Understanding Language | Language, Literacy, and Learning in the Content Areas

UNDERSTANDING LANGUAGE

Stanford University

Graduate School of Education

Assessing Students in Their Home Language

Guillermo Solano-Flores

Kenji Hakuta

May 5, 2017

Version 1.1

Table of Contents

Acknowledgements

The authors wish to express their gratitude to several colleagues who contributed to this paper. The paper benefitted from the discussions, feedback, and general support provided by Magda Chia, Rachel Kachchaf, and consortium state members. Finally, the paper was completed through a contract with the Smarter Balanced Assessment Consortium.

Resources produced by Understanding Language at Stanford University are available electronically at http://ell.stanford.edu.

Recommended citation:

Solano-Flores, G., & Hakuta, K. (2017). *Assessing Students in Their Home Language*. Retrieved from Stanford University, Understanding Language website: https://stanford.box.com/s/uvwlgjbmeeuokts6c2wnibucms4up9c2

Abstract

Assessing students in their home language is intended to produce more valid measures of academic achievement for English learners (ELs; students who are developing English as a second language). Provisions in the Every Student Succeeds Act (ESSA) offer a new set of opportunities for these students to demonstrate their knowledge by allowing state assessment systems to test them in their home language in the content areas of mathematics and English Language Arts. While, in principle, this new legislation is a significant step towards ensuring that valid inferences can be made about students based on test scores, it is limited in its ability to address the complexity of language processes and linguistic groups, according to current knowledge from the language sciences. In addition, testing English Language Arts in the ELs' first language poses serious threats to validity and fairness. Successful implementation of ESSA provisions largely depends on the extent to which assessment practices effectively address the nature of language development, the linguistic demands of each disciplinary content area, and the individual schooling histories of ELs. Using the perspectives of bilingualism, psychometrics, and educational policy, this paper discusses the possibilities and limitations of home language-based assessment. It discusses the factors that are critical to properly operationalizing home language-based assessment, and the sets of realistic expectations that policy and decision makers should have concerning policy and practice as critical to fair and valid assessment of ELs. Whether testing ELs with translated instruments is appropriate depends on whether they have received instruction in their first language, the type of translation, the availability of qualified translators, and the content assessed.

Assessing Students in Their Home Language

Guillermo Solano-Flores & Kenji Hakuta

Stanford University

Introduction

Attention to English learners (ELs) has moved from periphery to center in our national education reform effort, as the numbers of these students have increased their prominence in recent years, and their pattern of distribution across states has become more complex. Roughly one in nine or ten students in America's public schools is classified as an EL and one in five students in public schools comes from a home in which languages other than English are spoken (Kindler, 2002; NCES, 2016a).

Educational attainment of ELs, as measured by academic assessments in English, shows large gaps compared to non-EL populations. For example, 2010-2011 National Assessment of Educational Progress (NAEP) reading scores show a 36-point gap and a 44-point gap respectively in NAEP reading scores for 4th grade and 8th grade students (NCES, 2016b). Subgroup scores for ELs reported by states as part of their Title I accountability on their academic assessments also show large gaps whose magnitude depends on who is included in the EL subgroup (Hopkins, Thompson, Linquanti, Hakuta, & August, 2013).

Second language development is a long-term process. It takes between four and seven years for ELs to attain English proficiency at a level that allows them to meet district criteria for reclassification (Hakuta, Butler, & Witt, 2001; Thompson, 2015). Yet the Every Student Succeeds Act (ESSA) imposes the requirement that ELs be included in state assessment systems (typically available only in English) after a period of one year of schooling in English—which makes it difficult to validly assess any second language learner in the second language. Recognizing this inconsistency, the law has contained the following language since the inception of the notion of standards, assessment, and accountability was introduced in the Improving America's Schools Act of 1994 (Public Law 103-382, 1994):

> "English Learners shall be assessed … in a valid and reliable manner and provided appropriate accommodations on assessments administered to such students under this paragraph, including, to the extent practicable, assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English language proficiency"

By acknowledging the need for "assessments in the language and form most likely to yield accurate data," the legislation clearly signals the need for assessments in the native language of the student whenever appropriate. Yet most of the attention to this provision of the law has focused on accommodation practices in the administration of the assessment in English, not on the complexities of assessing students in the native language.

This paper contributes to understanding the complexity of assessing academic achievement among EL students in their home language. It makes an attempt to show that what counts as "assessments in the language and form most likely to yield accurate data" is shaped by the ways in which ELs are defined, the ways in which the accuracy of data is examined, and the ways in which "assessing in the home language" is understood. The paper provides a set of considerations intended to inform policy makers, decision makers, and practitioners in their efforts to comply with legal mandates according to current knowledge in the field of EL assessment.

More specifically, the paper provides conceptual considerations and examines practice implications of five major changes in ESSA:

- States have considerable flexibility in the construction of their Title I accountability system, although they continue to be held accountable for the EL subgroup, consistent with the civil rights origins of Elementary and Secondary Education Act (ESEA).
- The Title I accountability system must include an indicator of progress toward English language proficiency (ELP) in addition to an indicator of academic content achievement, bringing the two major academic indicators for ELs – content and language – within Title I.
- Up to seven states can be approved for "Innovative Assessment Pilots," granting flexibility to experiment with non-traditional assessments for accountability that may be locally developed and administered.
- Some flexibility is granted in how newly arrived ELs are assessed and counted in the accountability system during their first two years, but the provisions do not include the possibility of assessment in the native language for English Language Arts. [1]
- States are required to report in their state plan the languages other than English that are present to a significant extent in their student populations for which academic assessments are "not available and are needed" and the ways in which states "shall make every effort to develop such assessments."

Ultimately, the paper is concerned with valid assessment for ELs. For the purposes of this discussion, validity is understood as the extent to which adequate generalizations can be made about the skills and knowledge of these students based on their performance on tests. Thus, major questions underlying the discussion are: Under which conditions does testing ELs in their home language allow making more valid interpretations of their test scores than testing them in English? Also, what are the optimal characteristics of tests in the home language?

Following this introduction, the paper provides a short history of ESEA legislation concerning EL assessment. A trend in legislation towards more flexibility can be identified in the ways in which states assess ELs and towards a wider variety of testing practices. However, the vagueness of this legislation may lead to interpretations based on simplistic assumptions about language and linguistic groups.

The third section discusses the challenges of EL assessment. Many of these challenges stem from the fact that ELs are difficult to define and, in many cases, they are not properly identified. These challenges also stem from the fact that tests have unique linguistic features which are often underestimated.

The fourth section examines population misspecification as a potential threat to validly assessing EL individuals. It also discusses how translation may alter the constructs test items are intended to measure.

[1] Title I provisions allow for recently arrived ELs to be assessed in their L1 in Reading/Language Arts for 3 years, with an additional 2 years possible, and to use those results for accountability purposes.

The challenges of preserving the integrity of constructs across languages vary depending on content area, as language encodes knowledge in a unique manner for each discipline. Moreover, how students should be tested in the home language may be difficult to determine when the content area assessed is English Language Arts.

The fifth section discusses multiple ways in which “assessing ELs in their first language” can be implemented through different test translation formats. In terms of validity and practicality, each format (e.g., partial or full-text translation in the textual modality or in the audio or visual modality) has a unique set of advantages and disadvantages. Recent innovations in information technology in computer-administered tests allow designing translation formats that adjust to the unique set of needs of each EL student and which more effectively support ELs in gaining access to the content of items without altering the constructs those tests are intended to measure.

The sixth section discusses how decisions and practice concerning the assessment of ELs in their home language need to be guided by probabilistic views—as opposed to deterministic views. Probabilistic views recognize that there is always some uncertainty about the knowledge that is possible to gather about the proficiency of ELs in their first language and in English, and some uncertainty about the fidelity with which these students can be tested in their first language. While this error may be in some cases a reflection of poor assessment practices, it is mostly a reflection of the challenges that result from the complexity of language and linguistic groups.

The last section provides a summary and a short set of conclusions about the level of commitment needed from test users and decision makers concerning proper implementation, and the kinds of expectations it is reasonable to hold about assessments in the home language if these assessments are to contribute to more valid, fair assessment for EL students.

ESEA Legislation: The Language of Policy and the Policy of Language

Inclusion of ELs in Assessment Programs

The issue of inclusion of ELs in state assessment systems has been part of the Elementary and Secondary Education Act (ESEA) since the 1994 reauthorization as the Improving America’s Schools Act (IASA), carrying with it the inclusion provisions mentioned in the introduction. While the notion of identification of a student as an EL was carried in the Definitions section of the law, instrumentation was left to the operationalization of identification procedures required through the Office of Civil Rights as part of its enforcement of Title VI of the Civil Rights Act (commonly known as “Lau Compliance” after the U.S. Supreme Court decision in 1974, Lau v. Nichols).

The 2001 reauthorization of ESEA as the No Child Left Behind Act (NCLB) created strict accountability rules of adequate yearly progress, with a requirement for progress of different subgroups of students, including ELs in schools and districts. NCLB raised the stakes through required sanctions for schools and districts that failed to meet proficiency targets, set at 100% of the students—which virtually guaranteed failure for almost all schools unless the law was re-authorized.

NCLB also set up requirements for English language proficiency assessments aligned to a set of state standards corresponding to its academic standards. This accountability system was set up as part of Title III and required districts to set and meet targets for growth and status on their state English language proficiency assessment. However, because it was not part of Title I, this system of assessment for English language proficiency was not subjected to the scrutiny of the Title I peer review process.

By 2008, when President Obama came into office, in light of strict targets for growth toward academic proficiency, many schools and districts were failing to meet adequate yearly progress for Title I, suffering the fate of schools and districts identified as “failing” and facing strict sanctions. With political pressure mounting for relief, the administration under the leadership of Secretary of Education Arnie Duncan granted flexibility waivers for states agreeing to certain conditions, some of which are relevant to this paper. One of these conditions was the requirement that states adopt academic standards that are college- and career-ready, which was widely read to mean the Common Core State Standards developed under the leadership of the Council of Chief State School Officers and the National Governors’ Association.

Another notable requirement in the waiver process was that an applicant state adopt new English language proficiency standards that correspond to the college- and career-ready academic standards of the state. Note that the term “correspond,” rather than the term “align” (although the law still uses “align”) was used to signal that the ELP standards are not equivalent to academic content, but rather should speak to the language demands necessary for students to participate in and meaningfully engage with rigorous content. Many scholars and educators, including those involved with the Understanding Language initiative, welcomed this change as a way of bringing academic content and language development into coordination, rather than as separate instructional charges. An important document embracing this change was published by the Council of Chief State School Officers (2012).

Common Core State Standards and Assessment Consortia

The Common Core State Standards, which were finally released in 2010, received strong tailwind from the Obama administration through the funding of two state consortia—the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (henceforth referred to as Smarter Balanced)—to develop assessment systems aligned to the standards. Federal funding for these consortia is generally what critics mean when they refer to the Common Core as an example of heavy-handed federalism, since the standards themselves are adopted at the state level and nothing in federal law or policy specifies the Common Core.

The waiver process became de facto federal policy as the assessment and accountability portions of NCLB went into dormancy, and the inability of Congress to come to agreement on most matters, including education, cemented the policy. This environment also created an appetite for reform in assessment as the one-size-fits-all view of assessment and accountability in the style of NCLB became unpopular. Examples of innovative ideas can be seen in the California CORE districts and in the 51st State initiative (see Darling-Hammond, Wilhoit, & Pittinger, 2014).

The reauthorization of ESEA as the Every Student Succeeds Act (ESSA) of 2015 caught many by surprise. It reflected the extent to which unhappiness with the NCLB limbo became a bi-partisan consensus. While its focus on civil rights categories was regarded as a positive aspect, its rigid accountability perspective was reviled, because of both its one-size-fits-all approach and the excessive federal regulation that gave states limited flexibility.

Owing to its rigid focus on assessment for strict annual accountability, NCLB may have provided a long-term favor to advocates of the movement to view assessment "of, for, and as learning" (Black & William, 1998; Wiggins, 1998), who once again are questioning the value of annual summative assessment, raising the value of a continuum of assessment approaches that examine learning, and putting this into the policy framework for continuous improvement (Darling-Hammond, Wilhoit, & Pittinger, 2014). This broader perspective on assessment is reflected, for example, in the chapter on assessment of the California Department of Education's English Language Arts / English Language Development Framework (California Department of Education, 2014), which pays considerable attention to formative assessment and its role as a resource for teachers to inform their teaching and to provide feedback to their students.

The Challenge of Validly Assessing EL Students

Defining "English Learner"

There is no common national definition or set of criteria for defining an EL student, and there is no consistency even within a single state (Linquanti & Cook, 2013), which compounds the difficulty of getting a precise handle on the magnitude of the score gap between ELs and their non-ELs counterparts. Due to the complexities of language and the process of language development, efforts to define ELs are, to some extent, limited in their ability to capture the characteristics of these students that are relevant to valid testing. These limitations may lead to misspecification of EL populations and to errors in the identification of ELs (Solano-Flores, 2009).

ESSA defines "English learner" as an individual who, among other things, uses a language other than English at home and has difficulties in speaking, reading, writing, or understanding the English language that may be sufficient to deny the individual the ability to meet challenging state academic standards. In essence, the definition is basically the same as that provided by NCLB, which refers to "the ability of the individual to meet the state's proficient level of achievement on state assessments."

The definition acknowledges that ELs may vary in their levels of proficiency across different language modalities (i.e., speaking, reading, writing, and listening). This acknowledgement is important because speaking and listening skills are developed through social interaction, whereas reading and writing skills are typically developed through experience in formal instructional contexts (see Grosjean, 2001). The definition also recognizes the functional aspect of language (Mackey, 1968), especially in the context of school and disciplinary knowledge (see Halliday, 1993; Schleppegrell, 2004) as critical for a student to benefit from instruction by providing academic standards as the context for determining English proficiency. Finally, the definition recognizes that language is context-bound. How proficient an individual is in a language depends on the context in which language is used (Fishman, 1965).

Two limitations of ESSA's definition of EL relevant to making assessment decisions need to be discussed. The first limitation concerns decisions made about individual EL students. ESSA's definition is silent about the fact that ELs are bilingual—they continue developing their first language while they develop English, their second language. Typically, the use of a language other than English at home is employed to operationalize the identification of EL students, not to examine proficiency in the first language. This omission leads to inaccurate and fragmented views of any given EL individual student's language skills and to erroneous decisions concerning the assessment of EL populations.

One possible consequence of these fragmented views of ELs is underestimating language development. Bilingual individuals develop a language system that integrates their first and second language in a way in which their language skills are distributed in these two languages (Oller, Pearson, & Cobo-Lewis, 2007; Valdés, 2015). A parallel monolingual perspective (which focuses only on English proficiency) and a holistic monolingual perspective (which pays attention to proficiency in their two languages) render different perceptions of ELs' skills (see Hopewell & Escamilla, 2013; Kenner & Kress, 2003).

Another consequence of these fragmented views of ELs is overestimating ELs' proficiency in the first language. In the absence of information on reading and writing skills in the first language, faulty decisions may be made concerning the language in which it is best to assess a given individual EL student. Guided by the best of the intentions, but without proper information, EL students might be tested in their first language under the wrong implicit assumption that their reading and writing skills in the first language are as good as their oral skills in the first language.

The second limitation of ESSA's definition is its limited sensitivity to the fact that EL populations are heterogeneous (although various portions of the law do address differentiation within the EL definition, such as recently-arrived ELs, ELs with disabilities, and long-term ELs). Due to differences in culture, migration, schooling histories, and first language, among other factors, EL students vary considerably as to their speaking, listening, reading, and writing skills in both English and their first language (Batalova, Fix, & Murray, 2005). This heterogeneity exists even among ELs who are users of the same language. Yet the definition leads to assuming the existence of few and clearly distinguishable categories of English proficiency and, ultimately, to making blanket assessment policies that may produce invalid interpretations of test scores for some EL students.

The development of standards of English proficiency for ELs and the development of assessments of English proficiency based on these standards such as WIDA and the more recent ELPA21 (English Language Proficiency Assessment for the 21st Century), are major milestones in a series of efforts towards more valid assessment for ELs. Unlike previous efforts, intended to identify ELs' levels of English proficiency, WIDA is consistent with a vision of language, which "has expanded to encompass both social contexts associated with language acquisition and academic contexts tied to schooling in general, and particularly to standards, curriculum and instruction" (Gottlieb, Cranley, & Cammilleri, 2007, p. RG-6) and, consequently, with a vision that developing English as a second language is not exactly the same as learning English as a foreign language, or learning English Language Arts. Furthermore, it is a remarkable effort to provide information on English proficiency that is sensitive to the linguistic demands of academic contexts in the different language modes (see Cook, Boals, & Lundberg, 2011). ELPA21 extended the notion of academic language by building its standards directly on the disciplinary demands of language based on the Common Core and Next Generation Science Standards (ELPA21, 2014).

In spite of their significance, these accomplishments will not lead to more valid assessment practices for ELs if the information on ELs' English proficiency provided by WIDA, ELPA21, or similar, future English proficiency assessments is not used appropriately. For example, even if different measures of oral, reading, and writing proficiency in English are reported, treatment of EL students based on broad categories of English proficiency may persist unless proper professional development opportunities are made available to educators. These professional development opportunities should enable school administrators and educators to properly interpret those measures and make informed classification and assessment decisions for ELs.

Language as a Critical Component of Tests

In general, assessment instruments are administered under the assumption that the examinees have a minimum level of proficiency in the language of testing (AERA, APA, NCME, 2014). In addition to this assumption, the linguistic features of tests contribute to the complexity of validly assessing ELs. These linguistic features make tests different from other forms of text. For example, many test items contain complex problems stated in a relatively small number of words or sentences. Also, many items use certain grammatical forms, such as ellipsis (as in, Which of the following...), with higher frequency than other forms of text. Likewise, features such as an incomplete sentence followed by three or four complements preceded by capital letters (as is the case of the stem and its options in a multiple-choice item) are practically unique to the context of tests. Due to these linguistic features, tests pose students with a unique set of reading demands for both ELs and native English users.

Because of the dependence of tests on language as the medium through which they are administered, it is not a surprise that refining their linguistic features constitutes a great deal of the process of test development. A great deal of the process of developing tests consists of ensuring that their items' linguistic features (e.g., word frequency, sentence length, sentence structure, subordinate clauses, and presence of nominal phrases, among many others) do not affect comprehension adversely (Abedi, 2006). If tests are carefully developed, they should undergo a series of review iterations in which the wording of the items is refined based on comments from reviewers and the interpretations of responses from pilot students with whom these items are tried out.

This tremendous sensitivity of tests to language is even a more serious issue in the testing of ELs. For these students, the demands are more challenging not only because they are developing reading skills in a second language along with skills related to other domains but also because many items that include contextual information intended to make problems meaningful (for example, fictitious characters needing to compare their ideas, resolve a dilemma, or make a decision) use cultural referents and scenarios implicitly assumed to be familiar to all students but which may be specific to certain segments of the society.

While the potential lack of familiarity with these scenarios is a recurring concern in the literature on the testing of linguistic and cultural minority students, most of the information reported in the literature is anecdotal and based on specific items or small samples of items. Available empirical evidence indicates that students make sense of items based on their personal, everyday life experiences (Solano-Flores & Li, 2009a). Therefore, it is not unreasonable to argue that tests may potentially privilege white, middle-class, and suburban students over many EL students if their items tend to reflect middle-class and suburban contexts.

Language Proficiency and Fair Assessment

Decisions concerning who should and who should not be tested in their first language are to be made by educators. These decisions are likely to be erroneous in the absence of proper training, technical support, and appropriate information on EL students' school history. Unfortunately, because the definition of EL is silent about proficiency in the first language, schools are not obligated to obtain trustworthy information on the proficiency of their EL students in their first language.

Only 15% of EL students are immigrants (Zong & Batalova, 2015) and only about 30 percent of U.S. American schools provide bilingual education programs (American Federation of Teachers, 2002). Those programs may vary considerably in goals, ranging from transitional to dual immersion programs (Rennie, 1993). These pieces of information provide a gross indicator that the majority of ELs have not received instruction in their first language at any point in their lives. Thus, assuming that all of them have learned reading and writing skills in their first language and learned disciplinary content in their first language may do more harm than good.

A major challenge in assessing ELs in their home language is ensuring that test translations go through a thorough process of review and that sufficient time is allocated to make sure that this process is completed appropriately. Typically, the timelines for test translation are tight (say, one or two weeks), and leave little room for careful review (see Kopriva, 2008). The implicit assumption underlying this practice appears to be that, if the quality of the test in its original language version is good, those properties will be transferred to the target language as long as highly qualified translators are hired to do the job. Yet available empirical evidence indicates that the complexity of test translation has been underestimated. When a test is translated, the linguistic integrity of the items in the source language is not necessarily preserved in the target language (Turkan & Oliveri, 2014). Most importantly, there is evidence that translation can alter the constructs assessed by tests, to the extent that the skills and knowledge assessed may end up not being the same in the original version and in its translation (see Hambleton, 2005; Sireci & Faulkner-Bond, 2015). For example, the meaning and the grammatical complexity of the items may be increased or decreased and the words or terms used in the translation may be of lower or higher frequency among the population of users in the target language than among the population of users in the source language (Solano-Flores, Backhoff, & Contreras-Niño, 2009).

Altogether, these and many other factors contribute to making some items more difficult and some other items easier in the translation than in the original version of the test (see Arffman, 2013). Because of these complexities, high quality test translation cannot be achieved by simply allocating a few weeks for translators to produce the version of a test in a new language.

The current version of the Standards of Educational and Psychological Testing (AERA, APA, NCME, 2014) recognizes that, when tests are translated or adapted to be administered to students from populations for which the instruments were not originally created, the new versions need to be tried out with samples of students of the target populations. While the tone of the document does not emphasize sufficiently that test developers should be required to always take these actions, it is clear that impeccable translation does not suffice as validity evidence when ELs are tested in their first language.

Issues and Challenges in Home Language-Based Assessment

Heterogeneity of EL Populations

An important challenge for valid home language-based assessment is the linguistic heterogeneity of EL populations. This diversity concerns not only the number of languages spoken in the U.S. and the disproportionate numbers of their users, with nearly 80% of users of Spanish and dozens of languages spoken by small percentages of the total population of ELs (Kindler, 2002). It also concerns linguistic variation within broad linguistic groups due to dialect.

Dialects are varieties of the same language. These varieties are the result of differences in the frequency of oral and written modality features such as vocabulary, pronunciation, intonation, forms of speech, idiomatic expressions, and spelling, among many others (see Coulmas, 2013; Edwards, 2009; Phillips, 2006). Typical examples of dialects are the varieties of English used by the British and American press and broadcast television and the multiple varieties of Spanish used in the U.S.

Dialects are a reflection of social and demographic differences. Thus, dialect differences can exist between users of the same language but from different geographical areas, ages, genders, or socio-economic status. Because dialects are associated with their users, different dialects have different levels of social acceptability and prestige. A common misconception about dialects is that prestigious dialects are more complex whereas dialects with low prestige are “incorrect.” A large body of evidence shows that, regardless of social status, dialects are organized systems of conventions (see Wolfram, Adger, & Christian, 1999).

Although, in principle, dialects are typically regarded as mutually intelligible varieties of a language, it is probably more accurate to say that dialects are *mostly* mutually intelligible varieties of language. Therefore, the ability of individuals to deal with dialect differences in printed text may be shaped by experience. The performance of students in the low grades may be particularly sensitive to dialect differences, as the following mathematics word problem illustrates:

> *At what speed is an 18-wheeler moving if it takes 4.5 hours to move from Point A to Point B and these points are 300 km apart?*

In the U.S., older students may be more likely than younger students to have been exposed to the term, *18-wheeler* instead of *truck*—more frequently used by young students. Or, older students are more likely than younger students to have developed skills that allow them to infer the meaning of *18-wheeler* based on the context of the sentence, or to simply realize that knowing the meaning of the term is not critical to solving the problem successfully.

There is evidence that children who are native English users and have high levels of use of their non-standard dialects have more difficulty understanding words in the Standard English dialect than children who have low levels of use of their non-standard dialects (Gross, Chen, MacDonald, Kaplan, Brown, & Seidenberg, 2014). In the context of EL testing, there is evidence that dialect variation may affect students' abilities to properly interpret test items, regardless of whether they are assessed in English or in their native language (Solano-Flores & Li, 2006, 2009b). The lesson from this research is that, even if EL students are tested in their home language, variation due to dialect is a considerable source of error. This issue may be especially important in the testing of younger ELs. As a consequence, the use of a standard version of a language in a translated test is not a guarantee of accessibility to the content of items, especially at lower grades.

Different Linguistic Demands in Different Content Areas

As mentioned above, testing student populations in different languages poses a problem of equivalence—the extent to which there is certainty that two language versions of the same test measure the same constructs (Gierl, 2000; Hambleton, 1989). Equivalence can be examined through methods from item response theory—a psychometric theory of scaling—and, more specifically, the analysis of differential item functioning (DIF). An item is regarded to be differentially functioning across student populations (e.g., those tested in the original version of a test and those tested in the translation of that test) if samples of students of the two populations with similar performance on the overall test do not perform similarly on that item (Hambleton, 2005). Differential item functioning analysis may reveal that an item is biased either in favor of the focal group—in this case the sample of students who take the translated test—or in favor of the reference group—in this case the sample of students who take the test in the original language version (Sireci & Allalouf, 2003).

In an ideal world, all items to be included in a test should be tried out with samples of students from the focal and reference populations and scrutinized for differential functioning. Also in an ideal world, items detected as biased should be eliminated or revised and tried out again. In practice, however, given to budget limitations and tight timelines, it may be unrealistic to expect that test developers will perform all these analyses. Hence the need for improved and more rigorous test translation and test translation review procedures.

While translation quality plays a critical role in ensuring equivalence across languages, entire equivalence may be difficult to prove or reach because languages encode experience and meaning in different ways. Many examples can be given. One is the systems used to denote different kinship relations (e.g., see Kemp & Regier, 2012). Certain languages have specific terms to distinguish the order of birth of siblings or to distinguish relatives from one's father's side from relatives from one's mother's side. Also, languages have different grammatical ways of expressing time (e.g., perfective, imperfective, progressive, etc.). For example, a tense may involve a case in which something could have happened but did not happen or tenses may distinguish actions that occurred in the past progressively, without a clear ending, from actions that occurred in the past and which had a clear ending (e.g., Bybee & Dahl, 1989).

All these differences show that languages pose different set of resources for their users to communicate. When text is translated from one language into another, the translator has a twofold task. One is to preserve the intended meaning across languages; the other is to express that meaning in ways that are consistent with the rules and characteristics of the target language. In spite of translators' best efforts to preserve it, meaning may be altered, at least slightly. For example, while *brother-in-law* in English may mean either the husband of one's sister or the brother of one's wife, other languages have terms to make that distinction. In a translation of text written in one of those languages into English, that precision would be lost or the translator would need to make certain adjustments (i.e., adding some text) to preserve the original meaning. In contrast, in a translation from English into one of those languages, it would be impossible to resolve the ambiguity and the translator might not be able to determine the case of *brother-in-law* to which the original version in English refers.

In typical translation projects that do not involve tests (e.g., in literary translation), ambiguity due to these differences between languages is often mitigated by the content of the text, which may provide sufficient contextual information for the reader or listener to infer meaning expressed ambiguously. In test translation, that possibility is limited because test items are independent of each other and contain few (and usually short) sentences. With his colleagues, one of the co-authors of this paper (Solano-Flores, Backhoff, & Contreras-Niño, 2009b) has advanced the idea that error in test translation is inevitable (although it can be minimized) because the tension between preserving meaning across languages and doing it in ways that are consistent with the rules of the target language cannot be resolved entirely—even when translators do an impeccable job. Examples of translation error include an increase or a decrease in the complexity of the wording of an item or its reading difficulty, a slight modification of the meaning conveyed by the item, or a variation in the number of times that a critical term appears in the text of an item. Due to this impossibility, the fallibility of translations should be acknowledged as part of the efforts to validly test linguistically diverse populations.

The problem of equivalence poses different sets of challenges in different content areas because each discipline has a unique set of conventions for encoding knowledge through language (Lemke, 1998). In mathematics, the view of this discipline as a “universal language” may have originated from the high level of abstraction with which ideas can be synthesized and represented through a system of symbols. But a high level of abstraction does not mean that ideas cannot be represented in multiple ways or that those ideas have evolved in the absence of a culture or a community of users that shape usage (see Lemke, 2003).

The view of mathematics as a “universal language” may mislead into erroneously thinking that language issues do not pose problems for translation in mathematics assessment. To use a well-known example, different cultures use different conventions to separate thousands and decimals to represent numbers. A number that in the U.S. would be represented as *23,712.34* is represented in other cultures as *23.712,34* and in some other cultures as *23 712.34*.

These differences in conventions for representing information may pose different sets of challenges in different translation contexts. In the context of international test comparisons, the translations produced are adapted in ways that make them consistent with the system of notation conventions used in instruction in each country (see OECD—PISA, 2007). In contrast, in the context of EL assessment, the picture is more complex. Should the translation be consistent with the system of conventions used in the cultures associated with the student’s first language? Since most of these students’ schooling history is in the U.S., the answer is *no*, except for contexts in which the goal is to determine the level of proficiency in mathematics for newly-arrived ELs whose recent schooling has been outside the U.S. However, to complicate matters, even in those cases, variations in the systems of conventions used in mathematics may vary considerably across cultures that share the same language. For example, Spanish speaking countries vary considerably in the conventions they use to separate thousands and decimals. A simplistic view of mathematics as a universal language may lead to underestimating the complexities of translation endeavors. Also, translating a test into a language cannot be done appropriately without considering cultural variation in language in the context of a discipline.

The Impossibility of Validly Assessing English Language Arts with Translated Tests

Issues of equivalence across languages are considerably more serious in the case of English Language Arts—to the extent that it is impossible to make valid inferences about EL students' knowledge of this field based on their scores on English Language Arts administered in their native language. The reason is twofold. First, language arts are specific to the language in which they originate. Second, to a large extent, language arts involve more than meaning. Several examples are provided in this section.

The first example shows that there are subtle aspects that cannot be replicated and assessed across languages. Below is an English Language Arts item in English and its Spanish translation:

Read the following text:

Today you are you!
That is truer than true!
There is no one alive
who is you-er than you!

Explain three reasons why you think this text is a poem.

Lee el siguiente texto:

¡El día de hoy tú eres tú!
¡Esto es más cierto que la verdad!
¡No existe nadie
que sea más tú que tú!

Explica tres razones por las que este texto puede ser considerado como un poema.

The rhyming and the metrics of the poem written by Dr. Seuss (1959) have been altered in the Spanish version. Most importantly, the meaning conveyed through the deliberate use of grammatical absurdity (*you-er*, the view of a person as an adjective and, more specifically, a desirable quality) cannot be replicated in Spanish. Thus, while the meaning is kept in the translation of the text, its poetic nature is somewhat lost and the item cannot elicit from the student the analytical reasoning that it can potentially elicit in the English version. A student tested with the Spanish version would have much more difficulty than a student tested with the English version identifying rhyming and metrics, and appreciating the poetic tone of the text, simply because those features have been compromised.

The second example illustrates the challenges of scoring student performance due to the fact that the features that contribute to good writing are not valued in the same way across languages. Imagine a student who has received formal reading and writing instruction in her first language in a bilingual program and who writes an argument essay. This student uses discursive structures that differ from those used in English and are common and highly valued in her first language. For example, she consistently uses circular structures and does not use topic sentences as frequently as they are used in Standard English argument texts.

Due to these differences, the performance of this student would not be assessed fairly by simply using the scoring rubrics originally developed to assess students tested in English. On the other hand, a tremendous amount of work would be needed to develop or adapt scoring rubrics to ensure sensitivity to the characteristics of that student's written home language. Moreover, even if resources and time were not an issue, finding a sufficient number of student responses to use in the process of scoring rubric development and as benchmarks and training and calibration sample responses would be extremely difficult. Finally, needless to say, doing this for multiple languages would be practically impossible.

The third example shows the variety of conceptual issues need to be resolved before students' performance on translated English Language Arts tests can be validly interpreted. Suppose a task consisting of reading a literary passage, and then responding to a series of questions intended to assess understanding of certain words and the ability to interpret metaphors. Should both the literary passage and the questions be translated? If so, should it be assumed that understanding a given metaphor in English is the same as understanding that metaphor in the translated version? What should be done if the meaning and the intention of the metaphor are impossible to capture in the translation?

As stated above, a recurrent issue in EL assessment is that the knowledge of a discipline and the language in which that knowledge is encoded are extremely difficult to separate. As the examples above illustrate, English Language Arts poses an additional challenge for validly assessing EL students—that the knowledge of the discipline is difficult to separate from the natural language from which it originated. Using direct translations of English Language Arts assessment does not appear to produce interpretable scores. More serious than the threat of altering the constructs assessed by some items in the original language version, translated English Language Arts tests may assess a different knowledge domain.

The challenges discussed have serious implications concerning graduation practices—which need to be revised. If taking an English Language Arts assessment is a used as a high school graduation requirement, testing students classified as ELs in their home language might unfairly prevent them from accessing college.

Human Resources

As mentioned above, tests have linguistic features that make them different from other forms of text and which pose a unique set of translation challenges. Because of the challenges that stem from the complex intersection of language, content, and the characteristics of EL populations, careful attention needs to be paid to the selection of translation and translation review procedures to be used. Also, careful attention needs to be paid to the selection of translators and educators charged respectively with translating the instruments and providing ELs with a wide variety of translation-related accommodations and supports.

Regarding translators, it is important to keep in mind that, to a large extent, translation skills are context-bound. Whereas many translation skills are relevant to any kind of text regardless of genre and content, translators need to be familiar enough with the content of the text to be translated in order to be able to properly address register—the nuanced uses of language within a disciplinary area. Also, they need to be knowledgeable of the dialects (varieties) of the target language used by the specific target student populations.

While desirable, formal certification in the profession is not a guarantee that a translator has the skills mentioned above. In addition, for most of the dozens of language used by ELs, certified translators may be impossible to find because translator certificates are available for only a few language combinations. For example, the American Translators Association offer certificates for translators from English into Chinese, Croatian, Dutch, Finnish, French, German, Hungarian, Italian, Japanese, Polish, Portuguese, Russian, Spanish, Swedish, and Ukrainian (American Translators Association, 2016). While fifteen is a good number of target languages, only a few of them are among the dozens of languages most frequently used by ELs. Clearly, the conventional desktop translation approach that involves one or two trusted professional translators is limited in its effectiveness in generating translations of tests that contribute substantially to valid assessment for ELs.

Recent translation models for EL assessment emphasize the use of multidisciplinary translation and translation review teams that include translators, linguists, content experts, teachers who teach students from the target linguistic groups, and members of the speech communities (Solano-Flores, 2012; Solano-Flores, Backhoff, & Contreras-Niño, 2009; Solano-Flores, Contreras-Niño, & Backhoff, 2012). These new models contrast with conventional translation approaches in that they do not rely heavily on the work of one or two translators. In addition, they replace back translation with alternative translation review procedures intended to ensure that the content of tests is preserved across languages. (Note 1) As experience in assessment across languages accrues, mainly from international test comparisons such as PISA (Programme for International Student Assessment), this verification procedure tends to be abandoned in major endeavors involving translation, due to evidence showing that translators tend to "correct" inaccurate translation, thus masking translation errors made in the target language (Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007). Unfortunately, awareness of the limitations of back translation is not sufficiently widespread, and many individuals in charge of test translation projects may still consider it as an effective method for test translation review.

An important consideration regarding human resources and fidelity of implementation is that the selection of translators should not be based on the social status of the dialects of the target language. A common misconception in the field of assessment of linguistically diverse populations is that hiring highly qualified translators suffices to ensuring appropriate translations. Of course, highly qualified translators are always needed. However, quality should not be confused with social status. Attention should be paid to ensure that translations reflect the dialect or dialects used by the target populations, not the prestigious dialect. For example, hiring translators who use Iberian Spanish (Spanish Castellan)—a dialect that many consider as more socially prestigious than other Spanish dialects in the U.S.—to create Spanish test translations would do a serious disservice to EL Spanish users in the U.S., as the majority of them use other Spanish dialects.

Regarding teachers charged with providing a wide variety of translation-related accommodations and forms of supports for EL students (e.g., translating the text of tests partially or totally, reading aloud test directions for students), availability and quality are important issues to consider. First, teachers with the communication skills in the EL students' home language may not exist in the numbers state- or district-level decision makers expect or assume. Second, educators with those skills may not always have the level of skills needed to properly serve these students under testing conditions. Pronunciation and knowledge of both the academic language of the students' first language and the students' dialect in their first language are among the subtle aspect of educators' students' first language proficiency whose importance should not be overestimated. As with professional translators, proper selection and screening process should be put in place to ensure that educators in charge of providing translators have the skills needed.

Validity and Fairness Testing in the Home Language

Types of Test Translation

Up to this point, the discussion of the sources that may lead to inadequate translation practices has not been explicit about any particular form of translation. This section examines different forms of translation and discusses the advantages and limitations of each form from the perspective of validity and fairness.

Typically, the term translation is understood as full text translation, which is probably the most common form of translation in legal or literary text translation. However, in the case of tests, there is a wide variety of translation formats and translation devices that, for the purposes of this discussion, should be included as types of test translation. These translation formats vary from the conventional full text translation to the use of glossaries of specific words to making English-to-first language dictionaries accessible to students.

Table 1 shows some translation formats; they should be thought of as some of the many possible formats that could be or have been used in EL assessment. As seen in the first column, these formats have been classified into two broad categories—full text and partial text translation. For each format, the table indicates the medium (paper-and-pencil; computer) through which a test is administered.

Translation Type/Format	*Safety*	*Sensitivity*	*Fidelity of Implementation*	*Usability*
Full Text				
Monolingual-Printed: The text of each item is provided in the student's first language; the original text in English is not provided (P&P)*	4	4	1	4
Monolingual—Read-Aloud: The test administrator reads each item aloud in the student's first language; the original text in English is not provided (P&P)	4	4	4	4
Bilingual-Printed: The English and first language versions of each item are displayed side-by-side (P&P)	1	3	1	2
Bilingual-Screen: The English and first language versions of each item are displayed one on top of the other (CA)	1	3	1	2
Audio Available: The student has the option to listen a recorded audio version of the item in their first language; text in English provided (CA)	1	3	1	1
Partial Text				
English-to-First Language Dictionary: Student is given a commercially-available printed dictionary with translations of words (P&P or CA)	1	3	3	3
Printed Glossary: Each item shows the translation of selected words or terms next to it (P&P)	1	2	1	1

Translation Type/Format	*Safety*	*Sensitivity*	*Fidelity of Implementation*	*Usability*
Pop-Up Glossary: The text of each item in English highlights some words or terms as available for translation; when the student clicks on a word or term, its translation appears on the screen (CA)	1	1	1	1
Audio Glossary: The text of each item in English highlights some words or terms as available for translation; when the student clicks on a word or term, an audio translation of it is played (CA)	1	1	1	1
Directions Read Aloud in First Language: The test administrator reads the directions of the test aloud in the student's first language; English version provided (P&P)	4	4	4	4

Table 1. *Validity and Fairness Dimensions: Relative Ranking of Ten Test Translation Formats by Dimension (1=highest; 4=lowest). P&P = paper and pencil; CA = Computer-administered*

Adapted from Solano-Flores, G. (2012). Translation accommodations framework for testing English language learners in mathematics. Developed for the Smarter Balanced Assessment Consortium (SBAC). September 18, 2012. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Translation-Accommodations-Framework-for-Testing-ELL-Math.pdf

Five full text formats are shown. Monolingual formats come to mind when test translations are mentioned; two monolingual formats are shown—Monolingual-Printed and Monolingual—Read-Aloud. The Bilingual-Printed format shows the English version of an item and its translation next to it. The Bilingual-Screen format displays the translation below the original version of the item—a format that works better on screens in computer–based assessment, as is the case of Smarter Balanced assessments (see Smarter Balanced, 2016).

Five partial text formats are shown. The English-to-First Language Dictionary format consists of a simple, commercially-available dictionary that is given to students to look for words they do not know. The Printed Glossary format consists of translations of selected words or terms (strings of words) made available for each item. These words or terms are translated according to the context in which they appear. The Pop-Up Glossary shows on the screen of the computer the translation of selected words or terms when a student clicks on these words or terms. Likewise, in Audio Glossary, the recording of the translation of words or terms is played when the student clicks on those words or terms. The last partial text format shown consists of reading aloud the directions of the test.

Validity and Fairness Dimensions

Table 1 also shows the rank ordering of the ten translation formats on each of four validity and fairness dimensions—safety, sensitivity, fidelity of implementation, and usability (see Solano-Flores, 2012). These dimensions shape the effectiveness with which a given translation format provides linguistic support, thus contributing to making valid inferences about an EL student's knowledge and skills based on their performance on a test. For each dimension, a 1 and a 4 indicate respectively the highest (best) and the lowest (worst) ranking of a translation format.

Safety refers to how innocuous or harmless a translation format is for students who do not need it. Assumptions about a given student's proficiency in English may be inaccurate. This may be the case even if those assumptions are informed by tests of English proficiency, as measurement error is inevitable. If a student has been wrongly classified as an EL and is given a test in his home language without the version in English, the translation format may hamper rather than support his performance, which will not reflect his knowledge and the skills in the corresponding domain as accurately as it would do on the test in English. As shown in Table 1, translation formats are safe as long as they provide the original text of the test in English. The lowest safety ranking is for the Monolingual-Printed, Monolingual—Read-Aloud, and Directions Read-Aloud in First Language formats, which do not provide the English version.

Sensitivity refers to the ability of the translation format to react to the actions of the examinee in a way in which it adjusts to her needs. The best rankings are for translation formats that use information technology, thanks to which the computer provides linguistic support for specific terms. In the first place are Pop-Up Glossary and Audio Glossary, which provide translations for specific terms at the students' request (mouse click). In the second place is Printed Glossary, which provides translation for specific terms but not at the request of the student. In the third place are Bilingual-Printed and Bilingual-Screen formats, which react to examinees' actions but they do it only to the extent that the students switch back and forth between languages to access the entire text of the item in one language or the other. Also in the third place is Audio Available, which reacts to the examinees' actions to the extent that they can switch back and forth between language modalities to read or listen to the entire text of the item. Also in third place is English-to-First Language Dictionary, which may or may not react to the actions of the examinees. For example, the examinee may look up a word but the dictionary may not have that word—and if it does, the translation may not be sensitive to the context in which the word is used in the text of the item.

Finally, in fourth place are the Monolingual-Printed and Monolingual—Read-Aloud formats, which do not react at all to the examinee's actions. Also in fourth place is Directions Read Aloud in the First Language, a format that does not react at all to the examinee's actions and does not provide any translation (full or partial) of the text of the items.

Fidelity of Implementation refers to the extent with which the translation can be provided as intended by its creators. It is not possible to be certain about the effectiveness of a translation format if the way in which it is provided depends on the contexts in which students are tested or the idiosyncrasies of the individuals who administer it. Thus, the best ranking regarding fidelity of implementation is for translation formats whose administration does not assume any actions from the administrators. In contrast, a ranking of 3 is given to English-to-First Language dictionary. While this translation format does not require from administrators any other action than handing the dictionaries to the students, fidelity of implementation issues arises from the vagueness of state documents regulating the use of accommodations for ELs. In the absence of detailed guidelines, students from different schools may end up receiving different dictionaries of varying qualities and characteristics. Finally, the lowest ranking is for Monolingual—Read-Aloud and for Directions Read in First Language, as the individuals in charge of reading aloud the directions of the test may vary tremendously in their proficiency in the students' first language.

Usability refers to the extent to which students can use a translation format with ease. Usability depends on the skills needed to properly use and benefit from the translation format and on the extent to which a given EL is accurately assumed to have those skills. In first place in the ranking are Audio Available, Printed Glossary, Pop-Up Glossary, and Audio Glossary, which assume minimal mousing/clicking skills or minimal reading skills in the first language. The second ranking is for the Bilingual-Printed and Bilingual-Screen formats because, in order to benefit from this format, a student needs to be able to identify specific words or phrases she does not understand in English and identify them in the translation—an action whose complexity should not be underestimated. A ranking of 3 is given to the conventional English-to-First Language Dictionary, which assumes alphabetical word searching skills in the student. For some dictionaries, students may even need to be able to interpret or dismiss abbreviations and notes on usage and select the right translation for terms with multiple translations or usages. In fourth place is Monolingual-Printed, which assumes that EL students are able to read in English better than in their first language. Also in fourth place are Monolingual Audio and Directions Read in the First Language, formats which assume that the student has respectively better listening skills in the first language than reading and writing skills in the first language.

Note that these rankings might vary depending on both the specific set of circumstances in which test translation takes place and whether translation is understood as an outcome or a process. For example, the Bilingual-Printed and Bilingual-Screen formats is given the best ranking on the Fidelity of Implementation dimension based solely on considering the extent to which contextual and idiosyncratic factors may produce variations in the ways in which these formats are provided. A lower ranking could be given to these formats if the need for considering text size in test translation is considered in their evaluation. For example, to ensure that the two languages of an English-Spanish bilingual format provide the same kind of support to ELs, the design of the format needs to address the well-known (albeit often neglected) fact that Spanish takes 20%-25% more screen or paper space than English. Otherwise, the developers may end up using a smaller font size for the Spanish side or even reducing the space for students to provide their responses.

Properly acknowledging that these rankings may vary depending on which components of the process of translation are considered, the relative standing of a given format with respect to the other on each dimension tends to be the same; they provide an overall appraisal of the advantages and limitations of the different translation formats. While all of them (or variations of them) have been used in state large-scale assessment programs, some are ineffective and even perilous. Few have high rankings across the four validity and fairness dimensions. As can be seen, testing EL students in their home language cannot be assumed to be more valid or fair if proper actions are not taken to address these dimensions.

Response Processes

Current thinking in the field of assessment places an important value on understanding the cognitive processes underlying the ways in which students respond to items. These processes mainly refer to the cognitive activity elicited by an item. For example, imagine a multiple-choice item intended to assess geometrical reasoning. The item shows some quadrilaterals and four statements on their properties, from which the student has to identify the most accurate. Does the item really elicit geometrical reasoning among test takers? Does that reasoning involve comparing and evaluating the geometrical figures? Do students who are and those who are not proficient in geometry differ in the kind of reasoning and knowledge they use when they respond to the item? Can the problem be resolved only by using geometrical reasoning rather than another reasoning strategy?

The term cognitive validity refers to the extent to which evidence addressing these kinds of questions supports the assumption that the item elicits the mental processes intended by the developer (see Chi, 2006; Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Weir & O'Sullivan, 2011). Since cognitive activity can be inferred, not observed, collecting evidence on the cognitive validity of an item is based on asking samples of individuals to talk aloud while they are engaged in solving a problem. Also, evidence on cognitive validity is obtained by asking students to explain why they took certain actions when they were responding to the item (Ericsson & Simon, 1993; Leighton & Gierl, 2007).

Cognitive validity of test translations needs to be examined routinely if assessment in the home language is to contribute to valid and fair assessment for ELs (Solano-Flores, 2016). There are four main aspects that need examination. First is the obvious fact that limited proficiency in English may prevent students from gaining access to the content of test items. Evidence on cognitive validity should inform decisions concerning assessment in English, in the first language, or in both languages (Solano-Flores & Chia, In Press). In current testing practices, it is not customary to include ELs in the samples of students with whom pilot versions of translated tests are tried out, in spite of the valuable information they could provide on the linguistic challenges of the items and their wording. Underlying this unfortunate neglect may be the misconception that ELs cannot communicate in English. Yet the scant available evidence on this matter indicates that the majority of ELs can participate in talk aloud protocols and cognitive interviews conducted in English on tests administered in English (see Kachchaf, 2011; Kopriva, 2011).

Second, in spite of the large body of literature on cognitive validity, little research has included ELs and even less research has compared response processes when these students are tested in different languages. For example, questions like the ones listed above in the example of the item on geometrical reasoning could be asked to determine whether EL students reason in different ways depending on the language in which they are tested. There is evidence that some ELs may use different terms and even different problem solving strategies depending on the language in which they are tested—a difference that may reflect how comfortable they are using the language resources they have in each language in a formal, academic context (Solano-Flores, Lara, Sexton, & Navarrete, 2001). Also, there is evidence that ELs do not perform consistently on the same set of items administered in two languages. Moreover, there is evidence that ELs do not perform consistently when the same items are administered in different dialects of their first language (Solano-Flores & Li, 2009). This evidence indicates that, owing to the differences in the contexts in which they use and have acquired their two languages, ELs have different sets of strengths and weaknesses not only across languages but also across varieties within a language. This evidence also indicates that items administered in different languages or in different dialects of the same language pose different sets of linguistic challenges to ELs.

Third, examination of cognitive processes may reveal ways in which translations can be improved. A study that examined the think-aloud protocols of expert reviewers when they reviewed test translations of an international test found that their judgments of the translated items did not necessarily correspond to the categories of items classified as differentially and non-differentially functioning (Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013). These findings indicate that rigorous translation review procedures should be used in combination with DIF and other analytical procedures to evaluate the quality of test translations. Similar studies need to be conducted routinely in test translation for ELs, for example, to compare samples of translated items generated by several translation teams and determine which team generates fewer items that are differentially functioning, or to determine the extent to which differences in the dialect of the translated versions may contribute to differential item functioning. Regardless of their formal training and certification, translators may not be able to create translations of items that effectively support ELs to gain access to the content of items if they fail to address the complex interaction of content, dialect, and linguistic heterogeneity (Solano-Flores et al, 2007).

Fourth, while the use of information technology allows offering different translation formats to ELs, it also increases the complexity of test taking for these students. For example, hovering the mouse over a word identified as available for translation, clicking on it, and examining different dialect versions of its translation may increase the cognitive load of an item. Thus, in addition to the knowledge needed to interpret and respond to an item, examining cognitive process for ELs may include consideration of the metalinguistic skills students need to identify words they do not understand and interacting with the user's interface. Among many other, questions that need to be answered in projects involving testing ELs in their home language are: Do ELs use translations in the way test developers intend them to be used? How frequently is a given translation resource used by the student? What is the minimum amount of time students need to respond to an item if they are to use the translation resources made available to them? To what extent does the translation resource help students to gain access to the content of an item? While research intended to answer these questions is beginning to be conducted, policy makers need to be aware that, to ensure proper test translation, these kinds of issues should not be taken lightly—all of them are relevant to validity and fairness.

Operationalizing Home Language-Based Assessment for ELs

Recognizing Uncertainty, Inaccuracy, and Fallibility in EL Assessment Practices

A fundamental limitation of current testing practices for ELs is that they are influenced by deterministic views of language and language groups that follow the compliance needs generated by Civil Rights law that require the identification of a class of individuals without much regard to variation within class. (Note 2) For example, the use of a small number of categories of English proficiency fails to capture the heterogeneity of EL populations and limits the possibility of addressing each student's unique set of needs. Also, impervious to knowledge from the language sciences, test translation practices fail to recognize language variation (e.g., due to dialect differences among test takers or to idiosyncratic differences among translators) as a threat to validity in EL testing. Furthermore, analysis of differential item functioning analysis techniques, which allow detection of biased translated items, assume homogeneity in both the focal and reference groups—an assumption that is inappropriate in the assessment of EL populations (see Ercikan, Roth, Simon, Sandilands, & Lyons-Thomas, 2014).

In addition to these deterministic views, the effectiveness of testing practices is affected by error and uncertainty. For example, some students may be classified in the wrong categories of English proficiency. The reason is that, as with any instrument, tests of English proficiency are not perfect—even the best instruments have a margin of error within which wrong classifications of students are produced. Also, legislation and testing policies that allow the use of test translations can be interpreted in multiple ways concerning translation procedures and the characteristics of translation teams. Furthermore, while differential item functioning analysis techniques are often invoked as a resource that contributes to fair testing, it is not clear whether large-scale assessment programs use them systematically with all items or at least substantial numbers of items. High cost and tight timelines for development make it unlikely for these techniques to be used routinely, for example, to examine bias in translated items compared to their counterparts in the original English version. These high cost and tight timelines pose a limit to the extent to which potentially biased items can be detected and modified or revised.

In related publications (Solano-Flores, 2014; Solano-Flores & Gustafson, 2013), it has been proposed that more valid and fair assessment practices can be developed by recognizing the uncertainty that affects the process of assessment of EL populations at all its stages (e.g., identification of ELs and their native language, development or adaptation of instruments, interpretation of scores), rather than pretending that this uncertainty does not exist. More specifically, it has been proposed that probabilistic views should replace deterministic views of language and linguistic groups.

In the context of EL assessment in the students' home language, a deterministic view assumes that EL populations are homogeneous and does not challenge the assumption that EL individuals may be improperly classified in terms of their English proficiency. This deterministic view also fails to recognize language variation due to the students' dialects and relies and uses a top-down approach to translation, according to which the language used by one translator or a small team of translators is assumed to be as representative of the language used by an entire linguistic group. Moreover, generalizations about the students' skills and knowledge based on the scores may fail to incorporate error due to language as a factor to incorporate in score interpretation.

In contrast, a probabilistic approach assumes linguistic heterogeneity (even among native users of the same language) and fallibility in the categories of English proficiency within which a given EL may have been classified. This probabilistic approach uses a bottom-up approach to translation. Accordingly, translations are created by multidisciplinary translation teams with different kinds of expertise relevant to the content of the test to be translated. Thus, in addition to professional translators, the teams include content experts, teachers who teach students of the corresponding linguistic group, and sociolinguists (who provide expertise on language variation). Also, under a probabilistic view, test translation is used in combination with other accessibility resources, rather than relying on it as the only resource used to support ELs to gain access to the content of test items.

Examples of Practices Based on Probabilistic Views of Language

Pop-up textual and audio glossaries are an example of testing in ELs' home language guided by probabilistic views of language and linguistic group. These forms of accessibility resource are possible due to recent advances in information technology and computer-based assessment systems. Smarter Balanced has designed an interface which makes available for students the translation of selected words or terms (Smarter Balanced, 2016). For each item, a translation team identifies the words or terms that are likely to pose a challenge to EL students and which are irrelevant to the constructs measured (i.e., terms whose translation does not give away the content of the item). (Note 3)

This translation is item-specific—it is not taken from a dictionary and is not assumed to be the same for the same word or term when it appears in a different item. Words or terms available for translation are highlighted. When a student clicks on one of these highlighted words or terms, a translation pops up on the screen, showing, when appropriate, different dialect versions of the translation. The translation is sensitive to the grammatical context in which the target word or term appears in the text in English. For example, the translation of a noun appears in plural form if it appears in plural in the original text in English.

An audio version of this accessibility resource is available and can be used in combination with the pop-up version. This allows students to use the textual modality for some words and the audio modality for other words, or to access the translation of a word or term in one modality when the other modality is not helping them. This flexibility is important, as the reading skills in the first language may vary considerably across ELs. As can be seen, what makes this interface sensitive to the unique set of linguistic skills of each student is its ability to react to their actions.

A second example of use of probabilistic views concerns the evaluation of the effectiveness of test translations. In the conventional approach to evaluating the effectiveness of test translations, the performance of non-EL students is compared to the performance of EL students who are tested in English and to the performance of EL students who are tested with a translation of the test. Ideally, effectiveness would be observed when: (a) EL students tested with the translation score higher than their EL peers tested in English and (b) score differences between EL students tested with the translation and non-EL students are smaller than score differences between EL students tested in English and non-EL students. While this kind of comparison is necessary and should be conducted routinely, it has the limitation that it requires the use of large random samples of students to control for multiple extraneous factors—an action that is difficult to perform under tight timelines and budget restrictions.

A more powerful approach to evaluate effectiveness consists of deliberately examining the amount of measurement error obtained when EL students are tested with and without the translation. Generalizability theory, a psychometric theory of measurement error (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), makes this possible. Unlike the other two psychometric theories—classical theory and item response theory—generalizability theory allows partitioning measurement error into different sources of error.

In a perfect world, the variation of test scores observed among students would be attributable only to differences in their knowledge or skills. In the real world, test scores vary not only because of differences in the students' skills and knowledge but also because of many factors beyond control, such as differences in the characteristics of items that are irrelevant to the constructs measured (e.g., some items are unnecessarily more complex than others), and characteristics of the individuals who score their responses. Also, test scores vary across occasions—performance is not stable in time due, for example, to variations in students' moods or level of engagement or fatigue.

Generalizability theory allows examination of the extent of which the score variation observed is attributable to student (s)—the object of measurement—and facets (factors) such as item (i), rater (r), or occasion (o), and the interaction of all these sources (s x i, s x r, s x o, i x r, i x o, s x i r, and s x i x r x o). The theory also allows estimation of the extent to which the scores obtained can be regarded as generalizable—the extent to which appropriate generalizations about the students' knowledge can be made based on the scores of the specific test. In a simple design, if the translation is effective in supporting the students to gain access to the content of items, the magnitude of measurement error due to the main and interaction effect of facet, the item should be smaller for EL students tested with the translation than EL students tested without the translation (Solano-Flores, 2017; Solano-Flores & Li, 2013). The importance of this approach lies in the fact that it allows examination of validity and fairness based on the psychometric properties of the instrument (i.e., the validity of generalizations of test scores) not on comparing ELs with their non-EL counterparts.

Supporting Multiple Languages

In implementing a policy that mandates testing ELs in their home language, states face challenges that result from the wide variety of first languages used by EL populations in the U.S. For example, there are languages that have so small numbers of users or whose characteristics are so unique that it may be extremely difficult to find adequate translators. To some extent, those in charge of assembling translation teams may not have a way of ensuring that the individuals they are hiring for the job have the proper skills or qualifications. Ultimately, these challenges are relevant to assessment validity and fairness.

The number of speakers of languages other than English in the U.S. is disparate and these speakers' distribution is complex. The most common native language among ELs is Spanish (used by 71 percent of the EL students), followed by Chinese (4 percent). The remaining 25 percent of EL are constituted by users of many other languages (Ruiz-Soto, Hooker, & Batalova, 2015). In addition to this disproportionality, the ELs' first languages are not equally distributed throughout the country. Thus, in five states, the languages most frequently spoken by ELs are not Spanish. Also, trends in the distribution of ELs across the country are changing. ELs are now more widely distributed across states, with rapid growth in "non-traditional" states such as Arkansas, Kentucky, North Carolina, South Carolina, and Tennessee, contrasting to an earlier period when they were concentrated in a smaller number of big states such as California, Texas, New York, and Florida.

Given this complex linguistic makeup, careful, long-term planning is needed to determine which languages can be supported, especially because ESSA requires states to test EL students in their home language "to the extent practicable," ruling out low-incidence languages. More specifically, states need to make reasoned decisions on their priorities concerning the languages that are to be supported.

While the number of users is the criterion that first comes to mind to establish these priorities, there are other criteria that are critical to making fair decisions. Table 2 identifies factors that should be taken into consideration in deciding which languages are to be supported by an assessment system. These factors are part of a conceptual framework on item accessibility and language variation developed for Smarter Balanced (Solano-Flores, Shade, & Chrzanowski, 2014). One set of factors is called *relevance*, which includes frequency, proportionality, and criticality. Relevance factors define why a language is important to support. For example, while a high sheer number of users may be relevant to translating tests into a given language, another language may be relevant because, although its users are few, they are an ethnic or socioeconomic group vulnerable or historically underrepresented.

A second set of factors is called *viability*, which includes sustainability, the availability of human resources, cost, dependability of information, and fidelity of implementation. Viability factors define when the efforts to translate tests into a given language are likely to be successful. For example, for certain languages, assembling teams of translators can be a formidable task. Also, there should be a critical mass of potential users who can qualify to participate in translation teams, so that it is possible to make long-term projections of the translation efforts.

It is important to note that the reasoning behind these principles is not naive about the practical limits of test translation. Supporting all the EL students' first languages in the immediate short term may not be possible or sustainable, even if states have the best sets of financial resources at their disposal.

A form of an accessibility resource, currently being developed by Smarter Balanced, may contribute to support EL students whose home languages are difficult to support for the reasons discussed above. This resource consists of illustrations that pop up on the screen (in the same way translations of words or terms can pop up) when students click on the words or terms flagged as available for illustration. Since the illustrations offered are the same regardless of the students' home language, they may potentially become a cost-effective accessibility resource. While their effectiveness and technical properties—and the procedures for their proper design and production—are currently being investigated, available evidence indicates that the design of illustrations is a delicate process. Illustrations need to be conceived, scripted, crafted, and refined according to rigorous design specifications (Solano-Flores, Wang, Kachchaf, Soltero-Gonzalez, & Nguyen-Le, 2014). For example, in order to not increase the cognitive load of items unnecessarily, illustrations should be extremely simple. Also, in order to properly support EL students without giving away the content of items, the words and terms that can be illustrated need to be identified through a careful analysis of the linguistic features of items and their content.

Relevance	
Frequency	• *Sheer number of users*
Proportionality	• *Percentage of users with respect to users of other languages*
Criticality	• *Extent to which the language is used by a historically underrepresented group* • *Vulnerability of the group of users due to poverty or segregation* • *Limited access of the group of users to social programs* • *Scarcity of indicators of academic achievement for the group of users* • *Prevalence of low academic achievement among users of the language*
Viability	
Sustainability	• *Extent to which translation team members (including native speakers of the language) can keep doing translation work for a long time in the future* • *Percentage of students schooled in the language* • *Existence of a critical mass of teachers users of the language*
Human Resources	• *Availability of sufficient numbers of individuals who can act as translators* • *Ease with which translators can be identified and recruited*
Cost	• *Existence of financial resources needed to develop the translations* • *Existence of a well-established logistics for developing the translations*
Dependability of Information	• *Trustworthiness of the information about the language* • *Trustworthiness of the numbers of users*
Fidelity of Implementation	• *Extent to which the translation can be created according to established procedures* • *Availability of resources for evaluating and refining the implementation of the translation procedures*

Table 2. *Criteria for Determining Priorities in EL Students' Home Languages to Support: Relevance and Viability*

Adapted from Solano-Flores, G., Shade, C., & Chrzanowski, A. (2014). Item accessibility and language variation conceptual framework. Submitted to the Smarter Balanced Assessment Consortium, (pp. 57 and 60). October 10. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/11/ItemAccessibilityandLanguageVariationConceptualFramework_11-10.pdf

Assuming that these design requirements are properly met, pop-up illustrations rank along with pop-up glossaries and audio glossaries as the best translation formats do in their ability to meet the fairness and validity dimensions discussed above (see Table 1). It has to be noted that, while regarding illustrations as a form of translation may be unacceptable to some, current thinking in the field of semiotics holds that text and image are interacting, rather than separate, independent sets of semiotic modalities (Kress, 2010). But even if text and image are thought of as totally different, from a semiotics perspective, translation consists of representing information in a different semiotic modality. This includes representing text (or portions of text) in another language, in audio, or in images. According to this reasoning, illustrations are as defensible as a form of test translation as audio translations are as forms of translations for ELs—although with a different set of possibilities and limitations.

Final Remarks

New ESSA legislation opens up a new opportunity towards more fair and valid assessment for EL students by bringing to attention the need for assessment through the home language and reinforcing (as in prior legislation) the requirement for it to the extent practicable. However, the experience in test translation for ELs is relatively limited. As shown in the paper, many conventional translation practices (e.g., full text translation made by one or two translators under a tight timeline) may comply with legislation. Yet, given the vagueness of the wording of this legislation, they may fail to support valid interpretations of test scores for ELs.

Assessment in the home language of EL students should be considered accordingly within a broader view of assessment. This view should include the use of assessment through the native language for the traditional purposes of annual accountability, but also help conceptualize further the purpose of an assessment and accountability system for states and districts, especially for continuous improvement of systems, schools, and ultimately instruction.

From the perspective of accountability, the language used in legislation may be considered as straightforward and clear (e.g., “in a valid and reliable manner,” to “yield accurate data ... until such students have achieved English language proficiency”). But technically, many interpretations are possible, and each addresses different sets of students’ needs. For EL students in English-only and transitional programs, proper “assessing in the home language” needs to be conceived in multiple ways; for example, as providing partial text translation formats, such as pop-up and audio glossaries. For EL students in bilingual programs, proper “assessing in the home language,” for example, through full text translations such as Bilingual-Printed and Bilingual-Screen, should be regarded as viable only when those bilingual programs have a strong commitment and the means and resources necessary to support bi-literacy and the continuing development of the first language. “Assessing in the home language” in those programs can also be conceived of as a form of support for teachers to formatively assess their EL students.

To properly interpret and implement ESSA, decision makers and practitioners need to ensure that the translation practices they support are effective in addressing the nature of language development, the diverse linguistic demands of each disciplinary content area, and the multiplicity of individual schooling histories of ELs. Accordingly, assessment programs need to establish new sets of requirements for contractors and vendors in charge of creating test translations. More specifically, assessment programs need to:

1. Allocate the financial resources needed to create and offer those translation formats that have the highest levels of safety, sensitivity, fidelity of implementation, and usability;

2. Establish timelines that allocate reasonable amounts of time for the process of test translation as an important component in the overall process of test development;
3. Require the use of test translation and translation review procedures that involve teams comprised of not only professional translators but also educators who teach EL students the corresponding content areas, professionals who are familiar with the use of the target language in those content areas, and individuals who are native users of the target language;
4. Allocate resources to provide professional development for educators to support them to make appropriate decisions concerning the sets of accessibility resources that are appropriate for each of their EL students both instructionally and for assessment purposes; and
5. Commission research that informs decisions on issues yet to be resolved, such as test translation into low incidence languages or the extent to which constructs are altered in cross-language, cross-semiotic modality translation (e.g., translation of printed English into audio recording in the target language and translation of printed English into illustrations).

Ensuring valid, fair assessment for ELs entails much more than complying with ESSA legislation. It requires committing to the principles of social justice underlying that legislation and thinking critically and creatively about what test translation is.

Notes

Note 1. In the back translation method, the translated version of a test is translated back to the original language. Then the original and back translated versions of the test are compared to determine if the content is preserved across versions. Discrepancies are resolved by making the proper modifications in the target language version.

Note 2. See Department of Justice Civil Rights Division and Department of Education Office for Civil Rights Dear Colleague Letter (DCL) Jan. 7, 2015. Retrieved at http://www2.ed.gov/about/offices/list/ocr/letters/colleague-el-201501.pdf

Note 3. For a first-hand experience with the interface, visit the Smarter Balanced Practice Test: https://practice.smarterbalanced.org/

References

Abedi, J. (2006). Language issues in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-420). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

American Federation of Teachers (2002). *Teaching English-language learners: What does research say?* AFT Policy Brief Number 14.

American Translators Association (2016). *A guide to the ATA certification program*. Retrieved on December 21, 2016 from https://www.atanet.org/certification/aboutcert_overview.php

Arffman, I. (2013) Problems and issues in translating international educational achievement tets. *Educational Measurement: Issues & Practice, 32*(2), 2–14.

Batalova, J., Fix, M., & Murray, J. (2005). *English language learner adolescents: Demographics and literacy achievements*. Report to the Center for Applied Linguistics. Washington, DC: Migration Policy Institute.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1) pp. 7–71.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Bybee, J. L., & Dahl, O. (1989). The creation of tense and aspect systems of the world. *Studies in Language, 13*(1), 51-103).

California Department of Education (2014). *English Language Arts / English Language Development Framework* for California public schools: Kindergarten through Grade twelve (2014). Sacramento, CA: California department of Education.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance*. Cambridge, NY: Cambridge University Press.

Cook, H. G., Boals, T., & Lundberg, T. (2011). Academic achievement for English learners. What can we reasonably expect? *Phi Delta Kappan*, *93*(3), 66-69.

Coulmas, F. (2013). *Sociolinguistics: The study of speakers' choices. 2nd Edition*. New York, NY: Cambridge University Press.

Council of Chief State School Officers. (2012). *Framework for English Language Proficiency Development Standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: CCSSO.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Darling-Hammond, L., Wilhoit, G., & Pittinger, L. (2014). *Accountability for college and career readiness: Developing a new paradigm*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Dr. Seuss. (1959). *Happy birthday to you*. Random House, NY.

Edwards, J. (2009). *Language and identity: An introduction*. Cambridge, UK: Cambridge University Press.

Edwards, J., Gross, M., Chen, J., Macdonald, M., Kaplan, D., Brown, M., & Seidenberg, M. (2014). Dialect awareness and lexical comprehension of mainstream American English in African American English–speaking children. *Journal of Speech Language and Hearing Research,57*(5), 1883-1895.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education, 27*(4), 273-285.

Ericsson, K.A., & Simon, H.S. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts: The MIT Press.

Fishman, J. A. (1965). Who speaks what language to whom and when? *La Linguistique*, *2*, 67-88.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education / Revue Canadienne de l'Education, 25*(4), 280-296.

Gottlieb, M., Cranley, M. E., & Cammilleri, A. (2007). *Understanding the WIDA English language proficiency standards: A resource guide*. Madison, WI: WIDA Consortium.

Grisay, A., de Jong, J. H. A. L. Gebhardt, E. Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249-266.

Grosjean, F. (2001). The bilingual’s language modes. In Janet Nicol (Ed.), *One Mind, Two Languages: Bilingual Language Processing*. Oxford: Blackwell. 1–23.

Hakuta, K., Butler, Y. G., & Witt, D. (2001). *How Long Does It Take English Learners to Attain Proficiency?* (Policy Report 2000-1). The University of California Linguistic Minority Research Institute.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement, 3rd Edition* (pp. 147-200). New York: American Council on Education/Macmillan Publishing Company.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hopkins, M., Thompson, K. D., Linquanti, R., Hakuta, K., & August, D. (2013). Fully accounting for English learner performance: A key issue in ESEA reauthorization. *Educational Researcher, 42*(2), 101-108.

Hopewell, S., & Escamilla, K. (2013). Struggling reader or emerging biliterate student? Reevaluating the criteria for labeling emerging bilingual students as low achieving. *Journal of Literacy Research, 46*(1), 68-89.

Kachchaf, R. R. (2011). *Exploring problem solving strategies on multiple-choice science items: comparing native Spanish-speaking English language learners and mainstream monolinguals*. Unpublished doctoral dissertation. University of Colorado Boulder.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*(6084), 1049-1054.

Kenner, C., & Kress, G. (2003). The multisemiotic resources of biliterate children. *Journal of Early Childhood Literacy, 3*(2), 179-2-2.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services: 2000-2001* Summary Report. Prepared for OELA by the National clearinghouse for English Language Acquisition & Language Instruction Educational Programs. Washington, DC.

Kopriva, R. J. (Ed.) (2008). *Improving testing for English language learners*. New York: Routledge.

Kopriva, R. (2001). *ELL validity research designs for state academic assessments: An outline of five research designs evaluating the validity of large-scale assessments for English language learners and other test takers*. Paper prepared at the Council of Chief State School Officers Meeting, Houston, TX, June 22–23, 2001.

Kress, G. (2010). *Mutimodality: A social semiotic approach to contemporary communication*. New York: Routledge.

Leighton, J., & Gierl, M. (2007). Verbal reports as data for cognitive diagnostic assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (146-172). Cambridge University Press.

Lemke, J. L. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 87-113). New York: Routledge.

Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. In M. Andeson, A. Sáenz-Ludlow, S. Zellweger, S., & V. V. Cifarelli (Eds.). *Educational perspectives on mathematics as semiosis. from thinking to interpreting to knowing* (pp. 215-234). Ottawa, Legas.

Linquanti, R., & Cook, H. G. (2013). Toward a “Common Definition of English Learner”: Guidance for States and State Assessment Consortia in Defining and Addressing Policy and Technical Issues and options. Washington, DC: Council of Chief State School Officers.

Mackey, W. (1968). The description of bilingualism. In J. Fishman (ed.), *Readings in the sociology of language*, pp. 554-584. The Hague: Mouton.

National Center for Education Statistics. (2016a). *The Condition of Education 2016* (NCES 2016-144), English Language Learners in Public Schools.English Language Learners in Public Schools. Retrieved, November 2, 2016, https://nces.ed.gov/fastfacts/display.asp?id=96

National Center for Education Statistics. (2016b). *National Assessment of Educational Progress (NAEP), selected years, 2002–11 Reading Assessments, NAEP Data Explorer*. Cited in English language learners. Retrieved, November 2, 2016, https://nces.ed.gov/programs/coe/pdf/...CGF/COE_CGF_2013_05.pdf

Organization for Economic Cooperation and Development—Programme for International Student Assessment (2007). PISA 2009 translation and adaptation guidelines. Doc: NPM (0709)1. Dubrovnik, Croatia: National Project Managers’ Meeting.

Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics, 28*, 191-230.

Phillips, B. (2006). Word frequency and lexical diffusion. New York: Palgrave Macmillan

Public Law 103-382 (1994). *The Improving America's Schools Act of 1994.*

Rennie, J. ESL and Bilingual Program Models. ERIC Digest. ERIC Clearinghouse on Languages and Linguistics Washington DC.

Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. *International Journal of Science Education, 35*, 546–576.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99-141.

Ruiz-Soto, A. G., Hooker, S., & Batalova, J. (2015). Top languages spoken by English language learners nationally and by state. Migration Policy Institute, ELL Information Center Fact Sheet Series, No. 4. Retrieved, October 15, 2016 from http://www.migrationpolicy.org/research/top-languages-spoken-english-language-learners-nationally-and-state

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*, 148-166.

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review or Research in Education, 39*, 215-252.

Smarter Balanced (2016). *Smarter Balanced Assessment Consortium: Usability, Accessibility, and Accommodations Guidelines*. Prepared with the assistance of the National Center on Educational Outcomes. Retrieved, November 2, 2016, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwi4gfW1r5PQAhVkrlQKHSKICk4QFggbMAA&url=http%3A%2F%2Fwww.smarterbalanced.org%2Fwp-content%2Fuploads%2F2015%2F09%2FUsability-Accessibility-Accommodations-Guidelines.pdf&usg=AFQjCNHa9eRUoHxNLMIGiJEyN7gnfuI8gA&bvm=bv.137904068,d.cGw&cad=rja

Solano-Flores, G. (2009). The testing of English language learners as a stochastic process: Population misspecification, measurement error, and overgeneralization. In K. Ercikan & W. M. Roth (Eds.), *Generalizing from Educational Research* (pp.33-48). New York: Routledge.

Solano-Flores, G. (2012). *Translation accommodations framework for testing English language learners in mathematics*. Developed for the Smarter Balanced Assessment Consortium (SBAC). September 18, 2012. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Translation-Accommodations-Framework-for-Testing-ELL-Math.pdf

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education. 27*(4), 236-247.

Solano-Flores, G. (2016). *Assessing English language learners: Theory and practice*. New York: Routledge.

Solano-Flores, G. (2017). Chapter 46: Generalizability**,** In D. Wyse, N. Selwyn, E. Smith, & L. E. Suter, (Eds.), *Handbook of Educational Research* (pp. 937-956). London, UK. Sage.

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

Solano-Flores, G., & Chía, M. (In Press). Multiple language versions of tests. In K. Ercikan & J. Pellegrino (Eds.)**,** *Validation of score meaning in the next generation of assessments.*

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2012). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (Eds.), *Research in the Context of the Programme for International Student Assessment*. Springer Verlag.

Solano-Flores, G., & Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving Large Scale Assessment in Education: Theory, Issues, and Practice* (pp. 87-109). New York: Taylor & Francis: Routledge.

Solano-Flores, G., Lara., J., Sexton, U., & Navarrete, C. (2001). *Testing English language learners: A sampler of student responses to science and mathematics test items*. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice 25*(1), 13-22.

Solano-Flores, G., & Li, M. (2009a). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

Solano-Flores, G., & Li, M. (2009b). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment, 14, 1-15.*

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, *19*(2-3), 245-263.

Solano-Flores, G., Li, M., Speroni, C., Rodriguez, J., Basterra, M., & Dovholuk, G. (2007). *Comparing the properties of teacher-adapted and linguistically-simplified test items for English language learners*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL. April 9-13.

Solano-Flores, G., Shade, C., & Chrzanowski, A. (2014). *Item accessibility and language variation conceptual framework. Submitted to the Smarter Balanced Assessment Consortium.* October 10. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/11/ItemAccessibilityandLanguageVariationConceptualFramework_11-10.pdf

Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment 19*, 267–283.

Thompson, K. D. (2015). English learners' time to reclassification: An analysis. *Educational Policy*. Advance online publication. DOI: 10.1177/0895904815598394

Turkan, S., & Oliveri, M. E. (2014). *Considerations for providing test translation accommodations to English language learners on Common Core Standards-based assessments*. Educational Testing Service: Research Report ETS RR—14-05.

Valdés, G. (2015). Becoming bilingual and multilingual: Language acquisition and development. In G. Valdés, K. Menken, & M. Castro. (Eds.), *Common Core, Bilingual and English Language Learners: A Resource for Educators* (p. 38). Philadelphia, PA: Caslon, Inc.

Weir, C. J., & O'Sullivan, B. (2011). Language testing validation. In B. O'Sullivan (Ed.), Language testing: Theories and practices (pp. 13-32). Basingstoke, UK: Palgrave Macmillan.

Wolfram, W., Adger, C.T., & Christian, D. (1999). *Dialects in schools and communities*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers. Ch. 1: Language variation in the United States. pp. 1-34.

Zong, J., & Batalova, J. (2015). The limited English proficient population in the United States. *Migration Information Source*. Retrieved, November 2, 2016, http://www.migrationpolicy.org/article/limited-english-proficient-population-united-states