Listening. Learning. Leading.®

Research Report
ETS RR–14-05

Considerations for Providing Test Translation Accommodations to English Language Learners on Common Core Standards-Based Assessments

Sultan Turkan

Maria Elena Oliveri

June 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Considerations for Providing Test Translation Accommodations to English Language Learners on Common Core Standards-Based Assessments

Sultan Turkan & Maria Elena Oliveri

Educational Testing Service, Princeton, NJ

In this article, we review translation, adaptation policies, and practices in providing test accommodation for English language learners (ELLs) in the United States. We collected documents and conducted interviews with officials in the 12 states that provide translation accommodations to ELLs on content assessments. We then summarized challenges to ensuring fair and valid accommodations to ELLs and provided recommendations to address the challenges involved in translating a content test while preserving its validity.

Keywords English language learners; translation and adaptation as test accommodations; Common Core assessments

doi:10.1002/ets2.12003

The English language learner (ELL) population makes up 11% of K–12 students in the public schooling context of the United States. Since the passage of the No Child Left Behind (NCLB) Act of 2001 legislation, schools and states have been held accountable for the academic achievement of ELLs. While the ELL population is growing in number, it is academically at risk, as evidenced by the results of both the National Assessment of Educational Progress (NAEP) and state-specific assessments in mathematics and English language arts. As a result, the achievement gap between ELLs and non-ELLs on standardized content assessments is highlighted. In the current policy context, it is critical to ensure equal opportunity for ELLs to demonstrate their content knowledge on high-stakes tests. Relevant and adequate accommodations,[1] by definition, should offer ELLs equitable access to test content and achieve fairness in the assessment (Duran, 2008) by providing opportunities for ELLs to demonstrate their content knowledge. However, the current accountability system has not been engineered in a way to consistently execute valid accommodations of ELLs on state content assessments.

Two understudied ELL accommodation approaches are translation and adaptation of a content test. A *translated test* is one in which only the language changes between the source English and translated target language versions of the test, while the content or targeted constructs stay the same (Bowles & Stansfield, 2008). An *adaptation* "involves substantial changes to the original English test material, 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, 2011, p. 403). *Transadaptation*, on the other hand, is the process in which relatively minor changes are made to both versions of the test. Transadaptation processes might involve replacing items that are unsuitable for translation or adaptation in the target language. This article reviews state policies and practices in relation to translating tests, as well as the much more involved process of adapting tests. In this review, we highlight the issue of comparability between the source and target language versions of the test by presenting the state practices used in constructing a translated version of a test dependently or independently from the source language. We also present challenges and recommend solutions when using translation and adaptation as a test accommodation.

This review is timely, given the current efforts to develop assessments based on the Common Core State Standards (CCSS) by the Race to the Top Assessment Consortia. In 2010, the Council of Chief State School Officers (CCSSO) and the National Governors Association Center for Best Practices (NGA Center) announced the release of the K–12 CCSS to guide teaching, learning, and assessment practices at the national level. Historically in the United States, standards for student learning and assessments have been developed and applied at the state level. The passing of NCLB was a turning point in the history of American public schooling whereby states had to abide by the federal guidelines (or risk consequences to federal funding). Following the NCLB, a new era has been ushered in by the adoption of

Corresponding author: S. Turkan, E-mail: sturkan@ets.org

the CCSS, as well as an effort to develop national assessments led by two consortia: Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (Smarter Balanced). As of the writing of this report, 45 of 50 states have adopted the CCSS and signed on to participate in using new national assessments.

The large-scale assessment programs that are in development under PARCC and Smarter Balanced consortia promise improvement to the current practices for including ELLs in CCSS-based tests (Young, Pitoniak, King, & Ayad, 2012). Once the new assessment programs are operational, all the states' current policies with regard to assessing ELLs' academic achievement and progress will change. The promise in this change is that all students should be held to the same high expectations outlined in the CCSS. To uphold the fairness of these higher expectations, two issues must be addressed: (a) providing ELLs with equal access to assessments in English and (b) identifying valid and relevant accommodations for groups of ELLs with varying proficiency levels and educational backgrounds. Particularly with regard to the translation accommodations, though while the states participating in PARCC and Smarter Balanced consortia may support the administration the nonparticipating states may not (Gallagher-Geurtsen, 2013).

As new assessments are built and policies are revisited, states that wish to allow translation accommodations would benefit from the answer to the following question: What is the best way for ELLs to be accommodated on the assessments? Given the dynamic educational policies of the United States and growing number of ELLs in the shifting demographic landscapes across the country, offering translation as a test accommodation might be more essential than ever before. For this reason, framework documents have been written to guide the design of certain accommodations. For instance, to guide the assessments to be developed by Smarter Balanced, Solano-Flores (2012) identified challenges in developing effective translation accommodations for culturally and linguistically diverse students and discussed the limitations and potentials of four translation accommodations in relation to fairness and validity considerations. This framework document emphasizes the urgent need to understand the challenges and promises in pursuing translation accommodations when compared with other accommodations, such as linguistic modification/simplification and the provision of bilingual dictionaries (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005). Doing so could offer techniques for improving the validity of test translation practices, as well as insights for future empirical evaluation of translation accommodations.

The Current Article

To understand the challenges and benefits of pursuing translation accommodations, we focus on content assessments that are administered under the requirements of Title I of NCLB.[2] We present challenges and recommend ways to administer valid translation accommodations after reviewing research and state practices on these accommodations. More specifically, we first present a review of relevant research on how translation accommodations are typically provided to ELLs on standardized tests in the United States. The focus is on research that identifies challenges with translation accommodations and that elucidates how to optimize test translation. Then, we present a review of state practices and policies in providing translation as a test accommodation across the United States. In the review of state practices in translation accommodations, we aim to provide an update since the latest review conducted by Rivera, Collum, Shafer Willner, and Sia (2006); Shafer Willner, Rivera, and Acosta (2008); Young and King (2008); and Stansfield (2011). As noted, the update is timely and needed, given the likelihood that states will be re-evaluating their policies in light of the forthcoming consortia national assessments, which will raise the stakes concerning availability of valid and reliable ways of accommodating ELLs on the assessments. This review also extends and elaborates on the Stansfield (2011) review, which only focused on states that provide oral translation as an accommodation to ELLs. Stansfield focused solely on the pros and cons of providing simultaneous recorded oral translation of the test content versus sight translation. The current article contributes to the ongoing discussion about translation accommodations by focusing on written translations and also provides an update on the status of the state practices.

Providing Translation as an Accommodation

The use of translated tests as an accommodation is legitimated under the view that the native language test provision is an affirmation of ELLs' rights to their own languages (Ruiz, 1988). If the intention is to mitigate the challenges posed by linguistic and cultural factors, providing translation as an accommodation might provide equitable access to tested content.

However, there is little empirical research on translation accommodations (Hofstetter, 2003; Kieffer, Lesaux, Rivera, & Francis, 2009; Rivera et al., 2006; Shafer Willner et al., 2008). In the few empirical studies there are, findings are mixed with regard to effectiveness of translation provided as an accommodation on U.S. content tests (Hofstetter, 2003; Kieffer et al., 2009; Robinson, 2010). For instance, Kieffer et al. (2009) reviewed two empirical studies that reported on the effects of Spanish versions of mathematics tests on ELLs' performance. The authors found that ELLs scored lower on the Spanish versions of the tests. However, the positive effect size for those ELLs who received instruction in Spanish was notable when compared with another sample of ELLs who received instruction in English. This particular finding implied that a translated test might be a more valid indicator of learning if it corresponded to the language of instruction. In a study of kindergarten-level learners, Robinson (2010) found that native Spanish-speaking children in a classroom where the language of instruction was Spanish performed better on mathematics when the test was administered in Spanish. These two studies point to the role that the language of instruction plays in performance on the translated test, with similar results evidenced in both lower and higher grades.

To better judge effects of translation accommodations, we also need to first understand the issues and complexities involved in translation practices that contribute to constructing and administering a valid translation of a test as an accommodation. One of the main issues in the administration of translation accommodations deals with whether ELLs have been exposed to the test content through instruction in their native language (Liu, Anderson, Swierzbin, & Thurlow, 1999). Abedi, Lord, and Hofstetter (1998) found that translating a mathematics test into another language did not help ELLs if they did not receive mathematics instruction in that same language. Simply put, they may not have mathematics-specific knowledge in Spanish (Butler & Stevens, 1997; cf. Kieffer et al., 2009; Solano-Flores, 2008). In fact, some ELLs might not even have literacy skills in their first languages (Quest, Liu, & Thurlow, 1997). Alternatively, newly arrived ELLs who have received content and literacy instruction in their native languages may still face challenges with test content in their native language because of the possible curricular differences in the way content was presented in their home countries. This issue supports the assertion that relevant accommodations should be assigned to ELL students according to their particular educational backgrounds, needs, and challenges in the native language and English language (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007). For instance Pennock-Roman and Rivera (2011) showed that plain English accommodations might be effective for ELLs with intermediate levels of English proficiency. Similarly, the authors suggest that ELLs with high native language proficiency but low English proficiency might benefit from the translated test version.

Another fundamental complexity regarding translation practice is that the translated and/or adapted content test may not always be measuring the same set of intended constructs across the two languages. This could be because of variability across groups in dialect variation or because cultural knowledge is introduced during the translation process (Solano-Flores, 2008). Research regarding what standards for dialect or form of a target language and cultural knowledge should be adopted in building a translated test is limited (Solano-Flores, 2008; Solano-Flores & Trumbull, 2008). Further van de Vijver and Poortinga (1997) noted that versions of the same test, separately translated into two different target languages, might still end up assessing the intended construct in two different ways. This reasoning is somewhat supported by the finding that different language versions of a test may not yield test scores that are comparable across languages (e.g., Gierl & Khaliq, 2001; Hambleton, 1994, 2001; Sireci, 1997; van de Vijver & Tanzer, 1998). Thus, the same meaning may not be given similarly across languages (Greenfield, 1997) or even within the same language, subverting the goal of using translation to reduce threats to test validity.

Even though it is widely held that statistical analyses should be conducted to detect error possibly arising from translation practices, conducting appropriate statistical tests can be costly, as it often requires specific expertise and large sample data sets. For instance, differential item functioning (DIF) analyses are used to flag items that function differently among groups of test takers with the same latent traits (e.g., Ercikan, 2002; Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). However, DIF analyses require appropriate sample sizes and statistical expertise that might not be available in many states. It could be even more costly to couple DIF with other methods, for example, techniques used to detect lack of equivalence between two language versions of the test. A less resource-demanding method commonly practiced at the state level is to rely on reviews of human judges. Because these reviews are cost-effective, they are often used as part of the standard practices to ensure quality in test translations. However, research suggests that relying solely on human judgment is not sufficient. Solano-Flores, Backhoff, and Contreras-Niño (2009) recommended approaches combining qualitative and quantitative methods. They further noted that all efforts toward ensuring test equivalence and high-quality test translation

should be supported by a theory of test translation error that guides the reasoning of professionals who perform or review test translations.

Solano-Flores et al. (2009) provided a theory of translation that can be used to explain why ELLs perform differently from their non-ELL counterparts on test items flagged with DIF, thus helping the test developers to identify the errors caused by translation. The theory posits 10 translation error dimensions (style, format, conventions, grammar and syntax, semantics, register, information, construct, curriculum, and origin) classified under three broad categories: (a) item design, (b) language, and (c) content (see Solano-Flores et al., 2009). These dimensions may be modified according to the specific needs of each test translation project. The premise embedded in the theory is that errors inherent to test translation practice cannot be avoided, but can be minimized (Solano-Flores et al., 2009). Accordingly, the best translation is one that minimizes inevitable translation errors. Also, authors conceptualize test translation errors as multidimensional and systematize the process of test translation quality reviews. Use of the approach they advocate calls attention to minimizing the translation error types, helping to enhance equivalence between two tests. Systematizing the process of test translation review for a given test is also expected to be beneficial in raising stakeholders' awareness of appropriate test translation practices.

In sum, from our review, we conclude that the effects of translation accommodations on ELLs' performance are closely related to the quality of test translation and the approaches used to minimize translation error and maximize equivalence. Our review also makes it clear that valid and proper translation practices are widely emphasized in scholarly circles and that researchers advocate approaches to avoid translation errors and maximize equivalence between different language versions of the test. However, more research is still needed to understand which groups of ELLs would be best served by translation accommodations on content tests, as there are mixed findings about their effects.

Review of State Practices

Method

We conducted a survey of the U.S. state test accommodation policies and practices for ELLs in the summer of 2012. We collected official documents published on state departments of education websites. These documents described states' policies for providing translation accommodations. In addition, we invited state officials to participate in phone interviews to clarify the information collected from the electronic documents or sources. We sought answers to the following questions:

1. Do ELLs participate in statewide assessments?
2. Which state assessments include accommodation?
3. How is the ELL group that receives accommodations defined?
4. Is the state administering tests in languages other than English?
5. Are the tests translated or adapted?
6. Are the translated or adapted tests part of an accountability system?
7. What is the criterion for identifying the ELLs who need to benefit from translation accommodations?
8. What specific practices do the states that provide translation accommodations follow?
9. Are the translated versions of the tests constructed independently?
10. How are cutoff scores established?

We coded Questions 1, 4, 5, and 6 as binary categories (i.e., yes/no). For Questions 7, 8, 9, and 10, we first coded whether statements in the state documents identified which ELLs could qualify for receiving the translation accommodation. We then applied a second code to any statements that described practices to ensure comparability between multiple versions of a test. The first code is important to establish the criteria used to identify ELLs with the specific need for a translation accommodation. The second code was used to determine which practices are followed to ensure equivalence between two language versions of the test.

While presenting the translation practices led by states, we first identify which states construct the translated version independently of the English version and elaborate on the state practices that ensure construct equivalence between the translated and English versions of the test. States have two considerations, depending on the way in which test translation is offered. One concerns the criterion used to determine the groups of ELLs that qualify for the translation accommodation. The second pertains to whether the states construct the non-English version of the test independently of the English

Table 1 Translation Practices Based on Independent/Dependent Translation of English Test Version

States	Translated/ adapted versions of state content test(s)	Independent construction of the translated test	Direct translation from the English version
CA	√	√	
CO	√	√	
DE	√		√
KS	√		√
MA	√		√
NM	√	√	
NE	√	√	
NY	√	√	
OR	√		√
PA	√		√
TX	√	√	
WI	√		√

version. This particular consideration required following up with state officials on the specific translation implementation procedures, because this information was not often publicly available. We break up particular translation practices that states lead under the following three sections: (a) criteria for identifying ELLs for the purposes of translation accommodations, (b) independent construction of the translated test, and (c) direct translation from the English versions of the tests.

Findings From the Review of State Practices of Test Translation

The review revealed that 12 of the 50 states provided translated or adapted versions of their state academic content tests in languages other than English. These states are as follows: California (CA), Colorado (CO), Delaware (DE), Kansas (KS), Massachusetts (MA), New Mexico (NM), Nebraska (NE), New York (NY), Oregon (OR), Pennsylvania (PA), Texas (TX), and Wisconsin (WI). Table 1 shows translation practices based on independent versus dependent translation of English test version.

Criteria for Identifying ELLs for the Purposes of Translation Accommodations

The state documents revealed that the criteria to identify qualified groups of ELLs for translation accommodations lack specificity. For example, most of the documents were found to have language derived from the Title I proposition instructing the states to consider the duration of time that ELLs have spent learning English while administering state content assessments, as in the following:

> [Local Educational Agencies] can, in agreement with the [State Education Agency], conduct these assessments on an individual basis in a language other than English for up to two additional years for students who have not yet reached a level of English proficiency sufficient to yield valid and reliable information on what these students know and can do on an assessment written in English. (U.S. Department of Education, 2003, p. 6)

When a criterion so susceptible to different interpretations is used, practices to identify and classify ELLs and assess proficiency levels are likely to vary across states.

Construction of an Independent Test in the Native Language

The review of the policy documents revealed that in five states (California, Colorado, New Mexico, New York, and Texas), non-English versions of the state content tests are constructed independently of the English version. We briefly describe approaches of each state.

California offers standards-based tests in Spanish to ELLs who have been in U.S. schools less than 12 months or who are receiving classroom instruction in Spanish. The tests include reading and language arts tests administered in Grades 2 – 11, mathematics tests administered in Grades 2 – 7, and end-of-course tests including algebra I and geometry. Students taking the standards-based tests in Spanish, however, are still required to take the English version of the tests as well, for which they are allowed to receive accommodations such as having the directions read to them in their primary language and access to translation glossaries. California constructs the non-English version of the test independently of the English version.

Colorado offers two tests (writing and mathematics) at the third- and fourth-grade levels, but state officials note that the two tests are not viewed as compatible with the English versions, as they are constructed independently. Apparently, the number of third- and fourth-grade students who take the writing and mathematics tests in Spanish is not large enough to allow state officials to examine the comparability of these two tests through statistical analyses. New Mexico also constructs the state reading test in Spanish, quite independently of the English version, while transadapting the mathematics (Grades 3 – 8 and 11) and science tests (Grades 3 – 8) from English. The New Mexico reading and writing tests in Spanish are scored according to a separate rubric, because officials are concerned that the independently constructed tests might include culturally specific content that invalidates a common rubric approach.

New York provides up to 29 alternative language versions of high school level state global history and geography, U.S. history and government, mathematics, and sciences tests. These tests are constructed independently, and students are not allowed to take more than one language version of a test. Because of their restricted status, all alternative language versions (except Spanish) must be hand-scored.

Texas has instituted a new Spanish version of the state achievement test for Grades 3 – 5. This test was developed independently of the English version. The documents that we reviewed state that, during test development, state officials follow processes to assure comparability in content, rigor, and achievement standards.

Direct Translation From the English Versions of the Tests

Each of the other seven states constructed their non-English versions of the tests through direct translation of the English versions and/or transadapting from English. We now elaborate on the policies of the following states: Delaware, Kansas, Massachusetts, Nebraska, Oregon, Pennsylvania, and Wisconsin.

In Delaware, although the Spanish versions of the mathematics and science tests are translated directly from the English version, small sample sizes have impeded empirical studies showing that these tests measure comparable constructs. However, the cut scores used are identical to the English version. Additionally, the rubrics used in the Spanish version are taken from the rubrics of the English version.

Kansas provides a Spanish version of the mathematics assessment. The state department of education submits translation verification to the federal state department of education, and a state contractor deals with the technical aspects of comparability. The cut scores for the translated Spanish version are the same as those for the English version.

In Massachusetts, the Spanish and English versions of the 10th-grade mathematics tests are presented side-by-side in the test booklet. It is assumed that these tests are measuring the same constructs. Cut scores set for the English version are used for scoring the results of the Spanish version.

Spanish-speaking students in Oregon have the option to take the state's mathematics test in English or in a dual-language form (both Spanish and English). Although the translated versions are not constructed independently of the English test, the Oregon state department of education evaluates the comparability of test scores, using a three-pronged approach. First, an evaluation of the translations is conducted against the following four dimensions: syntactical accuracy, cognitive complexity, cultural relevance, and back translation. Second, in addition to a review of translation accuracy by Oregon teachers, an independent reviewer is contracted to troubleshoot any translation problems that might potentially influence the meaning of the language used in each item. The third prong consists of periodically conducted statistical tests, including DIF analyses, to examine differences at the item level, and multigroup confirmatory factor analysis to evaluate whether construct invariance can be established between the English-only and dual-language (English – Spanish) versions. The cut scores set to determine students' success are the same across both versions of the state test.

In Pennsylvania, the Spanish-language versions of the mathematics, algebra, science, and biology tests are also a direct translation of the English-language version, with some adapted items in cases where direct translation would not have

been appropriate**.** A translation verification study is performed by a testing company under contract to the Pennsylvania department of education. In addition, the company commonly performs (a) analyses of the ordering of items on the two versions based on logits and *p* values; (b) comparisons of student ability estimates run with Spanish items anchored on English item calibrations versus estimates run with Spanish item calibrations allowed to float; and (c) other analyses, including displacement and uniform DIF analyses. All these post hoc analyses are evaluated to demonstrate acceptable performance of the Spanish-language versions of the state's mathematics test.

In Wisconsin, the mathematics, science, and social studies tests are offered in Spanish and Hmong side-by-side with the English version. To enhance the comparability of the translated versions, the state's department of education follows a two-stage procedure. In the first stage, Milwaukee and Madison public school districts are, respectively, asked to translate the test into Spanish and Hmong. The translators recruited in each school district are designated as content experts in relation to the targeted construct and the characteristics of the student population. After the translations are complete, the state gathers a review panel of three to four people who are native speakers of Spanish and Hmong. They review each translation and make any linguistic changes as needed.

Discussion

From the review of state policies and practices, one conclusion we can draw is that there is no system in place to standardize translation practices across states, although a number of states adhere to guidelines for applying the peer-review guidance process. Peer review aims

> (1) to inform States about what would be useful evidence to demonstrate that they have met NCLB standards and assessments requirements; and (2) to guide teams of peer reviewers who will examine the evidence submitted by States and advise the Department as to whether a State has met the requirements. (U.S. Department of Education, 2009, p. 1)

The specific information pertaining to accommodations required from the states is as follows:

> Appropriate accommodations must be available for … [Limited English Proficient] students. The state should be able to meaningfully combine scores based on accommodated administrations with scores based on standard administrations … [and] provide documentation that a) appropriate accommodations are available and that the accommodations are used in a manner consistent with instructional approaches for each student and b) that valid inferences can be drawn from accommodated scores. Evidence may include, but is not limited to, procedures for training and monitoring, and reports from studies on the effect of specific accommodations. (U.S. Department of Education, 2005, p. 12)

Officials at 10 different state departments of education confirmed that the federal government monitors their compliance with NCLB through the peer-review guidelines at the level where test translation policy is brought to practice.

While it may be common practice to follow the guidelines, it should be noted that they do not specify what procedures and evidence are required at the policy and practice level to ensure the quality and fairness of the translation accommodations. In fact, an official at one state department of education observed that the only requirement is to submit evidence for translation verification. When asked what procedure the state follows to submit evidence of translation verification, the official's response was that translation verifications are mostly carried out through back translations by content experts who are preferably bilinguals. Perhaps not surprisingly, we observed that just about every practitioner and every entity responsible for test translation we interviewed or interacted with had a unique view of what valid translation is, how it should be implemented, and how seriously it should be taken into account as a factor of broader test validity.

Despite variations in purposes and practices of test translation, we contend that state officials and practitioners should be informed both of mainstream practices to deal with test translation challenges and ways to account for test translation accommodations in test validity. In that spirit, we next elaborate on the challenges and propose solutions with the hope

Table 2 Types of Flaws or Incidents Threatening the Validity of the Translated Tests at Different Levels of Analysis

Levels of analysis	Sources of threat to validity	Recommended solutions
Possible contributors to test translation error at the policy level	Selection of translators	Form a multidisciplinary team that is composed of curriculum experts, teachers with experience teaching relevant grades and subjects, linguists, translator, test developer, and psychometrician
	Checklists of translation review principles without specific quality control indicators	Supplement or replace checklists with a systematic evaluation of state practices
Item-level translation practices	Back translation practices	Supplement or replace back translation with concurrent or simultaneous test construction
	Lack of equivalence of targeted constructs between two versions	Follow a combination of both qualitative and quantitative methods to ensure construct equivalence between two tests either before or after the administration of the two versions of the test
	Lack of insight into different types of possible translation error	Consider various translation error dimensions laid out in a theory of test translation error

to contribute to scholarly discussion on valid test translation practices and the expanded use of relevant solutions in the near future.

Challenges and Solutions to Enhance Validity of Multiple Language Versions of Tests

There are challenges to the process of test translation at different levels: the test translation policy level and item-level translation practice level. Table 2 summarizes types of issues or challenges for both that could ultimately threaten the validity of the translated test versions. To counteract these difficulties, we propose a set of recommendations for improvement of test translation practices.

One common challenge faced in test translations is that test translators are mostly selected from a pool of individuals who are both content experts and bilingual or have native-like command of English (Stansfield, 2003). The risk here is that bilingual or native-like test translators may not be experts in translation, even if they are experts on the content covered on the assessment. Another related point that applies to translators of tests used across state (or national) boundaries is that the translators may not be familiar with how a particular content area is covered in the curriculum or instructional practices across different states (nations). For instance, Hambleton, Yu, and Slater (1999) found that the constructs of mathematics achievement differed between Chinese and the U.S. eighth-grade mathematics tests because the curricula and instructional practices across the two countries were not similar enough to test students on a single test. When the (U.S.-based) NAEP items were adapted to Chinese, it was observed that the mathematical concept of *estimation* (calculation of the rounded estimation of scores) was not recognized by the Chinese students, as it was not incorporated into their curriculum. As a predictable consequence, the Chinese students did not perform as well as their U.S. counterparts on items associated with this concept.

To tackle this challenge and ensure the quality of translation processes, one must deal with the misconception that native or near-native speakers or bilinguals can all be effective translators (Stansfield, 2003). Test translation review processes should adopt the practice of having a multidisciplinary team composed of "curriculum experts, teachers who taught the corresponding grades and subjects, a linguist, an American Translators Association–certified translator, a test developer, and a psychometrician" (Solano-Flores et al., 2009, p. 83). It is expected that a multidisciplinary team with such different kinds of expertise would be able to identify and resolve multiple dimensions of test translation errors. For instance, to eliminate the risk of selecting ineffective translators, Trends in Mathematics and Science Study (TIMSS, 2003) policy states that there will be both language and subject matter specialists on test translation

committees. The recommendation, therefore, is to diversify the kinds of expertise represented on test translation review committees.

Another issue resides in the Title I peer-review guidance policy mandating the evidence for ensuring accuracy of translation practices in the United States. Translators are tasked to show compliance with a checklist of principles. Interestingly, the validity of the translated version of the test, no matter where in the world and for what purpose it is offered, tends to be ensured by compliance with checklists. Review checklists that a state adopts are often brief lists that provide proof that some action has been taken to, for example, (a) minimize cultural differences, (b) ensure that the essential meaning has not changed after translation, or (c) ensure that the words and phrases are equivalent. We recommend that short lists of fundamental review principles be expanded, supplemented, or replaced with more systematic evaluations of state practices. A systematic evaluation of state practices would not only standardize the test translation practices but also provide a venue for states to communicate and consult with one another about the resolutions of the particular challenges they have faced. We find the need for further attention and possibly standardization of the checklist approach, as currently applied within the context of U.S.-based test translation policies and practices.

Further, while most states in the United States might follow a uniform set of guidelines to check for translation quality, the practices associated with the guidelines currently vary significantly from state to state. In fact, owing to the diversity of ELLs within different states, best practices for some states may not necessarily relate to other states. In this sense, when accounting for translation quality in one state in relation to another, population composition — specific characteristics of ELLs in the particular state — needs to be considered. Lack of quality control in test translation implementation adds another layer of variation that limits the validity of translated tests. The survey of state test translation practices presented in this article revealed that the most common practice of quality control was to conduct a qualitative evaluation of the translated versions of the test vis-à-vis a peer-review guidance process mandated by the federal guidelines for test translations, an approach that itself can be highly variable.

Another challenge emerges from using back translation as a translation method. As defined by Brislin (1986), back translation first involves the process of translating the scale from the original or baseline test into the target language and then reversing this process with the help of two or more bilingual individuals. Unless additional quality-control checks are carried out by independent translators, back translation as the sole method of test translation is likely insufficient to address threats to test validity and score interpretation, owing to variance that might be introduced that adversely affects construct equivalence between the two tests. Another source of error in back translation can be that the source language of the test is driving the final evaluation of the test, rather than also evaluating errors most apparent in the target language (Rogers, Gierl, Tardif, Lin, & Rinaldi, 2003). In sum, errors made in back translation might constitute threats to the validity of the scores of the translated test.

Issues around back translation practices could be minimized by leading either concurrent or simultaneous adaptation approaches. Ercikan, Simon, and Oliveri (2012) recommended a simultaneous test development approach because it enables formulation and conceptualization of the underlying construct and its measurement to the target languages at early stages of test development. Moreover, by using a simultaneous test development approach, unidimensional views of test development could be reconstructed. These views need to be reconstructed because, when adapting tests across languages, there might be linguistic and cultural features that cannot be directly translated. Thus, modifications to the source language version might be more amenable to transadaptation.

Another translation challenge concerns the procedures used to ensure construct equivalence among multiple versions of a test. Under the NCLB mandates, test translation and adaptation is supposed to make the scores of translated tests more comparable to those of the original test. If construct equivalence cannot be established between two tests, as van de Vijver and Poortinga (1997) noted, two language versions of the same test might end up assessing different sets of skills and knowledge. To enhance construct equivalence between multiple versions of a test, one should first decide whether it is more valid to adapt, rather than translate or transadapt, the test into the native language (L1) of the test takers (Hambleton & Li, 2005). Usually, the common practice is to translate the test following the English version (Solano-Flores & Trumbull, 2008). Second, while the construction of the test in L1 is dependent on the original English version, the procedures in test construction should be equivalent as well. That is, while the assessments constructed in English are tested on students and the wording of items is carefully worked out, translated tests are not usually subjected to the same level of rigorous, standardized processes (Stansfield, 2003). Third, stakeholders should utilize a combination of qualitative and

quantitative methods either before the two versions of the test are operationally administered or after administration at periodic intervals. DIF analysis, for example, is an appropriate quantitative method to help identify inaccurate translation of terms across languages (e.g., Ercikan, 2002; Ercikan et al., 2004). As an example of mixed methods, international test programs such as the Program for International Student Assessment (PISA) and TIMSS periodically adopt review and quantitative procedures to ensure effective and quality test translation. Implementing such methods, of course, can be costly because they require expertise and resources. Consequently, states often rely on less costly methods, such as human judgment. However, despite costs, states could adopt some compromise quality controls that are less costly but increase quality. We recommend that a combination of qualitative and quantitative methods should be conducted, at least at periodic intervals.

One last challenge with multiple translated versions of a test concerns the errors embedded in the process of translating. The challenge here is to acknowledge that there are errors but that the errors could be minimized through systematizing the methods followed to avoid errors. Language is not a fixed category or aptitude, but a dynamic phenomenon (Solano-Flores & Trumbull, 2003). By default, two different languages operate on different lexical, syntactical, and discursive structures. This situation might result in a lack of equivalence across two language versions of the test items, often defined as *test translation error* (Solano-Flores et al., 2009). Test translation errors inevitably arise from the various dimensions such as style, format, conventions, grammar and syntax, semantics, register, information, construct, curriculum, and origin. To tackle the translation errors, a committee of test translators should be prepared to meet the challenges associated with errors inherent to the translation process. The larger point relevant for this article is that language-related measurement errors are worth consideration and examination in testing ELLs to minimize inevitable errors through following a systematic framework such as the theory of translation error by Solano-Flores et al. (2009).

Finally, one general recommendation we can offer is that states should consider two important criteria when offering translation accommodations to a large group of ELLs: (a) an accurate classification of ELLs' academic language proficiency in L1 and English (Solano-Flores & Trumbull, 2008) and (b) the language in which they received instruction. Accurate and valid methods for classifying ELLs according to their literacy and academic language proficiency in English and their native language should be adopted. To this end, appropriate language tests and classification procedures are demanded when deciding when and whether to provide language accommodations. Also, before the decision to offer translation accommodation, it is important to know whether ELLs have received instruction in their native language. Assessment in the native language would not necessarily be a better option if the language of instruction has been English.

Conclusion

ELLs are a heterogeneous group with varied ethnic backgrounds, first languages, socioeconomic statuses, quality of prior schooling, and levels of English language proficiency. In the wake of the common core standards-based tests, we drew attention to several issues in the provision of test translation as an accommodation on standardized state tests in the United States. First, our review of the research on test translation revealed that there is little empirical research on the benefits of test translation accommodations. We reemphasized the necessity of appropriate assignment of accommodations according to particular needs of ELLs.

Second, a review of state practices showed that test translation varies by purposes, policies, and practices across states. This review also revealed that few statistical measures are taken to ensure equivalence between the translated and English versions of the test. In addition, the prevalence of checklists was noted as a challenge to ensuring valid test translation practices, because they do not specify a consistent standard for human judges to apply uniformly. Because the scope of linguistic quality control might be limited, it is advisable to minimize cultural differences and attempt to evaluate whether essential meaning has not changed after translation. We recommended, as an alternative, that a combination of qualitative and quantitative methods should be employed to enhance construct equivalence. It is our hope that our review (even though it may not stand the test of time), discussion of challenges, and recommended solutions will contribute to the ongoing process of enacting policies and practices that enhance provisions for providing translation accommodations, especially in the current context of common core standards-based assessments.

In conclusion, the three recommendations we consider are most critical. One, test translation should be assigned to the relevant groups of ELLs with consideration of whether ELLs have received instruction in the language of the test.

Second, researchers and practitioners need more sophisticated test translation processes that help the field move beyond the limitations of relying solely on back translation. Finally, it is imperative that researchers, practitioners, and policy makers prioritize the channeling of resources to more quality control processes to ensure construct equivalence between the two tests.

Notes

1 Though the definition varies, an accommodation is a change made to an assessment without altering the underlying construct. According to Butler and Stevens (1997), accommodation refers to the "support provided to students for a given testing event either through modification of the test itself or through modification of the testing procedure to help students access the content in English and better demonstrate what they know" (p. 5).

2 Excluded from consideration are international assessments and NAEP.

References

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report No. 666). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/r666.pdf

Abedi, J., Lord, C., & Hofstetter, C. H. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/TECH478.pdf

Bowles, M., & Stansfield, C. W. (2008). *Standards-based assessment in the native language: A practical guide to the issues.* Retrieved from the Education Week website: http://www.edweek.org/media/maz-guide%20to%20native%20language%20assessment%20v15-blog.pdf

Brislin, R. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137–164). Newbury Park, CA: Sage.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Technical Report No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://cse.ucla.edu/products/reports/TECH448.pdf

Duran, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education*, *32*, 292–327.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing*, *2*, 199–215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*, 301–321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2012). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues and practice* (pp. 110–124). New York, NY: Routledge/Taylor & Francis.

Gallagher-Geurtsen, T. (2013). *Common Core test accommodations approved for language learners*. Retrieved from the Every Language Learner website: http://www.everylanguagelearner.com/blogs/news/tagged/test-accommodations

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164–187.

Greenfield, P. (1997). You can't take it with you. Why ability assessments don't cross cultures. *American Psychologist*, *52*(10), 1115–1124.

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

Hambleton, R. K. (2001). The generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*(3), 164–172.

Hambleton, R. K., & Li, S. (2005). Translation and adaptation issues and methods for educational and psychological tests. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 881–903). Hoboken, NJ: John Wiley & Sons.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC Guidelines for Adapting Psychological Tests. *European Journal of Psychological Assessment*, *15*, 270–276.

Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education*, *16*(2), 159 – 188.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168 – 1201.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice*, *26*(3), 11 – 20.

Liu, K., Anderson, M., Swierzbin, B., & Thurlow, M. (1999). *Bilingual accommodations for limited English proficient students on statewide reading tests* (NCEO State Assessment Series, Minnesota Report No. 20). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.* (2002).

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*, 10 – 28. doi:10.1111/j.1745-3992.2011.00207.x

Quest, C., Liu, K., & Thurlow, M. (1997). *Cambodian, Hmong, Lao, Spanish-speaking and Vietnamese parents and students speak out on Minnesota's Basic Standards Tests* (NCEO State Assessment Series, Minnesota Report No. 12). Minneapolis, MN: National Center on Educational Outcomes.

Rivera, C., Collum, E., Shafer Willner, L., & Sia, J. K., Jr. (2006). An analysis of state assessment policies addressing the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *A national review of state assessment policy and practice for English language learners* (pp. 1 – 173). Mahwah, NJ: Erlbaum.

Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher*, *39*(8), 582 – 590. doi:10.3102/0013189X10389811

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *Alberta Journal of Educational Research*, *49*(3), 290 – 304.

Ruiz, R. (1988). Orientations in language planning. In S. McKay & S. Wong (Eds.), *Language diversity: Problem or resource?* (pp. 3 – 25). Cambridge, MA: Newbury House.

Shafer Willner, L., Rivera, C., & Acosta, B. (2008). *Descriptive study of state assessment policies for accommodating English language learners*. Arlington, VA: The George Washington University Center for Equity and Excellence in Education. Retrieved from http://ceee.gwu.edu/AA/DescriptiveStudy.pdf

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

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37*(4), 189 – 199.

Solano-Flores, G. (2012). *Smarter Balanced Assessment Consortium: Translation accommodations framework for testing English language learners in mathematics.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Translation-Accommodations-Framework-for-Testing-ELL-Math.pdf

Solano-Flores, G., Backhoff, E., & Contreras-Nińo, L. A. (2009). Theory of test translation error. *International Journal of Testing*, *9*, 78 – 91.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3 – 13.

Solano-Flores, G., & Trumbull, E. (2008). In what language should English language learners be tested?. In R. J. Kopriva (Ed.), *Improving testing for English language learners: A comprehensive approach to designing, building, implementing and interpreting better academic assessments* (pp. 169 – 200). New York, NY: Routledge.

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

Stansfield, C. W. (2011). Oral translation as a test accommodation for ELLs. *Language Testing*, *28*(3), 401 – 416.

Trends in Mathematics and Science Study. (2003). *Translation and cultural adaptation of the TIMSS 2003 instruments.* Retrieved from http://timss.bc.edu/PDF/t03_download/T03_TR_Chap4.pdf

U.S. Department of Education. (2003). *Part II: Final non-regulatory guidance on the Title III state formula grant program — Standards, assessments and accountability.* Retrieved from http://www.ed.gov/programs/nfdp/NRG1.2.25.03.doc

U.S. Department of Education. (2005). *A user's guide to preparing submissions for the NCLB standards and assessments peer review.* Retrieved from http://www2.ed.gov/admins/lead/account/peerreview/usersguide.doc

U.S. Department of Education. (2009). *Standards and assessments peer review guidance*. Retrieved from http://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29 – 37.

van de Vijver, F. J. R., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263–279.

Young, J. W., & King, T. (2008). *Testing accommodations for English language learners: A review of state and district policies* (College Board Research Report No. 2008–6). New York, NY: The College Board.

Young, J. W., Pitoniak, M., King, T., & Ayad, E. (2012). *Smarter balanced assessment consortium: Guidelines for accessibility for English language learners.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/Guidelines/AccessibilityandAccommodations/GuidelinesforAccessibilityforELL.pdf

Action Editor: John Sabatini

Reviewers: Alexis Lopez and Danielle Guzman-Orth