters voted in favor of Proposition 203 (English Language Education for Children in Public Schools), adopting English as the official language of the state. The law stipulates that all public school education in the state be conducted in English. Students with limited English proficiency are to be placed in a structured English immersion program for not more than one year, at which point they are mainstreamed into English classrooms. Furthermore, the bill also mandates that all students take the statewide assessments in English, regardless of their proficiency. Since the law took effect in 2001, all statewide assessments have been offered exclusively in English.

As these examples demonstrate, the decision to offer assessments in the native language is not solely a question of numbers. Despite the fact that Arizona and New Mexico both have large Spanish-speaking ELL populations, New Mexico's laws support the provision of Spanish language instruction and assessment, whereas Arizona's newly-enacted Proposition 203 prohibits instruction or assessment in any language other than English, bringing it in line with other states, such as Virginia and North Carolina, where laws are in effect that directly or indirectly prohibit testing in other languages.

But laws, cost, numbers, and psychometric concerns are not the only factors that influence the decision to offer native language assessments. States such as California, Illinois, and Florida have large numbers of ELLs but do not offer assessments in the native language. The reasons may be complex and ultimately may relate to the attitudes of individuals who are in a position to

make a decision on this issue. Clearly, the basic attitudes, orientation, and beliefs of the state's leadership can also influence what happens in the state.

20. What quantitative evidence could a state provide to demonstrate that a translated assessment is comparable to the English version?

Once an assessment is rendered in a non-English language, the question of score comparability arises immediately. Translation/transadaptation raises concerns about the equivalence of the constructs assessed by the original and target language tests. These methods also raise concerns about the need for statistical adjustments to eliminate any change in the test score that may result from the translation or adaptation. There is considerable literature in the educational and psychological measurement fields addressing the concerns about how to assess construct equivalence and score comparability. This literature was the subject of a review by Sireci (1997). Sireci's review departs from proposed methods reported on by Hambleton.

Hambleton (1993, 1994) spearheaded an international effort to come up with guidelines for the translation and adaptation of tests. The members of the team who produced these guidelines were psychometricians with a strong orientation toward statistical analysis methodology. The guidelines they produced outlined research designs for testing the comparability of scores. The designs they proposed included 1) having bilinguals take both the original and transadapted versions, 2) having monolingual speakers of the source language take both the original and back-translated versions, and 3) having monolingual speakers of each language take the version of the test that corresponds to their language. In each design the two sets of scores would be compared and statistical adjustments would be made, based on the assumption that the adjustments correct error in measurement introduced by the use of different languages in the assessment.

However, each of these designs has significant flaws. The bilingual groups design creates a major hurdle, identifying individuals who are equally bilingual. Sociolinguists insist that such individuals are extremely rare, or even that it may be impossible to be equally bilingual in two languages. Wainer (1999:12) has noted that in test translation situations "we can never assume that anyone could compete equally on both forms." Sireci (1997) has indicated that another problem with this design is that the bilingual subjects are not representative of the monolingual cohort they are supposed to represent.

In the second design, which involves having monolingual speakers of the English language take the original and back-translated versions, no data involving the translated version is actually collected or analyzed; nor are the examinees from the language group that actually takes the translated version included in the design.

In the third design, samples from two different language groups take the tests. In this design, the monolingual examinees represent two different populations with different educational backgrounds. As a result, differences in item difficulty and in mean test scores may be due to differences in the two test populations in the mastery of the educational content and constructs assessed. Elder (1997) has noted that group differences may be explained in two ways: 1) There is a real difference in the ability being tested, or 2) There are confounding variables within the

test that systematically mask or distort the ability being tested. Similarly, Sireci stated that the major drawback of this design is "the inability to separate group proficiency differences from differences due to the tests themselves" (1997:14). He notes that the design could be improved by a matching procedure that would select subjects for the two groups on some related criteria, such as intelligence, socioeconomic status, etc. However, he notes that the results of studies on the effects of using related criteria as opposed to random selection are mixed. He also notes that such designs are usually impractical. That is, data on related criteria are not available or the researcher has a limited ability to sample from each group.

It is rarely possible to collect data that permits us to know whether mean differences are due to differences in language of the test or to differences in content mastery in the two examinee populations taking the test. Therefore, test program administrators cannot legitimately make statistical adjustments to compensate for these differences. As a result, almost by default, translated versions of state assessments are normally scored on the same scale as the English version. Still, in such situations it is possible to present some quantitative evidence that the test items function similarly. For example, one can correlate the difficulty values of test items in each language. If the correlation is high, then one can conclude that the items function similarly across language groups, even if they are not of equal difficulty across groups. One can also present qualitative evidence that the NLAs are measuring the same content.

21. What qualitative evidence could a state provide to demonstrate that a translated assessment is comparable to the English version?

Despite the fact that psychometric comparability is difficult to establish with assessments in the native language, qualitative measures can be used as evidence of comparability. Specifically, a procedure known as translation verification, first used to our knowledge in the early 1990s as part of the Third International Mathematics and Science Study (TIMSS), provides strong evidence of the comparability of original and translated assessments (Mullis, Kelly, & Haley, 1996). A translation verification serves as a quality control measure on the translation and consists of an independent, side-by-side, line-by-line comparison of the original and translated assessments. The translator reviews the overall layout of the two assessments, the translation of the student instructions, and the translation of each item. The translator compares each translated item with the English version and documents any adaptations in a translation verification report. The translator assigns two codes to each modification, one for the *type* of modification and one for the *severity* of the modification. This second code is intended to convey the extent to which the given modification is a threat to the validity of the item or the comparability of the English and translated items. Generally, those that are deemed necessary and do not affect the intent of the item are considered ‘minor modifications’ and those that affect the intent of the item (e.g., mistranslations) are considered ‘major modifications’.

In the translation verification report, the translator also provides an explanation of each modification and makes suggestions about how the modification should be addressed. It is important to note that in some cases, minor modifications are necessary to ensure that meaning is conveyed appropriately in the target language, and in those cases, the translator’s notes would indicate that the modification was minor and no action should be taken. In other cases, the

translator may find mistranslations or other modifications that affect the interpretation of the item. For instance, the translator may find that some text that appears in the English item has not been conveyed in the translated version. In cases such as those, the translator would recommend an alternative translation to correct the error.

In addition to the qualitative comments just described, a translation verification study also provides limited quantitative data. After reviewing all items on the test, the translator tallies the number of major and minor modifications and determines what percentage of the total number of items is affected. This overall percentage can be used as a yardstick by which to measure the quality of the translation, and the qualitative comments can be used to improve upon the existing translation or translations of subsequent assessments.

At the time of publication, three states -- Kansas, New Mexico, and Pennsylvania -- have conducted translation verification studies on their assessments in Spanish (Lopez & Stansfield, 2006). Other states are likely to follow their lead.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8(3),* 231-257.

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

Auchter, J. & Stansfield, C.W. (2001). A process for translating achievement tests. In C. Elder, A. Brown, et al. (Eds.), *Experimenting with uncertainty: Essays in honor of Alan Davies. Studies in language testing 11* (pp. 73-80). Cambridge, UK: Cambridge University Press.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*, 261-277.

Garcia, T., Parent, L.R., Chen, W.H., Ferrara, S., Johnson, E., Oppler, S. & Shiefh Y.Y. (2005). Study of a dual language test booklet in 8th grade mathematics. *Applied Measurement in Education, 18(2),* 129-161.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal for Psychological Assessment, 9(1),* 57-68.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal for Psychological Assessment, 10(3),* 229-244.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services: 2000-2001 summary report*. Washington, DC: The George Washington University National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs. Also available online at http://www.ncela.gwu.edu/policy/states/reports/seareports/0001/sea0001.pdf

Liu, K.K., Anderson, M.E., & Swierzbin, B. (1999). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 1*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Lopez, A. & Stansfield, C.W. (2006). *Translation verification study of the New Mexico mathematics standards-based assessments.* Rockville, MD: Second Language Testing, Inc.

Mullis, I.V.S., Kelly, D.L., & Haley, K. (1996). Translation verification procedures. In M.O. Martin & I.V.S. Mullis (Eds), *Third international mathematics and science study: Quality assurance in data collection.* Chestnut Hill, MA: Boston College. ERIC Document Reproduction Service, ED 406417.

Rivera, C., Collum, E., Shafer Willner, L., & Ku Sia, J. (2006). Study 1: An analysis of state assessment policies regarding the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S.G. (1997). Problems and issues in linking assessment across languages. *Educational Measurement: Issues and practice, 16(1),* 12-19.

Solano-Flores, G., Trumbull, E. & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-130.

Stansfield, C.W. (2003). Test translation and adaptation in public education in the USA. *Language Testing, 20*(2), 189-207.

Stansfield, C.W. & Bowles, M. (2006). Study 2: Test translation and state assessment policies for English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Stansfield, C.W. & Kahl, Stuart R. (1998). *Lessons learned from a tryout of Spanish and English versions of a state assessment.* Paper presented at a symposium on multilingual versions of tests at the annual meeting of the American Educational Research Association. San Diego, CA. ERIC Document Reproduction Service, ED 423 306.

Tanzer, N.K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. Hambleton, P. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235-263). Mahwah, NJ: Erlbaum.

Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and practice, 18(4),* 10-16.