CHAPTER NUMBER
USING TRANSLATION AS A TEST ACCOMMODATION WITH CULTURALLY AND LINGUISTICALLY DIVERSE LEARNERS

SULTAN TURKAN
MARIA ELENA OLIVERI
JULIO CABRERA

1. Introduction

In this chapter, we discuss three main issues associated with using language translation[1] as a testing accommodation in content assessments administered to diverse learners. By "diverse learners" we refer to: a) English language learners (ELLs) schooled in the United States (U.S.) and b) culturally and linguistically diverse students (CLDs) around the world. Although the two acronyms (i.e., ELLs and CLDs) might refer to diverse student populations with different characteristics, we use the broader term CLDs to refer to these two types of students. We use examples from the U.S. to illustrate particular points associated with translation as an accommodation. However, we conjecture that these issues apply to other countries and contexts with large populations of CLD test-takers.

We highlight examples from the U.S. because the heterogeneity of the CLDs taking standardized content assessments is remarkable, yet there are monolithic approaches to providing accommodations to these students. The "one size fits all" approaches may not be conducive to meeting the linguistic and cultural needs of CLDs. Heterogeneity is evidenced by the multitude of native languages, cultural backgrounds and diverse ways of schooling in the home countries of the students. During the course of their

[1] We use the term test translation to align this chapter with the rest of the book, wherein the term test translation is used. However, the term adaptation might be more suitable for the types of accommodations conducted at the state level, as they go beyond literal translation to include cultural and other adaptations.

U.S. schooling, CLDs display traits of heterogeneity in relation to their English language proficiency in four areas of language: listening, reading, speaking, and writing. There is also diversity in their English proficiency in dependent on context (e.g., at school, at home, with friends, with relatives, etc.; Solano-Flores 2008, 190). These factors lead to heterogeneity of CLDs and pose continuous challenges to test developers in relation to finding ways to accommodate diverse of CLDs in content assessments. Translation accommodation is promising for those CLDs who could benefit from demonstrating content knowledge in their native (or strongest) language.

Testing accommodations are changes made to an assessment without altering the underlying construct. Accommodations are intended to provide increased access to content by making the assessment tasks comprehensible to students with limited language proficiency. Factors precluding access to the assessed construct might be related to cultural and linguistic differences between the culture of the test takers and the culture of the test developers, or the mainstream culture of the students for whom the test is developed. Language can also introduce construct irrelevance or underrepresentation in assessments. Construct irrelevance happens when the test performance of an individual is influenced by factors other than the measured construct. Construct irrelevant variance might occur in translated assessments that were developed for one mainstream population and administered to another population. As language carries cultural beliefs, values, and practices, the group of linguistically and culturally diverse test takers may not have shared knowledge or experience with the mainstream culture, resulting in measurement issues that are not related to the test content. Moreover, when test items are translated across languages, sources of incomparability might arise when the standard dialect used in the test does not match the dialect used at home (Solano-Flores 2008). Construct underrepresentation occurs when relevant content that makes up the targeted construct is not included in an assessment (Messick 1994). Messick argues that the assessment should authentically represent the targeted construct. These two issues might constitute major threats to the validity of inferences made from the content assessment. The objective of translation accommodation is to reduce construct irrelevant variance that might arise from the linguistic and cultural difficulties CLDs face in demonstrating knowledge on standardized content assessments.

Difficulties with measurement comparability between multiple language versions of a test constitute another threat to test validity and fairness because this might introduce bias for or against certain groups of test takers. Translation committees need to provide evidence

demonstrating that the meaning intended for the targeted construct is equivalent between two language versions of the test and does not favor any group of test takers over another for whom the test structure is similar. An item or instrument that is biased will produce non-equivalent scores, and is a concern when tests are translated across languages (Oliveri and Ercikan 2011). This may take place because particular words or terms may be used in one language version as compared to another, changing the meaning and differential difficulty of an item. For example, in a Swedish-English comparison (Hambleton 1994, 235), English-speaking examinees were presented with the following item:

> Where is a bird with webbed feet most likely to live?
> a. in the mountains
> b. in the woods
> c. in the sea
> d. in the desert.

In the Swedish translation, "webbed feet" was translated to "swimming feet," providing a clue to the correct choice and giving an advantage to the Swedish examinees. This illustration exemplifies the type of issues one needs to consider when translating items because a difference in test performance may be incorrectly attributed to differences in ability across examinee groups when the difference is related to test adaptation.

The main benefits of using translation accommodations are that they provide CLDs with increased access to tested content and equitable opportunity to demonstrate their content knowledge in the language over which they have most command. These benefits are particularly important for CLDs in the U.S. due to their large representation in the student population. CLDs represent approximately 11 percent (above five million) of the total PK-12 U.S. public school population (NCES 2012), with some states having at least 15 percent of CLDs (i.e., Texas, Arizona, California, and Nevada; Batalova and McHugh 2010). These students often lag behind in many content areas compared to their English monolingual peers, as their English language literacy skills take longer to develop (Lesaux and Geva 2006). Further, under the U.S. No Child Left Behind (NCLB 2002) policy, schools and districts are accountable for the academic achievement of CLDs. After NCLB, the new era of American public schooling is being defined by Common Core State Standards (CCSS), an effort being led by the *National Governors Association and the Council of Chief State School Officers.*

Adopted by 45 states, CCSS provides national standards for developing national assessments. Similar to the assessments administered

under the NCLB era, the new common core standards-based assessments will hold all students to the same expectations as guided by the CCSS. To meet these expectations, states will need to adopt effective ways of providing CLDs with relevant testing accommodations, taking into account the effectiveness and relevance of different types of accommodations to the linguistic and cultural needs of the heterogeneous characteristics of CLDs. Unfortunately, there is little empirical research about the effectiveness of different types of accommodations on the performance of CLDs to guide the design of the new common core based tests. Despite the lack of empirical research, framework documents have been written to guide the design of certain accommodations. For instance, to guide the assessments to be developed by Smarter Balanced Assessment Consortium, Solano-Flores (2012) identifies challenges in developing effective translation accommodations for CLDs and discusses the limitations and potentials of four translation accommodations in relation to fairness and validity considerations. Still, there is no sufficient empirical research on the effectiveness of translation accommodation (Robinson 2010; Hoffstetter 2003). This could be due to the limited use of translation accommodations across the U.S. In fact, findings from our review of the state test translation practices (Turkan and Oliveri, forthcoming), revealed that 12 of the 50 states provided adapted versions of content tests in languages other than English. One striking finding was the large variability in the measures taken to ensure comparability between multiple language versions of the tests.

In this chapter, we approach the fundamental issues of comparability between multiple language versions of the tests by discussing test design, administration, and validity of inferences made from results of assessments translated into a language other than English (i.e., translation accommodation). First, we discuss design and administration issues and provide recommendations to minimize translation error. Also, to increase comparability of items and the test, we offer particular examples of practices which can enhance test adaptation efforts, such as identifying the right expertise to translate assessments and consider the dialect of the target language of translation. We also provide strategies to improve the test review process to minimize inevitable translation errors. We emphasize the principle of assigning translation accommodations to CLDs by utilizing a mix of qualitative and quantitative methodologies to increase measurement comparability and minimize threats to test validity in translated assessments. In terms of validity-related considerations, we describe post-test administration strategies to eliminate threats to validity of translated tests.

2. Test Translation Design and Administration Issues

In this section, we provide suggestions about which approaches and methods to use when designing fair and valid test translation accommodations. These are important as the process of test translation involves more than just rewriting items in another language (Solano-Flores and Gustafson 2013).

2.1 Finding the right expertise to do the translation job

The first step of translating a test into a targeted language involves determining who will be part of the translation committee. Test translators are selected from a pool of individuals who are content experts and bilinguals or have a native-like command of English (Stansfield 2003). The risk here is that bilingual or native-like test translators may not be good at translation, even if they may be experts on the content covered on the assessment. We argue that test translators with expertise in content matters and/or test development ought to do the test translation. Moreover, test translators, content specialists, and test developers ought to form the core of the test translation review committees. If the right expertise is not recruited to translate the test, principles of comparability and test validity might be compromised. We thus suggest dispelling the myth of inferring high quality translation capabilities of individuals with expertise in bilingual education and teaching English as a foreign language without additional training in translation (Solano-Flores 2008). Similarly, bilingual translators might not be qualified to translate tests in all four language skill areas (i.e., reading, speaking, listening, and writing).

Measures ought to be taken to ensure selecting the right expertise. There are exemplar cases reflecting selection of the right expertise. For instance, the Trend in International Mathematics and Science Study's (TIMMS) guidelines require that there will be language experts, subject matter specialists, and external highly qualified reviewers on test translation committees that oversee the translation of a test (Technical Standards for IEA 1999; Yu and Ebbs 2011). More specifically, TIMSS employs a group of qualified translators for an initial translation of the international English version of the TIMSS. Following this translation, a panel of subject experts reviews the translation, identifies any translation and adaptation issues, and highlights adaptation needs. After the field test, the translated assessment is reviewed and updated again by the same teams of translators and reviewers. Then, an external review is completed by the International Association for the Evaluation of Educational Achievement.

In assembling this team, the qualifications and background knowledge of the translators should be considered. As highlighted in a study conducted by Roth et al. (forthcoming), translators may notice different issues in translations depending on their educational background and work experiences. In their study, three experts were asked to think out loud while reviewing adaptations of a science test administered to English- and French-speaking students. The experts differed in (pedagogical, pedagogical and content, or linguistic) knowledge. Study findings revealed that experts with content and pedagogical knowledge identified the type of differences potentially impacting responses to items by the students (e.g., the interaction between the item, its content, and the translation issues), whereas the language expert solely identified linguistic differences across test language versions. A committee of test translators, content specialists, and test developers should thus conduct reviews to ensure the comparability in structural, semantic, cultural and mechanical use of language between test versions. Instead, we suggest considering the proficiency of bilingual translators in the four skill areas of language and including linguistically qualified experts with pedagogical knowledge and pedagogical and specific knowledge in the content area of translation. These expert reviews are essential to ensuring fair and valid test translation. Once the right expertise is properly selected, another key issue to keep in mind when translating tests is the minimization of translation error.

2.2 Minimizing the inevitable translation error

To provide evidence for comparability between multiple language versions of a content test, sources of test translation error should be investigated and minimized at the item-level. When translating tests, errors in conveying the targeted meaning are inevitable. The investigation of cultural and linguistic factors underlying test translation error is important because it threatens comparable and valid measurement of students' skills. Lack of equivalence across multiple language versions of the test items is defined as 'test translation error' (Solano-Flores, Backhoff, and Contreras-Nino, 2009). To minimize translation error, comprehensive item reviews should be conducted at the design and test piloting stages. To help with these reviews, Solano-Flores et al. (2009) discuss a theory of translation error and guide test translators to utilize the theory in their immediate context and for the particular purpose of test translation.

The theory of test translation error is based on the understanding that language is not a fixed category or aptitude, but a dynamic and

probabilistic phenomenon (Solano-Flores and Trumbull 2003). According to this view, language encapsulates cultural norms and cognitive schemas. This view of language acknowledges that language proficiency is not deterministic but is influenced by multiple factors including levels of native- and English-language proficiency. Within this theory, the process of test translation is systematized in relation to quality control (Solano-Flores et al. 2009). For effective systematization of test translation, it is beneficial to bring all stakeholders to a level of understanding of the process of test translation, following a standard theory for translating tests.

Test translation theory identifies two main types of translation error: internal and external. Solano-Flores et al. (2009) identify ten dimensions of translation error: style, format, conventions, grammar and syntax, semantics, register, information, construct, curriculum, and origin. These dimensions are classified into internal or external dimensions. Internal dimensions (style, semantics, register, conventions, grammar and syntax, information, construct) are concerned with the work of the translators while the external dimensions (format, origin, and curriculum) are outside the scope of translators' work. Translation error types are identified under the specific type of dimensions mentioned above. For instance, under the internal dimension of "conventions" there are error types such as: *grammatical inconsistency between stem and answer options in multiple-choice items*, among others. Within the external dimension, namely "format," the following error types are identified: c*hange of size, font*, among others (see Solano-Flores et al. 2009, for more discussion). The theory does not universalize the set of test translation error dimensions and the error types included within each dimension. Rather, theory informs that these dimensions be modified according to the specific needs of each test translation project.

This theory sets an example for content reviews that every translation committee ought to complete during test translation design and piloting stages. If possible errors originating from the factors related to item sampling and content, which are somewhat outside the test translators' control, are recognized, then the focus could be more readily given to the language-related sources of translation error. Otherwise, reviews could be haphazardly random and may not capture the core sources of errors residing from internal language related dimensions of test translation.

2.3 Addressing dialect variation

Dialect variation, a cultural and linguistic factor, is another potential source of threat to the validity and fairness of a translated test. There may

be inevitable differences in the meaning conveyed between the two language versions of the translated test using different dialects in the same language. Solano-Flores (2006) defines dialects as "a variation of a language that is characteristic of the users of that language." He argues that dialects *within* a language (e.g., Cantonese and Mandarin) might be more distant than those *across* languages (e.g., Danish and Norwegian).

A dialect does not merely encompass variation in the linguistic structure of a language but also in cultural expressions. Moreover, dialects spoken by the populations of CLDs might not always match the standard dialect used in test development; therefore, CLDs may be somewhat unfamiliar with words and expressions used in the test (Solano-Flores and Gustafson 2013). These types of dialect differences were identified in a study conducted by Oliveri and Ercikan (2011) wherein words that were more appropriate for a test developed for French speakers from France were used in a French version of a test used in Quebec, Canada. Moreover, an interaction among the dialect, items, and student were reported in a study conducted by Solano-Flores and Li (2008). Study findings imply that dialect of the translated test influences the way CLDs respond to the items which ultimately have an effect on their test performance. Failure to account for dialect variation in the target language of the test, might make the inclusion of CLDs in large-scale assessments invalid and unfair.

Consequently, Solano-Flores discusses that different dialects should be represented in test translation review committees. Dialect variation might constitute a source of measurement error in tests administered in the U.S. due to the diversity of Spanish speakers. Approximately 80 percent of ELLs are Spanish speakers (NCES 2004). They come from multiple countries including Cuba, Ecuador, Mexico, and Puerto Rico, all of whom speak different versions of Spanish. Dialect variation should thus be considered when translating tests for test takers from these various Spanish-speaking countries.

One of the ways to eliminate any threats to validity resulting from the dialect variation of the target language is to pilot test the translated items to a representative sample of the target student population and follow up with in-depth interviews. Doing so would introduce both quantitative and qualitative evidence for ensuring validity and fairness of the translated test. Moreover, ensuring that items do not contain words that are particularly associated with one dialect (e.g., "chaval" in Spain to indicate boy, since "chico" might be more readily understood by all or most Spanish speakers) should also be considered in pilot testing. Another way is to conduct extensive reviews. For instance, PISA translation guidelines recommend forming a committee of two independent translators and a third individual

who develop the test through a "double-translation and reconciliation" procedure. This procedure involves translating the source material into the target language and reconciling the source languages into one single national version.

Dialect variation in the target language remains a source of threat to the validity of the test scores. We have recommended that the translated version of the test be piloted with students representing diverse linguistic backgrounds to ensure that the dialect the test was constructed in applies to a majority of the target test takers. The process of piloting and revising the test translation calls for time and resources. If the essential time and resources are not invested, the design of translated tests might compromise from the value of piloting and revising the translated versions of the tests (Solano-Flores and Gustafson 2013), which could, in turn, threaten the validity of the translated tests.

2.4 Assigning translation accommodation to relevant groups of diverse learners

Test translation accommodation is important to be assigned to students with the relevant linguistic and cultural needs (Pennock-Roman and Rivera 2011). The first question in relation to assigning relevant test accommodation in the form of a translated version of the test is: What makes a student eligible to take the translated version? One of the main criteria for assigning translation accommodation is that CLDs have received instruction in their native language or the target language of the translated test. If not, written translated tests should not be administered to CLDs, but individual student cases could be evaluated to determine if oral test translation might be appropriate and relevant to administration (Stansfield 2011). In the U.S., oral translation is provided as an accommodation, usually by bilingual teachers, for those ELLs who cannot read the translated script of a test in their native language. Therefore, test translation accommodation might be irrelevant to those who do not have literacy skills.

Blanket approaches to assigning translation accommodation to CLDs should be avoided. In assigning testing accommodations, it is important to consider not only the educational backgrounds of CLDs, but also their linguistic and cultural needs in their native language as well as in English. In other words, the decisions made about whether to assign translation accommodation should be based on a range of variables such as a student's literacy levels in the native language and English, the predominant language of instruction, length and degree of formal

schooling in the native language and so on.

Unfortunately, the vast literature in misclassification/misassignment continues to suggest that identification of CLDs remains a challenge not only due to the linguistic and cultural differences of students but also to the arduous differentiation between learning disabilities and second language acquisition (Klingner, Artiles, and Barletta 2006). Cultural differences between school personnel and CLDs increase inappropriate CLD referrals to special/remedial classes (Brown 2004), and are further compounded by unconscious stereotypes (i.e., racial biases). These inappropriate referrals are further increased because of the similarities between the characteristics of second language acquisition and learning disabilities and difficulties (Lesaux 2006). Poor assignment practices (e.g., lack of proper teacher preparation to identify ELLs) may result in being more harmful than beneficial (Artiles et al. 2005; Brown 2004). For instance, Brown (2004) highlights the over representation of CLD students in special education compared to their representation in the general population (14.2% compared to 5%, respectively), and compared to their non-CLD counterparts (62.5% compared to 63.1%). Misdiagnosis and misclassification of CLDs may result in misassignment of translation accommodation because CLDs may not have the literacy skills and/or may not have received instruction in the language of the translated test.

One way to address the issue of misassignment is to identify CLDs' needs. Based on a meta-analysis on accommodation effects, Pennock-Roman and Rivera (2011) found that CLDs with high native language proficiency and low English language proficiency benefit from a translated test more than CLDs with low native language proficiency and intermediate levels of English proficiency, because they can process the items translated in their native language. Alternatively, newly arrived CLDs who have received content and literacy instruction in their native languages may still face challenges with test content in their native language because of possible curricular differences in the way content was presented in their home countries. In other words, the curricular differences, or methods of learning a particular subject (e.g., the use of calculators), may be different from the mainstream approach.

It is important to translate a test in a target language the students have been exposed to during instruction of the tested content (Liu et al. 1999) and to minimize misclassifications. One key difference between the U.S.-based tests and international tests is that students who take the international tests, like the ones listed above, receive instruction in the language in which the test is administered; however, in the U.S., the language in which CLDs receive instruction is predominantly English.

Hence, it is more important that those who assign translation accommodations to CLDs in the U.S. follow these criteria than their international counterparts.

3. Identification of Sources of Threats to Validity of Translated Tests

Due to the manifold sources of errors, there ought to be evidence indicating that a translated test as an accommodation is valid in terms of its comparability with the language of the original test. The evidence should demonstrate that the meaning intended, as part of the construct assessed, is equivalent between two or multiple versions of the test in any language. If the evidence is not collected, there may be threats to test validity.

Specifically, in review of the practices followed across the U.S., the issue of comparability was not taken seriously. Most states do not have procedures to systematically ensure comparability and measurement equivalence between multiple language versions of a test. Common practice is to check the translated version of the test against a few criteria: 1) minimize cultural differences, 2) confirm that the essential meaning has not changed after translation, and 3) confirm that the words and phrases are equivalent. To illustrate a more specific case, the Oregon State Department of Education takes three measures to ensure the comparability of test scores driven from the translated versions (see for a larger discussion, Turkan and Oliveri, forthcoming). First, they ensure the accuracy of the translation according to four dimensions: syntactical accuracy, cognitive complexity, cultural relevance, and back translation. In addition to the review of every item translated by Oregon teachers, an independent reviewer is contracted to troubleshoot any translation problems that potentially influence the meaning of the language used in each item. The third step is to periodically conduct statistical tests such as DIF and multi-group confirmatory factor analyses to evaluate whether construct invariance can be established between the English only and dual language (English-Spanish) versions. Examples as such indicate how different measures could be taken to conduct investigations of comparability during and after test administration.

In this section, we identify sources of threat to test validity and discuss three strategies that could be taken in the post-test administration to minimize or eliminate the threats (see Ercikan, Simon, and Oliveri 2012,

for a more comprehensive review). The strategies are: a) psychometric (e.g., DIF, classical test theory and dimensionality) analyses, b) expert review of test items, and c) conducting student cognitive processes using think aloud protocols.

Strategies for eliminating threats to test validity in the post test administration.

To detect potential differences in performance across examinees on an item, DIF analyses are conducted. It is recognized that DIF might help identify the sources of translation error across multiple language versions of the test (e.g., Ercikan et al. 2004). In DIF analyses, examinees of the same ability are matched; DIF is identified if matched examinees have differential probabilities of responding to a test item. There may be different ways and methods of detecting DIF, such as contingency tables (e.g., Mantel-Haenszel), regression equations (e.g., logistic regression), or unidimensional item response theory (IRT) (for a review of these and other methods, see Roussos and Stout 2004). Once DIF is detected, further analyses (e.g., linguistic reviews of items) are typically conducted by panels of experts to identify bias and sources of DIF (Oliveri and Ercikan 2011). These reviews often focus on examining comparability of a) test item content, b) cognitive complexity, c) cultural load and d) linguistic differences across test versions. These reviews are conducted to examine whether responses not only reflect ability on the measured construct but also signal construct irrelevant variance. Reviews are important because bias threatens score comparability and reduces the validity of inferences made based upon test scores.

To illustrate, previous studies conducted to investigate differences in adapted versions of tests reveal difficulties associated with this process and highlight difficulties encountered in cross-cultural studies involving translated tests. For example, a study conducted by Angoff and Cook (1988) using verbal and mathematical items from the Scholastic Aptitude Test (SAT) and the Prueba de Aptitud Academica (the Spanish language equivalent of the SAT) revealed low correlations for the difficulties in verbal items between the two language versions. Moreover, Solano-Flores, Contreras-Niño, and Backhoff-Escudero (2006), using data from the 1995 Spanish version of TIMSS administration in Mexico, also found high percentages of items with translation errors and a high correlation between the severity of translation errors and item difficulty for the Spanish-speaking students, suggesting the test items were more difficult for

Spanish-speaking students. Furthermore, results of a study conducted by Sandilands et al. (forthcoming), using data from the 2006 PIRLS administration to English and Spanish speakers from the U.S. and Colombia, revealed the following differences across the Spanish and English versions of tests: a) item instructions and phrasing of the questions, b) grammatical structure, c) sentence length, d) vocabulary complexity, and e) word usage in 8 out of 24 items.

One of the constraints described above in providing translation as an accommodation is related to limited financial resources and qualified translators. Conducting analyses of DIF (particularly with small sample sizes) might be costly and require psychometric expertise (Ercikan, Simon, and Oliveri 2012). The use of expert reviews on test translations, which could be cost effective, might be useful in such cases to minimize DIF.

Other psychometric analyses can also be conducted to investigate potential differences across different language versions of a test. For example, classical test theory based analyses of test data involve examining item discrimination indices such as point-biserial correlations, and internal consistency reliability indices (Bowles and Stansfield 2008). Item difficulty values (or *p*-values) and conditional *p*-values (Muñiz, Hambleton, and Xing 2001) might be conducted with relatively small sample sizes, particularly, for low to medium stakes tests. For example, item statistics (difficulty and discrimination indices) can be compared and correlations of these indices can be calculated to obtain evidence of the degree to which items are ordered similarly for the comparison groups. If differences occur for particular items, those items are flagged and can be further examined by panels of experts. Dimensionality analyses can also be conducted to investigate similarity of factor structure at the test-level (Ercikan and Koh 2005, 25) and can be conducted at exploratory (Oliveri and Ercikan 2011) or confirmatory levels. Using these analyses (as well as those at the item-level) to investigate DIF require large sample sizes, which may not be feasible in states with small numbers of ELL students.

A third approach is based on utilizing think aloud protocols to investigate student cognitive processes. These analyses are conducted to examine whether differences in particular item attributes (response format, item type) and item content lead to differences in item responding across examinee groups. A study conducted by Ercikan et al. (2010) suggests the usefulness of think aloud protocols to examine measurement comparability of test items. In the study, think aloud protocols were used to investigate students' thought processes as they responded to test items using two language versions of tests (English and French) and to confirm whether results from think aloud protocols confirmed differences identified by

expert reviews. Results indicated that the think aloud study confirmed differences identified by expert reviews for 10 out of 20 DIF items. Findings suggest item attributes identified in expert reviews may not be the actual sources of DIF and expert reviews should not be used in isolation to detect bias. Results from the Ercikan et al., (2010) study suggest combining what students verbalize, observations from the test administrator while students are problem solving before and after task completion, and students' (correct and incorrect) responses as evidence of students' thought processes. Thus, the use of these methods might serve to complement expert reviews and be viable even in cases of small sample sizes (Ercikan et al. 2012).

4. Conclusion

In this chapter, we have discussed test translation as an accommodation along with the design, administration, and validity issues in using translation as an accommodation with culturally and linguistically diverse learners. We first presented issues related to designing and administering translated tests, in which four issues were discussed. First, it is essential that test translation is conducted by individuals with the right expertise, including bilingual, content and pedagogical experts. Unless the right expertise is selected for test translation, multiple language versions of the test might render incomparable, which constitutes a threat to the validity of the test. Second, the design of the test translation is inevitably error prone but could be minimized through extensive reviews against a common set of error dimensions. Third, the target language of translation might not represent the dialect that students learn and speak at home, and unless the language of the test is familiar to the students, the results of the test and score interpretations might render invalid and unfair. Lastly, we discussed the importance of assigning the translation accommodation to the relevant groups of diverse learners. The testing accommodation would not be of any use if the students did not have the literacy skills of the targeted language or did not receive instruction in that language.

Test translation is a promising type of test accommodation allowing diverse learners to demonstrate content knowledge in their native language. However, CLDs are heterogeneous groups coming from varied ethnic backgrounds, first languages, socioeconomic statuses, quality of prior schooling, literacy skills, and levels of English language proficiency, which brings along complexities in designing, administering, and ensuring validity of test translation. In terms of design, the first and foremost measure to take is to form an expert group of translators with expertise

both in the language and content. To avoid any adverse consequences of not having the right expertise to the translation, test translations should be designed by 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, Backhoff, and Contreras-Nino 2009, 83). It is highly recommended that a multidisciplinary team of differing expertise be formed to identify and resolve multiple dimensions of test translation errors. As the inevitable test translation errors are minimized, results and score interpretations of the test translation accommodation become more valid and fair.

Another major consideration related to test administration is that translation accommodations should be assigned to those culturally and linguistically diverse test takers who have content knowledge but not the language proficiency to take the test in its original language version. However, it should be required that these diverse test takers receive instruction in the language of translation. Otherwise, administering the translation accommodation would not serve its intended purpose. A systematic classification program that identifies the linguistic, cultural, and educational backgrounds of CLDs would help to detect which groups of CLDs would be most eligible to qualify for test translation accommodation.

Moreover, in terms of ensuring test validity, it is absolutely essential that test translation accommodation not favor any group of test takers over others who fall under the same underlying trait structure and manifest similar probabilities of correctly responding to an item. Furthermore, multiple versions of the test must be comparable in terms of construct representation. To ensure construct equivalence, 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. For instance, it is recognized that DIF analyses might reveal the inaccurate translation of terms across languages (e.g., Ercikan et al. 2004; Ercikan 2002). To ensure comparability at the development and design stage, simultaneous test development approaches should be followed, as Ercikan, Simon, and Oliveri (2012) argue, as 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, decentralized views of test development can be resorted to. These views are important because when adapting tests across languages there might be linguistic and cultural features that cannot be directly

translated. This might lead to modifications of the test version in the source language to include references that are not centered within one particular (dominant) culture, but instead, are more amenable to transadaptation.

The discussion presented in this chapter is timely, especially when the U.S. faces the challenge of developing the next generation of standardized assessments that measure higher order cognitive skills of *all* students. With large percentages of ELLs in the U.S. and linguistically diverse learners across the world, addressing ways to enable learners to accurately express their content knowledge is imperative. Further, this discussion contributes to understanding the mechanisms of test translation as an accommodation, as the available literature mostly focuses on the effects of translation accommodations on ELL performance (Hofstetter 2003; Kieffer et al. 2009). Very few conceptual framework papers (Solano-Flores 2012) exist that discuss major design, administration, and validity issues in translation accommodations for both U.S.-based and international tests. Discussion, as such, would enhance understanding of translation as test accommodation, which could then lead to increased fairness and validity of inferences made based upon results of the next generation assessments administered to linguistically diverse learners. Also, the review presented in this paper about issues and opportunities in administering translation accommodation is intended to serve next generation student assessments that are administered to diverse groups of students in the U.S. and around the world.

To further our understanding of translation accommodation, future empirical research directions might include investigating which content tests are most conducive to providing this accommodation with minimal test translation error. Also, future investigations could be pursued to understand what types of CLDs with interrupted formal schooling would best benefit from translation accommodations. Some might migrate to a new country with a solid background in content knowledge while others might arrive with no native language literacy nor any content knowledge. This type of research could be replicated with different groups of CLDs representing various native languages.

Works Cited

Angoff, William H., and Linda Cook. 1988. *Equating the Scores of the College Board Prueba de Aptitud Academica and the College Board Scholastic Aptitude Test. (College Board Report No. 88-2).* New York: College Entrance Examination Board.

Artiles, Alfredo J., Robert Rueda, Jesus J. Salazar, and Ignacio Higareda. 2005. "Within-group Diversity in Minority Disproportionate Representation: English Language Learners in Urban School Districts." *Exceptional Children* 71 (3): 283-300.

Batalova, Jeanne, and Margie McHugh. 2010. *Number and Growth of Students in US Schools in Need of English Instruction.* Washington, DC: Migration Policy Institute.

Schools in Need of English Instruction. Washington, DC: Migration Policy Institute.

Bowles, Melissa, and Charles W. Stansfield. 2008. *A Practical Guide to Standards-based Assessment in the Native Language.* Illinois: NLA—LEP Partnership.

Brown, Clara L. 2004. "Reducing the Over-referral of Culturally and Linguistically Diverse Students (CLD) for Language Disabilities." *NABE Journal of Research and Practice* 2 (1): 225-243.

Ercikan, Kadriye. 2002. "Disentangling Sources of Differential Item Functioning in Multi-language Assessments." *International Journal of Testing* 2: 199–215.

Ercikan, Kadriye, Rubab G. Arim, Danielle M. Law, Jose. F. Domene, France Gagnon, and Serge Lacroix. 2010. "Application of Think-aloud Protocols in Examining Sources of Differential Item Functioning." *Educational Measurement: Issues and Practice* 29 (2): 24-35.

Ercikan, Kadriye, Mark J. Gierl, Tanya McCreith, Gautam Puhan, and Kim Koh. 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, Kadriye, and Kim Koh. 2005. "Examining the Construct Comparability of the English and French Versions of TIMSS." *International Journal of Testing* 5 (1): 23-35.

Ercikan, Kadriye, Marielle Simon, and Maria E. Oliveri. 2012. "Score Comparability of Multiple Language Versions of Assessments within Jurisdictions." In *Improving Large-scale Assessment in Education: Theory, Issues and Practice,* edited by Marielle Simon, Kadriye Ercikan, and Michael Rousseau, 110-124. New York: Routledge/Taylor and Francis.

Hambleton, Ronald K. 1994. "Guidelines for Adapting Educational and Psychological Tests: A Progress Report." *European Journal of Psychological Assessment* 10: 229-244.

Hofstetter, Carolyn H. 2003. "Contextual and Mathematics Accommodation Test Effects for English-Language Learners." *Applied Measurement in Education* 16 (2): 159–188.

Kieffer, Michael J., Nonie K. Lesaux, Mabel Rivera, and David J. Francis. 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.

Klingner, Janette K., Alfredo J. Artiles, and Laura Méndez Barletta. (2006). "English Language Learners who Struggle with Reading: Language Acquisition or Learning Disabilities?" *Journal of Learning Disabilities* 39: 108-128.

Lesaux, Nonie K. 2006. "Building a Consensus: Future Directions for Research on English Language Learners at Risk for Learning Difficulties." *Teachers College Record* 108: 2406-2438.

Lesaux, Nonie K., and Esther Geva. 2006. "Development of Literacy in Language-minority Students." In *Developing Literacy in Second-language Learners: Report of the National Literacy Panel on Language-Minority Children and Youth*, edited by Diane August and Timothy Shahan, 53-74. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Liu, Kristin K., Michael E. Anderson, Bonnie Swierzbin, and Martha L. Thurlow. 1999. *Bilingual Accommodations for Limited English Proficient Students on Statewide Reading Tests: Phase 1*. (*State Assessment Series, Minnesota Report 20.*) Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Messick, Samuel. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Research* 23 (2): 13-23.

Muniz, Jose, Ronald K. Hambleton, and Dehui Xing. 2001. "Small Sample Studies to Detect Flaws in Item Translations." *International Journal of Testing* 1 (2): 115-135.

No Child Left Behind (NCLB)Act of 2001. 2002. 20 U.S.C. 70 *§* 6301 *et seq*.

Oliveri, Maria. E., and Kadriye Ercikan. 2011. "Do Different Approaches to Examining Construct Comparability Lead to Similar Conclusions?" *Applied Measurement in Education* 24: 1-18.

Martin O Micheal, Rust Keith, Adams J Raymond. 1999. "Technical Standards for IEA Studies." International Association for the Evaluation of Educational Achievement.

National Center for Education Statistics (NCES). 2004. "English Language Learner Students in U.S. Public Schools: 1994 and 2000."

Issue Brief, December 12, 2013. http://nces.ed.gov/pubs2004/2004035.pdf

———. 2012. "English Language Learners in Public Schools. (Indicator 8-2012)" *The Condition of Education*, December 12, 2013. http://nces.ed.gov/programs/coe/indicator_ell.asp.

Pennock-Roman, Maria, and Charlene Rivera. 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.

Robinson, Joseph P. 2010. "The Effects of Test Translation on Young English Learners' Mathematics Performance." *Educational Researcher* 39 (8): 582-590.

Roth, Wolff-Michael, Maria E. Oliveri, Debra Sandilands, and Juliette Lyons-Thomas. Forthcoming. "Tracking Sources of DIF Using Expert Think-aloud Protocols." *International Journal of Science Education.*

Roussos, Louis A., and William Stout. 2004. "Differential Item Functioning Analysis." In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, edited by David W. Kaplan, 107-116. Thousand Oaks, CA: Sage.

Sandilands, Debra, Maria E. Oliveri, Bruno D. Zumbo, and Kadriye Ercikan. (Forthcoming). "Investigating Sources of Differential Item Functioning in International Large-scale Assessments using a Confirmatory Approach." *International Journal of Testing*.

Solano-Flores, Guillermo. 2006. "Language, Dialect, and Register: Sociolinguistics and the Estimation of Measurement Error in the Testing of English Language Learners." *Teachers College Record* 108 (11): 2354-2379.

———. 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, Guillermo, Eduardo Backhoff, Luis A. Contreras-Niño. 2009. "Theory of Test Translation Error." *International Journal of Testing* 9: 78-91.

Solano-Flores, Guillermo. Luis A. Contreras-Niño, and Eduardo Backhoff-Escudero. 2005. "The Mexican Translation of TIMSS-95: Test Translation Lessons from a Postmortem Study." Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada, April 12-14.

Solano-Flores, Guillermo, and Martha Gustafson. 2013. "Academic Assessment of English Language Learners: A Critical, Probabilistic,

Systemic View." In *Improving Large Scale Education Assessment in Education: Theory, Issues, and Practice,* edited by Marielle Simon, Kadriye Ercikan, and Michael Rousseau, 87-109. New York: Routledge/Taylor and Francis.

Solano-Flores, Guillermo. 2012. "Smarter Balanced Assessment Consortium: Translation Accommodations Framework for Testing English Language Learners in Mathematics." *Smarter Balanced Assessment Consortium (SBAC).* September 18, 2012.

Solano-Flores, Guillermo, and Min Li. 2008. "Examining the Dependability of Academic Achievement Measures for English-Language Learners." *Assessment for Effective Intervention* 33 (3): 135–144.

Solano-Flores, Guillermo, and Elise Trumbull. 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.

Stansfield, Charles W. 2003. "Test Translation and Adaptation in Public Education in the USA." *Language Testing* 20 (2): 189–207.

———. 2011. "Oral Translation as a Test Accommodation for ELLs." *Language Testing* 28 (3): 401-416.

Turkan, Sultan, and Maria E. Oliveri. Forthcoming. "Considerations for Providing Test Translation Accommodations on Common Core Standards-Based Assessments." *International Multilingual Research Journal*

Yu, Alana, and David Ebbs. 2011. "Translation and Translation Verification." In *Methods and Procedures,* edited by Michael. O. Martin and Ina V.S. Mullis, 1-13. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.