8

Comparability in Multilingual and Multicultural Assessment Contexts

Kadriye Ercikan, *Educational Testing Service/University of British Columbia*
Han-Hui Por, *Educational Testing Service*[1]

CONTENTS

INTRODUCTION

A basic tenet of the validity of inferences from assessments is that scores reflect the underlying knowledge and abilities that the test is designed to measure, and the score meaning is consistent for individuals from different language and sociocultural backgrounds. The validity of interpretation of performance on assessments in multicultural and multilingual contexts is critically tied to whether (1) the assessments are tapping the

[1] The authors are grateful to Randy Bennett, Maria Elena Oliveri, and Donald Powers for reviewing an earlier draft of this chapter; to Guadalupe Valdés and Christian Faltis for their feedback; and to Emily Pearce for editorial support.

knowledge and skills we are interested in assessing, (2) the constructs being assessed are comparable for different sociocultural groups, and (3) the scores are comparable across languages and cultures (Ercikan & Lyons-Thomas, 2013). These criteria for score comparability are at the heart of fairness in the interpretation and use of assessment results; they require us, as assessment developers, users, and specialists, to examine and verify what constructs the assessments are targeting and whether they are assessing the same construct with the same psychometric properties for different groups.

Ensuring that assessments provide consistent score meaning is crucial when students have different language and sociocultural backgrounds. In international assessments of learning outcomes, such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA), consistency of score meaning across countries, languages, and cultural groups is central to making accurate and meaningful inferences. Chapter 5 (Comparability Across Different Assessment Systems) elaborates on the validity of the comparisons made across different assessment systems. In addition to international assessments, the issue of consistent score meaning is also a concern for countries with populations from diverse language and sociocultural backgrounds such as in countries with large immigrant populations. In the United States, students have varied sociocultural and language backgrounds, with large proportions speaking a language at home and in their community that differs from the language used in school. The recognition of diversity and its implications for validity and fairness led states to develop assessments in multiple languages and to provide language tools and accommodations to assess students' performance. For example, in New York State, mathematics assessments are adapted into such student home and community languages as Spanish, Traditional Chinese, Haitian, Korean, and Russian (Tabaku, Carbuccia-Abbott, & Saavedra, 2018).

The purpose of this chapter is to highlight the complexity of comparability issues when assessments are administered in multiple languages to students from diverse backgrounds, to describe research on the comparability of assessments and scores, and to discuss guidelines and processes in optimizing comparability of multiple language versions of assessments. The first section describes the sociocultural and language diversity in the United States and countries around the world and discusses the impact of such diversity on interpretation and use of assessment results. The next section introduces the concepts of measurement equivalence and the methodologies used in examining measurement equivalence and score comparability. The third section describes the guidelines for optimizing score comparability across adapted versions of assessments and provides recommendations to create comparable scores. The final section addresses score comparability challenges and potential solutions for the next-generation assessments that involve administration in digital environments.

Throughout this chapter, we distinguish between *adaptation* and *translation*, with the latter term used to denote creating different language versions with a focus on linguistic equivalence. *Adaptation*, however, refers to the broader process of creating language versions that may include changes made to create greater cultural relevance in addition to linguistic equivalence. The term *adaptation* is preferred in assessment contexts because the task goes beyond the literal translation of the assessment content and more accurately reflects the process that is expected to lead to a greater validity of the assessment for the targeted populations. Furthermore, *measurement/score comparability* and *measurement equivalence* are used interchangeably to broadly define the comparability aspects of assessments that include comparability of score interpretation as well as statistical notions of measurement equivalence.

SOCIOCULTURAL AND LANGUAGE CONSIDERATIONS IN ASSESSMENT

Sociocultural and language diversity is a reality shared by all countries around the world. In the U.S. context, information about languages spoken has been obtained from Census respondents since 1980. According to the Census data, the estimated percentage of people speaking *only* English at home has steadily fallen, declining from 89.1 percent in 1980 to 78.2 percent in 2017. Other widely used languages include Spanish (41.0 million), Chinese (3.5 million), Tagalog (1.7 million), Vietnamese (1.5 million), Arabic (1.2 million), French (1.2 million), and Korean (1.1 million) (U.S. Census Bureau, 2017a). In the 2017 Census, 80 percent of the school-age children who spoke a different language at home and in their communities also spoke English "very well"[2] (U.S. Census Bureau, 2017b). The extent of language diversity raises the issue of how best to assess students in ways that lead to valid and fair interpretation and use of assessments. In particular, questions arise regarding what language students should be tested in: the home/community language versus the language of schooling? Or should both languages be included in the assessment (López, Turkan, & Guzmán-Orth, 2017)? What kinds of language accommodation tools should be provided to students, and how can score comparability be established across multiple language versions of assessments?

The inherent differences between languages can make test adaptation a challenging task. Adapted versions of assessments must reflect equivalent meaning, format, relevance, intrinsic interest, and familiarity of the item content (Ercikan, 1998, 2003; Hambleton, Merenda, & Spielberger, 2005). Languages vary in the frequency of word use and word difficulty. Moreover, grammatical forms in one language may not have equivalent forms in other languages, or may possibly have many of them. There is also the difficulty of adapting syntactical style from one language to another. Languages may also differ in form (alphabet versus character based) and direction of scribe (left-to-right, right-to-left, or top-to-bottom). Some languages, such as German and French, also require much more text to convey the same intended meaning compared to English.

Sociocultural factors must also be taken into account when developing assessments and in interpretation and use of scores (Geisinger, 1994; McQueen & Mendelovits, 2003; Wu & Ercikan, 2006). Learners from diverse sociocultural backgrounds have different experiences with schooling and learning. For example, cultural norms can affect learning and world views, including how success is perceived. How students are taught, and how achievement is defined in the educational system, generally reflects the perspectives of the mainstream society and not necessarily those of all its cultural groups. For example, Yup'ik children in rural Alaska learn critical community practices, such as fishing and navigation, from observing and participating in these activities with experienced adults. Because verbal interactions are part of this key learning process,

[2] The available response options to the survey item "How well does this person speak English?" were "Very well," "Well," "Not well," and "Not at all."

a school system that expects passive listening with little interactions may put these students at a disadvantage (Lipka & McCarty, 1994).

Diverse sociocultural backgrounds also include the students' socioeconomic status (SES), which affects their experiences with schooling and learning, which in turn affect how the students interact with assessments. Research has provided ample evidence that higher SES is associated with higher achievement (Berliner, 2012; Lee & Burkam, 2002; Perry & McConney, 2010; Sirin, 2005; Tate, 1997), such as in reading (Grover & Ercikan, 2017; Silva, Verhoeven, & van Leeuwe, 2011). The positive association between SES and achievement is consistent with its relations to other life outcomes, including health status (McEniry, Samper-Ternent, Flórez, Pardo, & Cano-Gutierrez, 2019). High-SES students, as a group, attend schools that provide them with more resources, such as a higher teacher-to-student ratio (Shifrer & Fish, 2019). Indeed, SES has been shown to affect students' access to academic preparations (Carnevale & Rose, 2003), the identification and availability of aids for students with learning disabilities (Elder, Figlio, Imberman, & Persico, 2019), opportunity to learn (Bachman, Votruba-Drzal, El Nokali, & Castle Heatly, 2015; Blömeke, Suhl, Kaiser, & Döhrmann, 2012), and eligibility for test accommodations (Zirkel & Weathers, 2016). SES is also related to one's eligibility for unwarranted time extensions. Quealy and Shapiro (2019) reported that white, middle-class students were much more likely than students of other races and SES backgrounds to receive section 504 plan provisions, unfairly allowing students twice as much time to complete the New York specialized high school entrance examinations. The interaction between assessments and students' socioeconomic backgrounds suggests that taking these contexts into account in developing and interpreting assessment results can prevent further disadvantaging teachers and students from low-SES schooling contexts. A more detailed discussion of the disparities in access to test accommodations is found in Chapter 7, Comparability When Assessing Individuals with Disabilities.

Issues of displacements should also be taken into consideration when interpreting and using assessment results. Studies have shown that students who are displaced—whether due to school closures (Kirshner, Gaertner, & Pozzoboni, 2010) or natural disasters (e.g., Hurricane Katrina; Ward, Shelley, Kaase, & Pane, 2008) or because they are refugees (Gahungu, Gahungu, & Luseno, 2011)—showed declines in academic performance. In these instances, scores from these students may not truly reflect their actual ability, and their scores have to be interpreted with those considerations.

Impact of Sociocultural and Language Factors on Score Interpretation and Use

An important issue often overlooked is that students' performances on assessments are the outcomes of complex interactions between the students' experiences (such as their language and sociocultural backgrounds) and the knowledge and skills targeted by the assessment and other properties of the assessment. When students interact with or participate in assessments, multiple psychological, social, and cognitive factors are at play, including their home/community language; their familiarity with contexts, objects, words, and how students relate to them; how they function in an assessment context; and their anxiety levels, among many other affective and conative variables (Snow, 1993). Solano-Flores and Nelson-Barber (2001) added other sociocultural influences prevalent in cultural groups, such as communication patterns and socioeconomic conditions.

While responses to assessment questions may reflect what students know and can do, their cultural and language practices outside of school also affect how they interpret assessment questions and formulate their responses. Ercikan, Roth, Simon, Sandilands, and Lyons-Thomas (2014) noted that the nature and frequency of access to white mainstream cultural practices outside of school contribute to students interpreting items differently. Parents who do not understand the purpose or nature of formal assessments or who have not been invited into mainstream school cultural practices may also contribute to students' confusion and to an increase in test anxiety. Likewise, teachers who are unable to communicate well with the parents or all of the children in their classes due to language differences may also contribute to students' confusion and test anxiety. Furthermore, differences in sociocultural norms can lead to differences in how students engage with assessments and what responses they provide to test items (Solano-Flores, Lara, Sexton, & Navarrete, 2001). For example, students from some cultural groups may have been socialized to not provide lengthy or elaborate answers to interviewers who are considered to be of higher status or maturity, or they may hold back when indicating confidence and success level or when being requested to disclose personal information. Possible sociocultural differences may also be found in motivation, experience with psychological assessments, and speed of responding (Talento-Miller, Guo, & Han, 2013).

Another important consideration that can affect the students' assessment performance is their access to opportunities to learn and engage with similar kinds of assessments (Ercikan, Roth, & Asil, 2015). The opportunity to learn the curricular content which is subsequently assessed, develop test-taking strategies, and become familiar with the assessment technology, as in the case of digitally-based assessments, can all contribute to the students' ability to engage with the assessment. Hambleton (2005) highlighted item format as a potential threat to score comparability, which suggests that currently novel item types (e.g., hot zone selection, drag and drop) should be used with caution when assessing students with nonwhite and nonmainstream language and sociocultural backgrounds to minimize introducing construct-irrelevant demands in assessment questions. Access to test preparation classes or test time extension can further contribute to measurement incomparability. Given that the scores should reflect primarily the competencies being assessed (Messick, 1989, 1995), the interpretations of scores across different language and sociocultural groups should account for the differences in access to, participation in, and benefit from learning opportunities.

Policies Addressing Language Diversity

The recognition of language and cultural diversity among the student population in the United States and its implications for validity and fairness moved states to provide language tools and accommodations to assess students' performance. These language tools and accommodations are intended to optimize student performance and minimize the impact of language proficiency on performance on assessments that are not intended to assess language proficiency. Policies dealing with language diversity within the United States vary widely among jurisdictions (Tabaku et al., 2018). The

Smarter Balanced Assessment Consortium provides language support for students in the form of glossaries and translations of test directions and items in several languages commonly spoken as home and community languages. Standard language glossaries are available in Spanish, Arabic, Cantonese, Mandarin, Filipino (Ilokano and Tagalog), Korean, Punjabi, Russian, Ukrainian, and Vietnamese (Smarter Balanced, n.d.). However, in each state students' access to such tools and support depends on state laws and regulations.

Currently, some states allow alternative standard language versions of some assessments (New York State Education Department, 2016; Ohio Department of Education, 2018, 2019; Oregon Department of Education, 2019). In New York, alternative languages include Spanish, Chinese, Haitian Creole, Russian, Polish, Korean, Bengali, Arabic, Urdu, Vietnamese, Amharic, Portuguese, and several others. Other states translate test directions, but not the assessment itself, into commonly spoken home/community languages (e.g., South Carolina Department of Education, n.d.; State of New Jersey Department of Education, 2019). More frequently, students who need language support are provided with word-to-word or translation dictionaries, which give standard language counterparts for specific terms but not definition, use, or explanation (Florida Department of Education, n.d.; Ohio Department of Education, 2018, 2019; South Carolina Department of Education, n.d.; State of New Jersey Department of Education, 2019). Chapter 6, Comparability When Assessing English Learner Students, provides more details on accommodations designed to help students from non-English and/or bilingual backgrounds.

Growing language and sociocultural diversity is not unique to the United States. Linguistic and sociocultural diversity exists to varying degrees throughout the world, and the recognition and treatment of such diversity also vary. South Africa, which has 11 official languages, administers assessments in these languages to students who come from backgrounds that include dozens of other languages spoken in the community. In Canada, which has two official languages, English and French, assessments are given in the two official languages, and students are assessed in the language of instruction (Ercikan, Oliveri, & Sandilands, 2013).

Placing value on language and sociocultural diversity necessitates policies to ensure that, regardless of background, students have the same opportunities to demonstrate their knowledge, skills, and competencies on assessments. This may necessitate exempting students whose English language proficiency has not advanced enough to allow them to demonstrate their knowledge, skills, and competencies using assessments conducted entirely in English. Only when bilingual students have developed the required level of language proficiency should they be tested in English and be provided the tools and accommodations to support their performance on assessments. Another possibility is to provide education and assessment in the students' home/community language until they have developed enough English proficiency to fully participate in and benefit from an education in English. In such cases, students will be administered adapted versions of assessments in their home language.

MEASUREMENT EQUIVALENCE

Developing assessments that capture the intended set of constructs for a language and sociocultural group requires extensive research to provide insights on how the construct is operationalized and developed in different contexts and empirical evidence that the assessment captures the intended constructs. The challenges are multiplied in multilingual and multicultural assessment contexts when the targeted constructs differ across and within cultural groups and social contexts, or the assessments are adapted to different languages. The appropriate interpretation of scores and comparisons for students from different sociocultural contexts and language backgrounds requires establishing empirical evidence of measurement equivalence for the considered groups. Measurement equivalence includes (1) *construct equivalence*, (2) *test equivalence*, and (3) *equivalence of testing conditions* (Ercikan & Lyons-Thomas, 2013; also see Chapter 5, Comparability Across Different Assessment Systems).

Construct equivalence is defined as the equivalence of meaning of the construct in terms of its theoretical definition, the way it is operationalized, and the way it is developed for the cultures in which the assessment will be administered. In addition to demonstrating similar psychometric properties, the evidence for whether a construct (e.g., reading proficiency) is conceptualized the same way across language and sociocultural groups also needs to be grounded in empirical research based on the considered language and sociocultural groups.

Test equivalence refers to the equivalence of test content, and the language and sociocultural equivalence at the item and the overall test levels. This includes the equivalence of text, graphics, formatting, language meaning, language demands, and cues for responding to test questions, and the sociocultural relevance in the words and contexts provided in the items.

Equivalence of testing conditions refers to the equivalence of test administration conditions such as the communication between test administrator and examinees, which includes test directions, instructions, and training sessions for the test administrators. It includes whether the different language versions of tests were administered in an identical fashion, whether the test format was equally appropriate in each language version, whether the speed of response was not more of a factor in one language than the other, and whether other response styles such as acquiescence, tendency to guess, and social desirability did not vary significantly across groups (Hambleton, 2005; Hambleton & Patsula, 1999). In addition, in a different language or sociocultural setting, test administrators should be drawn from the local language communities; be familiar with the culture, language, and local dialects; have adequate test administration skills; and know the importance of following standardized procedures associated with the assessment. Broader testing conditions should also be considered in interpreting scores of students from diverse cultural and language backgrounds. These conditions may include societal context for testing (such as the emphasis given to testing), which may affect how students perceive the testing situation and the role of testing as well as students' motivation to perform and how they engage with the assessment.

In addition to measurement equivalence, score comparability requires measurement unit or scalar equivalence, which refers to whether units on the score scales based on different assessment versions have equivalent units. Even when measurement equivalence requirements are met, in order to compare performance levels of students from different language and sociocultural backgrounds, or test forms taken in different languages, scalar equivalence is required. For example, a score difference of 10 score points on one scale based on the source-language version of the assessment can only be considered equivalent to 10 score points on a scale based on the non–English language version when the scores are on the same scale and using the same measurement units.

While measurement equivalence of items across subgroups is ideal, true equivalence is difficult to achieve in practice. In reality, variations across assessment versions or across language and sociocultural groups are inevitable. For instance, items may exhibit varying degrees of differential item functioning (DIF). The classification scheme developed at the Educational Testing Service assumes functional equivalence if the items exhibiting DIF do not exceed statistical significance threshold values (Dorans & Holland, 1992). Similarly, empirical evaluations of equivalence require determining the levels of difference for different language and sociocultural groups and establishing what levels of differences may be tolerated without compromising the comparability of scores for these groups.

Degrees of Measurement Incomparability

The growing interest in international assessments and comparisons, and a recognition of the complexity of score comparability across language and culture groups, have led to extensive research on the comparability of assessments and scores across languages, the procedures that optimize the performance of students from different language backgrounds, and the development of guidelines for test adaptation. Research on item-level comparability focuses on "unexpected" performance differences for examinee subgroups that are matched in terms of their overall ability or performance on the test. Regardless of group membership (e.g., based on gender, ethnicity, socioeconomic, or language backgrounds), students with the same ability level should have the same likelihood of receiving the same test score. The great majority of this research has focused on using DIF methods to examine item equivalence (e.g., Dorans & Kulick, 1986; Ercikan, 1998; Ercikan & Lyons-Thomas, 2013; Gierl, Rogers, & Klinger, 1999; Oliveri, Olson, Ercikan, & Zumbo, 2012; Padilla, Benítez, & Castillo, 2013; Sireci & Allalouf, 2003) and measurement equivalence based on test data factor structure (e.g., Ercikan & Koh, 2005; Güzel & Berberoglu, 2005; Sireci, Patsula, & Hambleton, 2005; Zumbo, 2003). More comprehensive discussions of DIF detection methods can be found in Holland and Wainer (2012) and Chapter 3, Comparability of Aggregated Group Scores on the "Same Test."

Research using DIF methodology has demonstrated extensive incomparability between language versions of assessments within and across countries. Studies conducted in Canada comparing the French and English versions of large-scale assessments found that 18 to 60 percent of items functioned differently for the two language groups (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004), pointing to high levels of incomparability at the item level. Research on international assessments identified high levels of incomparability between different language versions of items and tests (Byrne & van de Vijver, 2010; Ercikan, 1998; Ercikan & Koh, 2005; Ercikan & Lyons-Thomas, 2013; Ercikan & McCreith, 2002; Gierl, 2000; Gierl et al., 1999; Grisay, 2003; Hambleton et al., 2005; Marotta, Tramonte, & Willms, 2015; Oliveri & Ercikan, 2011; Solano-Flores, Backhoff,

& Contreras-Niño, 2009). For example, Ercikan and Koh (2005) compared English and French versions of TIMSS administered in the United States and France and identified that 79 percent of the science items functioned differentially between the two countries and language groups. These findings highlight the importance of the quality of the adaptation process on the validity of measurement and the comparability of scores, and the importance of establishing measurement comparability of assessments across languages.

Some research evidence demonstrates that measurement incomparability identified at item levels does not necessarily lead to observable scale-level differences. Previous studies using various methods such as exploratory factor analyses (Arim & Ercikan, 2005), confirmatory factor analyses (Ercikan & Koh, 2005; Oliveri et al., 2012; Zumbo, 2003), and test characteristic curve (TCC) comparisons (Ercikan & Gonzalez, 2008) identified little to no differential test functioning (DTF) despite large proportions of differences identified at the item level. For example, Ercikan and Gonzalez (2008), using item response theory–based TCCs, found small score-scale differences between different language versions of PIRLS assessments despite the presence of large percentages of DIF items. Also, as shown by a study conducted by Zumbo (2003) using confirmatory factor analysis, a similarly negligible DTF was seen even when the test contained large amounts of high-level DIF against a group. The inconsistency between item- and scale-level differences in measurement equivalences point to the importance of examining measurement comparability both at the item level and the scale level. Furthermore, one of the purposes of DIF analysis in multilingual assessments is to help identify possible differences created by the test adaptation process. Identifying such possible adaptation problems before assessments are finalized can provide opportunities for correcting adaptation problems and enhance comparability.

Sources of Measurement Incomparability

Even though one purpose of DIF analyses is to identify adaptation problems, the sources of statistical differences identified by DIF are often difficult to pinpoint (Ercikan, 2002; Haberman & Dorans, 2011; Hambleton et al., 2005). In particular, DIF found in items between language versions of an assessment does not always indicate problems in test adaptation. Sources of DIF are often explored using bilingual expert reviews and think-aloud protocols with students from non–English language groups (Ercikan et al., 2004). For example, Ercikan and McCreith (2002) examined sources of DIF using bilingual experts for comparing items in the English and French versions of TIMSS and demonstrated that in some booklets as few as 36 percent of the cases evidencing DIF were due to adaptation problems. Even when differences are identified by bilingual experts, these sources of DIF must be treated as hypotheses, as other research has shown that adaptation differences do not necessarily lead to differences in how students read, interpret, and solve test items in different languages (Ercikan et al., 2004).

Researchers have recognized the complexity of identifying potential sources of score incomparability in assessing students from diverse backgrounds (Ercikan, 2002; Hambleton et al., 2005). Psychometric differences between language versions can be due to multiple factors, including sociocultural and curricular differences between groups (Ercikan et al., 2004; Solano-Flores & Nelson-Barber, 2001; Solano-Flores, Trumbull,

& Nelson-Barber, 2002), educational policies and standards, values, and motivation to take the assessment (Arffman, 2010; Gee, 2013; Greenfield, 1997; John-Steiner & Mahn, 1996). Cultural differences can influence intrinsic interest in, familiarity with, and the interpretation of item content. In addition, word meaning is created through social and cultural interactions (Campbell, 2003; Derrida, 1998; Greenfield, 1997), and social context and cultural experiences are expected to affect interpretations of words, which in turn may affect the trajectory of thought processes and ultimately responses to assessment questions (Ercikan & Lyons-Thomas, 2013; Ercikan & Roth, 2006; Roth, 2009, 2010; Solano-Flores, 2006). Even within the same language context, such as in the United States or Canada, students from some sociocultural groups speak structurally and semantically different varieties of English. Examples of these include some but not all indigenous students, African-American students, Mexican-American students, and students from nonmainstream-SES backgrounds. For some but not all of these students, written standard English is an added difficulty (Roth & Harama, 2000), thereby creating linguistic incomparability for students from different sociocultural backgrounds.

Limitations of DIF Methodology

Research results increasingly point to the limitations of methodologies used in examining measurement comparability when the diverse sociocultural context is not taken into account. This research indicates that neglecting the within-group heterogeneity for DIF methodology has validity implications (Grover & Ercikan, 2017). Using simulated data, Oliveri, Ercikan, and Zumbo (2014) varied the heterogeneity within the focal groups from 0 to 80 percent and found that, as heterogeneity increased, the rates of correct DIF detection decreased. Ercikan and colleagues (2014) conducted DIF analyses on heterogeneous linguistic groups created using the information on the language of instruction, dominant language setting, and home/community language. They demonstrated that English and French language learners are heterogeneous groups, and that the DIF results do not necessarily apply to all members of a given language group in the same way, suggesting that issues other than test adaptations also play a role.

Researchers have attempted to account for heterogeneity by crossing two manifest groups (e.g., gender and ethnicity) to create more specific groups for DIF analysis. DIF analyses conducted at the subgroup level allow the researcher to determine whether and to what extent DIF items detected at the population level are the same as the DIF items detected at various subpopulation levels. This has been referred to as a "melting pot" DIF (Dorans & Holland, 1992) or DIF dissection approach (Zhang, Dorans, & Matthew-López, 2005). Ercikan and Oliveri (2013) also proposed a two-step approach in conducting DIF on heterogeneous samples in which latent class analysis is conducted within groups of the manifest variable of interest (i.e., males and females), as opposed to on the whole sample, which is typically done.

GUIDELINES AND CONSIDERATIONS FOR MULTIPLE LANGUAGE VERSIONS OF ASSESSMENTS

Two sets of guidelines dedicate significant attention to test adaptation and evaluation of the quality of test adaptations. These are (1) the standards developed by the American Educational Research Association (AERA), the American Psychological

Association (APA), and the National Council on Measurement in Education (NCME), and (2) the International Test Commission (ITC) guidelines. Relevant sections of these guidelines for multilingual and multicultural versions of assessments are summarized below.

AERA, APA, and NCME Standards

In the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014), four standards are particularly relevant to test adaptation:

- **Standard 9.4** highlights the need for assessment developers to explain and provide justification for linguistic modifications that they deem to be appropriate in specific situations. These modifications should be taken into account in score interpretations.
- **Standard 9.5** recommends that if there is evidence that scores are not comparable across multiple versions of assessments, additional information should be provided to help users interpret assessment scores correctly.
- **Standard 9.7** calls attention to the need to describe the approaches used in establishing the adequacy of adaptation, and for empirical and logical evidence to be provided for score reliability and the validity of the inferences based on the target assessment for all linguistic groups. For example, if an assessment adapted into Spanish is meant to be used with Mexican, Cuban, Spanish, and other Spanish-speaking subgroups, it is the responsibility of the assessment developer to provide independent reliability and validity evidence for each of those subgroups.
- **Standard 9.9** recommends that assessment developers provide evidence of the comparability of different language versions of an assessment. For example, the assessment developer should present evidence that the same construct is being measured in both assessments.

The Standards also advise against using back-translation, which involves comparisons of the source version with the target-to-source translation, as the sole method for verifying linguistic comparability. The comparability of the source- and back-translated target versions is not sufficient to establish that the two language versions have the same meaning and provide similar information to students. The use of interpreters or translators who are not familiar with proper testing procedures or purposes of testing may make inadequate translation and adaptation of the assessment and provide inappropriate test administration.

ITC Test Adaptation Guidelines

The second set of guidelines was developed by the ITC (Hambleton, 2005; ITC, 2018). The ITC emphasizes these guidelines as being instrumental in conducting and evaluating the adaptation or the simultaneous development of assessments for use with different populations. The 18 guidelines, along with suggestions for practice, were organized around six broad topics: pre-condition, test development, confirmation/empirical analyses, administration, score scales and interpretation, and documentation.

The guidelines in the section titled “Pre-Condition” highlight the decisions that are made before the translation/adaptation process begins. The second section, “Test Development,” focuses on the process of adapting an assessment that includes items, test instructions, and scoring rubrics. The third section, “Confirmation,” contains guidelines on documenting empirical evidence addressing score equivalence, as well as reliability and validity of the assessment in multiple languages and cultures. The section “Administration” pertains to the preparation of materials and instructions to minimize language and sociocultural issues in the administration process. The fifth section, “Score Scales and Interpretation,” discusses score comparisons. The final section, “Documentation,” contains guidelines detailing the technical aspects of test adaptations and the appropriate use of the test scores.

Assessment Development and Adaptation Processes

The quality of adaptation is optimized when assessments in the source language are developed with the test adaptation goal in mind. Brislin, Lonner, and Thorndike (1973) suggested using short, simple sentences of fewer than 16 words, employing the active rather than the passive voice, repeating nouns instead of using pronouns, and using specific rather than general terms (e.g., “cows, chickens, and pigs” rather than “livestock”). For an assessment to be adaptable, the source assessment should also avoid the language structures that are unlikely to have equivalents in other languages, such as metaphors or colloquialisms, certain modals (e.g., verb forms with “could” or “would”), adverbs and prepositions telling “where” or “when” (e.g., frequent, beyond, or upper), possessive forms (e.g., mine or Tim’s), probabilistic words (e.g., “probably” and “frequently”), and sentences with two different verbs if the verbs suggest different actions.

The most commonly used method for creating adapted versions of tests is *successive* test adaptation, where the assessment is developed in a source language and one or more bilingual translators adapt the assessment to the target language and culture. *Simultaneous* and *concurrent/parallel* assessment development are alternatives to the traditional approach of translating assessments created in a single source language, typically English in the North American context. In *simultaneous* assessment development, the emphasis is on the use of a multidisciplinary committee of experts in the languages, psychometrics, and the content domain for developing items (Tanzer & Sim, 1999). Test items are developed by *bilingual item writers* in one language and are immediately adapted into the other language. The *concurrent/parallel* assessment development model (Solano-Flores et al., 2002) utilizes *shells* or *templates*, which define item structure and the cognitive demands of each item. Using these templates for item development, the language groups work jointly in all stages of the assessment development process. The different language versions are developed by experts from each language group based on a common assessment blueprint. Using this approach, each assessment originates in the language it is targeted for and is developed by content experts from the particular sociocultural group, except for a small portion of the assessment that is adapted from the source language for linking purposes. Different development processes may have trade-offs between *comparability* and *cultural authenticity* of adapted assessments. While concurrent/parallel development prioritizes cultural authenticity, successive development prioritizes comparability, and simultaneous assessment developments target a compromise between comparability and cultural authenticity (Ercikan & Lyons-Thomas, 2013).

Creating Comparable Scores

There has been significant research on creating comparable score scales for multilingual versions of assessments (Cook, 2006; Cook & Schmitt-Cascallar, 2005; Sireci, 1997, 2005). This research indicates that, in the absence of sufficient evidence for measurement equivalence across groups, score scales should be based on separate language/country calibrations and comparability should be established through a linking procedure. Cook and Schmitt-Cascallar (2005) provide an overview of score comparability and describe four methods of linking scores on assessments given in different languages.

Currently, some of the most well-known assessments with multiple language versions include international assessments such as the TIMSS, the PISA, and the PIRLS. In these assessments, a single international score scale allows for comparisons of performances across countries and language groups. The single score scale is computed with item parameters calibrated on an international sample that consists of a randomly selected subsample from countries that take the assessments in different languages. In some assessments, country-specific parameters are used when the international item parameters do not fit the scale well because of some level of measurement incomparability.

In establishing comparability of scores across language versions of assessments, consideration of comparability must start from the conceptualization of the assessment. Ercikan and Lyons-Thomas (2013) identify seven key steps in developing and adapting assessments for use in different languages and cultures:

1. **Examine the equivalence of the constructs.** This step requires examining and comparing the construct definitions in the source and other cultures and languages. It may involve a review by cultural and language expert groups who can evaluate the appropriateness of the construct definitions and identify aspects of the construct that may be different for the two language and cultural groups.
2. **Select a test adaptation and development method.** The next step involves deciding on the approach to developing multiple versions of assessments. If a source version already exists, successive test adaptation may need to be used. If, however, there is an opportunity to build multiple language versions from the beginning, parallel or simultaneous development may be employed.
3. **Perform the adaptation of the test or measure.** There are several factors that will affect the quality of the adaptation. First is the linguistic features of the source version that might affect translatability. These include using short sentences, repeating nouns instead of pronouns, and avoiding metaphors and a passive voice in developing assessments. Other factors include language background and proficiencies of translators in the relevant languages, such as whether translators are fluent in both languages and knowledgeable about both source and target cultures, and they have clear understanding of the construct being assessed.

4. **Evaluate the language equivalence between the two assessment versions.** A necessary step in developing multiple language versions of assessments is an evaluation of equivalence by bilingual experts. Reviews of equivalence can determine differences in language, content, format, and other aspects of items in the comparison languages. Insights from such reviews can help inform revisions of adaptations to optimize comparability.
5. **Document changes made in the adaptation process.** Documenting changes and the rationale for these changes between the language versions of assessments is critical for informing test users for potential impact on comparability.
6. **Conduct a field test study to examine measurement equivalence.** Establishing measurement equivalence requires empirical evidence to support such an evaluation. Field test data can be used to examine the reliability and validity of both language versions of assessments, as well as measurement equivalence using classical test theory-based analyses, factor analyses, DIF analyses, and comparisons of TCC curves. Additional follow-up studies can include a second round of expert reviews and cognitive analyses to provide further support for comparability of the language versions of assessments.
7. **Conduct linking studies.** In order to create comparable scales with measurement unit equivalence, a linking study is needed once measurement equivalence has been established.

NEXT-GENERATION ASSESSMENTS

As assessments increasingly include multiple modes of administration—using both paper and digital delivery, and sometimes delivered on multiple types of digital devices within the same test administration—as well as increasingly complex forms of interactivity, assessment developers need to consider ways student backgrounds may affect how students engage with assessments. The technology-enhanced environments provide growing opportunities for interactivity between the student and the assessments, and usually involve multiple modes of engaging with the assessment. To understand the questions, students may be required to read excerpts, listen to audio segments, view video clips, or manipulate diagrams with their mouse. They then respond by typing text, speaking, drawing diagrams, plotting graphs, or dragging items. While the first generation of computer-based assessments typically mimics its paper-based counterparts in using text-based items, advancing technology has made new item formats possible. An example is the integration of videos in assessment items, which allows for the assessment stimulus to be delivered through an acted-out scenario. Videos (or animations) are useful when a detailed description can be too lengthy and the video can more effectively demonstrate and elaborate on what is being described. In assessments of language proficiency, students can be asked to verbally describe the interactions they see in a video, which may approximate how language is used in day-to-day life. Mechanical reasoning can also be tested when students are presented with videos of machine components and given verbal or written questions on how the machine works.

Some performance-based assessments require students to work with others. Such tasks often include multimodal presentation of the stimuli and responses are not restricted to writing. While such performance-based assessments, including assessments modeled after games, can create an environment where students acquire and demonstrate skills not typically captured in traditional assessments (e.g., communication and collaborative problem solving), the multimodal aspects of tasks and responses create challenges for measurement equivalence. Some newer assessment types also require students to interact with on-screen actors or avatars. Assessments with interactive elements may introduce behaviors that are not observed in traditional assessments (Zapata-Rivera & Bauer, 2012), such as the tendency to explore features of the assessment platform, or to engage in behaviors that push its boundaries (e.g., deliberately providing irrelevant responses), in addition to variations of sociocultural differences already observed in traditional assessments that may result in different response times, constructs being assessed, and measurement properties of the assessments.

Game-like elements (e.g., use of avatars in animations and actors in videos) that have been introduced into assessments to engage students present a different set of challenges. For example, a text item can refer to a "fellow student" without mentioning race, hair color, gender, age, or dressing style, whereas the use of avatars and actors will inevitably reflect these characteristics (Popp, Tuzinski, & Fetzer, 2016). Lee and Park (2011) found that minorities reported a lower sense of belonging and less desire to participate in a game with more white avatars than when the ethnic diversity of the avatars was more balanced. Such physical attributes can introduce social identity threats and affect student motivation (Baylor, 2011) and score comparability. The representation of diversity is often necessary and desired in state assessments with a diverse population and can be achieved in videos and animations.

Score comparability issues can also arise when sociocultural groups interpret pictures and videos differently based on their experiences. The extensive use of images in some new item types, such as hot-spot items (i.e., identification of correct or incorrect zones) and drag-and-drop image matching, requires that the images and videos represent the same notion to students from different sociocultural backgrounds (Solano-Flores & Nelson-Barber, 2001).

Despite the challenges, next-generation assessments can confer comparability advantages that are lacking in traditional assessments. Videos offer advantages in cases where some language and sociocultural groups might otherwise require test accommodations. In some cases, administering items through a video (or through audio) can be akin to the *human read-aloud* accommodation provided for some state assessments such as the LEAP 2025 Assessment, offered by the Louisiana Department of Education (Data Recognition Corporation, 2016), where items are read verbatim to individual students needing test accommodations. The readers may not clarify, provide additional information, assist, or influence the student's response in any way. For mathematics read aloud, readers may read the title, provide a general overview of the image (e.g., graphs, equations), and describe the details in a succinct manner. The process requires the recruitment and the training of many readers. With video or audio test items, variation due to readers is eliminated as all examinees experience the same stimulus.

The current guidelines for test adaptations have yet to consider the possibilities and limitations of the new assessment types, partly due to the dearth of research studies. The decision to use a new format should be aligned to the assessment's goals and fit with the intended construct and the ease of test adaptations. Also, in representing diversity, the designs of avatars should not default to stereotypes. A diverse group of reviewers should evaluate the appropriateness of the diversity representation of avatars and actors. Finally, students' familiarity with technology should be considered (Ercikan, Asil, & Grover, 2018). Students who regularly use video chatting platforms or are accustomed to receiving news or information online may have an edge over students who have little or no access to computers outside the school environment.

CONCLUSION

In this chapter, our goal has been to highlight the importance of following carefully designed procedures for designing and developing assessments for multilingual and multicultural contexts, establishing comparability of assessments and scores across language groups, and taking multiple societal factors into account in using and interpreting scores.

Comparability of scores for students from different language and sociocultural backgrounds is central to the validity of inferences from assessments. Validity is compromised when scores from multiple language versions of assessments are compared, implicitly or explicitly, without establishing comparability. For example, aggregating scores from different language versions of assessments, such as English and Spanish, at the class, school, or higher level makes an implicit assumption of score comparability of these language versions of assessments and the resulting scores. The incorrect assumption of score comparability compromises the validity of inferences when comparing students' performances at the class, school, or higher level. More explicit comparisons, such as comparing performance levels of students who took the assessment in English versus Spanish, may disguise or exaggerate performance differences when comparability has not been established.

Research on comparability issues in multiple language versions of assessments has been evolving to consider different assessment contexts. As assessments increasingly include multiple modes of administration (e.g., paper, digital, and multiple devices) and interactivity, assessment developers must also consider the expanded ways student backgrounds may contribute to how differently students engage with assessments and the resulting comparability issues across languages. It is important to highlight that the comparability of assessments and scores across languages is expected to be sensitive to population, cultural, and societal contexts. The validity and comparability evidence need to be updated periodically given potential changes in the society, education systems, and sociocultural context of assessments over the years.

REFERENCES

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

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research, 54*(1), 37–59.

Arim, R., & Ercikan, K. (2005, April). *Comparability between the US and Turkish versions of the Third International Mathematics and Science Study's (TIMSS) mathematics test results*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Bachman, H. J., Votruba-Drzal, E., El Nokali, N. E., & Castle Heatly, M. (2015). Opportunities for learning math in elementary school: Implications for SES disparities in procedural and conceptual math skills. *American Educational Research Journal, 52*(5), 894–923.

Baylor, A. L. (2011). The design of motivational agents and avatars. *Educational Technology Research and Development, 59*(2), 291–300.

Berliner, D. C. (2012). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord. org/Content.asp?ContentID=16889.

Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education, 28*(1), 44–55.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods*. New York: Wiley.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107–132.

Campbell, L. (2003). How to show languages are related: Methods for distant genetic relationship. In B. D. Joseph & R. D. Janda (Eds.), *The handbook of historical linguistics* (pp. 262–282). Malden, MA: Blackwell Publishing.

Carnevale, A. P., & Rose, S. J. (2003). *Socioeconomic status, race/ethnicity, and selective college admissions.* Century Foundation. Retrieved July 18, 2019, from https://files.eric.ed.gov/fulltext/ED482419.pdf.

Cook, L. L. (2006). *Practical considerations in linking scores on adapted tests*. Keynote address at the 5th International Meeting of the International Test Commission, Brussels, Belgium.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139–170). Mahwah, NJ: Lawrence Erlbaum Associates.

Data Recognition Corporation. (2016). *LEAP 2025 accommodations and accessibility features user guide.* Retrieved May 30, 2019, from https://www.louisianabelieves.com/docs/default-source/assessment/leap-accessibility-and-accommodations-manual.pdf?sfvrsn=12.

Derrida, J. (1998). *Monolingualism of the other, or, the prosthesis of origin*. Palo Alto, CA: Stanford University Press.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series, 1992*(1), i–40.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355–368.

Elder, T. E., Figlio, D. N., Imberman, S. A., & Persico, C. I. (2019). *School segregation and racial gaps in special education identification* (No. w25829). National Bureau of Economic Research.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543–553.

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

Ercikan, K. (2003). Are the English and French versions of the Third International Mathematics and Science Study administered in Canada comparable? Effects of adaptations. *International Journal of Educational Policy, Research and Practice, 4*, 55–76.

Ercikan, K., Asil, M., & Grover, R. (2018). Digital divide: A critical context for digitally based assessments. *Education Policy Analysis Archives, 26*(51), 1–24. Retrieved from https://epaa.asu.edu/ojs/article/view/3817.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301–321.

Ercikan, K., & Gonzalez, E. (2008, March). *Score scale comparability in international assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*, 23–35.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 545–569). Washington, DC: American Psychological Association.

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In *Secondary analysis of the TIMSS data* (pp. 391–405). Dordrecht, the Netherlands: Springer.

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity, fairness and testing of individuals in high stakes decision-making context* (pp. 69–86). Bingley, UK: Emerald.

Ercikan, K., Oliveri, M. E, & Sandilands, D. (2013). Large scale assessments of achievement in Canada. In J. A. C. Hattie and E. M. Anderman (Eds.), *The international handbook of student achievement* (pp. 456–459). New York: Routledge.

Ercikan, K., & Roth, W-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*, 14–23.

Ercikan, K., Roth, W-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record, 117*(1), 1–28.

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.

Florida Department of Education (n.d.). *Florida NAEP 2015 English language learners (ELL) inclusion guidelines*. Retrieved May 30, 2019, from http://www.fldoe.org/core/fileparse.php/5423/urlt/ELLAGPPA.doc.

Gahungu, A., Gahungu, O., & Luseno, F. (2011). Educating culturally displaced students with truncated formal education (CDS-TFE): The case of refugee students and challenges for administrators, teachers, and counselors, *International Journal of Educational Leadership Preparation, 6*(2), 1–19. Retrieved from https://eric.ed.gov/?id=EJ973832.

Gee, J. P. (2013). Reading as situated language: A sociocognitive perspective. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 136–151). Newark, DE: International Reading Association.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280–296.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

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

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178–195.

Güzel, Ç. I., & Berberoglu, G. (2005). An analysis of the programme for international student assessment 2000 (PISA 2000) mathematical literacy data for Brazilian, Japanese and Norwegian students. *Studies in Educational Evaluation, 31*(4), 283–314.

Haberman, S. J., & Dorans, N. J. (2011). Sources of score scale inconsistency. *ETS Research Report Series, 2011*(1), i–9.

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* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1–30.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. New York: Routledge.

ITC (International Test Commission). (2018). ITC guidelines for translating and adapting tests (2nd ed.). *International Journal of Testing, 18*, 101–134.

John-Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational Psychologist, 31*(3–4), 191–206.

Kirshner, B., Gaertner, M., & Pozzoboni, K. (2010). Tracing transitions: The effect of high school closure on displaced students. *Educational Evaluation and Policy Analysis, 32*(3), 407–429.

Lee, J. E. R., & Park, S. G. (2011). "Whose second life is this?" How avatar-based racial cues shape ethno-racial minorities' perception of virtual worlds. *Cyberpsychology, Behavior, and Social Networking, 14*(11), 637–642.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Lipka, J., & McCarty, T. L. (1994). Changing the culture of schooling: Navajo and Yup'ik cases. *Anthropology & Education Quarterly, 25*(3), 266–284.

López, A. A., Turkan, S., & Guzmán-Orth, D. (2017). Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students. *ETS Research Report Series, 2017*(1), 1–12.

Marotta, L., Tramonte, L., & Willms, J. D. (2015). Equivalence of testing instruments in Canada: Studying item bias in a cross-cultural assessment for preschoolers. *Canadian Journal of Education, 38*(3).

McEniry, M., Samper-Ternent, R., Flórez, C. E., Pardo, R., & Cano-Gutierrez, C. (2019). Patterns of SES health disparities among older adults in three upper middle- and two high-income countries. *The Journals of Gerontology: Series B, 74*(6), 25–37.

McQueen, J., & Mendelovits, J. (2003). PISA reading: Cultural equivalence in a cross-cultural study. *Language Testing, 20*(2), 208–224.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

New York State Education Department. (2016). *Testing accommodations for students with disabilities and English language learners*. Retrieved May 30, 2019, from http://www.p12.nysed.gov/assessment/accommodations/testingaccomell-16.pdf.

Ohio Department of Education. (2018). *Accessibility features for students taking the paper-based Ohio's state tests.* Retrieved May 30, 2019, from http://education.ohio.gov/getattachment/Topics/Testing/Accommodations-on-State-Assessments/OHAccessManual.pdf.aspx?lang=en-US.

Ohio Department of Education. (2019). Revised assessment accommodations for English learners. Retrieved May 30, 2019, from http://education.ohio.gov/getattachment/Topics/Other-Resources/English-Learners/Revised-Assessment-Accommodations-for-English-Lear/Announcement-EL-Accommodations-and-OGT-Retakes.pdf.aspx?lang=en-US.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education, 24*(4), 349–366.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education, 27*(4), 286–300.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing, 12*(3), 203–223.

Oregon Department of Education. (2019). *2019-20 Oregon accessibility manual*. Retrieved September 4, 2019, from https://www.oregon.gov/ode/educator-resources/assessment/Documents/accessibility_manual.pdf.

Padilla, J.-L., Benítez, I., & Castillo, M. (2013). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Methodology, 9*, 113–122.

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137–1162.

Popp, E. C., Tuzinski, K., & Fetzer, M. (2016). Actor or avatar? Considerations for selecting appropriate formats for assessment content. *Technology and Testing: Improving Educational and Psychological Measurement*, 79–103.

Quealy, K., & Shapiro, E. (2019, June 17). Some students get extra time for New York's elite high school entrance exam. 42% are white. *The New York Times.* Retrieved from https://www.nytimes.com/interactive/2019/06/17/upshot/nyc-schools-shsat-504.html.

Roth, W. M. (2009). Phenomenological and dialectical perspectives on the relation between the general and the particular. In K. Ercikan & W. M. Roth (Eds.), *Generalization in educational research* (pp. 235–260). New York: Routledge.

Roth, W. M. (2010). *Language, learning, context: Talking the talk*. London, UK: Routledge.

Roth, W. M., & Harama, H. (2000). (Standard) English as second language: Tribulations of self. *Journal of Curriculum Studies, 32*(6), 757–775.

Shifrer, D., & Fish, R. (2019). A multilevel investigation into contextual reliability in the designation of cognitive health conditions among US children. *Society and Mental Health.* https://doi.org/10.1177/2156869319847243.

Silva, S. M., Verhoeven, L., & van Leeuwe, J. (2011). Socio-cultural variation in reading comprehension development among fifth graders in Peru. *Reading and Writing, 24*, 951–969. https://doi.org/10.1007/s11145-010-9242-2.

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

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, NJ: Lawrence Erlbaum Associates.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417–453. https://doi.org/10.3102/00346543075003417.

Smarter Balanced Assessment Consortium. (n.d.). *Embedded designated support—glossaries*. Retrieved August 9, 2019, from https://portal.smarterbalanced.org/library/en/instructions-for-using-embedded-glossaries.pdf.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record, 108*(11), 2354.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing, 9*(2), 78–91.

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., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107–129.

South Carolina Department of Education (n.d.). *SC READY online testing tools and supports*. Retrieved May 30, 2019, from https://ed.sc.gov/scdoe/assets/File/districts-schools/special-ed-services/SC%20Ready%20Accommodations%20Charts_12-31-15.pdf.

State of New Jersey Department of Education. (2019). *Testing accommodations*. Retrieved May 30, 2019, from https://www.nj.gov/education/assessment/accommodations.

Tabaku, L., Carbuccia-Abbott, M., & Saavedra, E. (2018). *State assessments in languages other than English*. Retrieved August 8, 2019, from https://files.eric.ed.gov/fulltext/ED590178.pdf.

Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing, 13*(2), 89–104.

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations. *European Journal of Psychological Assessment, 15*, 258–269.

Tate, W. F. (1997). Race, ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652–680.

U.S. Census Bureau. (2017a). *2017 American Community Survey 1-year estimates.* Retrieved May 30, 2019, from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_B16001&prodType=table.

U.S. Census Bureau. (2017b). *Characteristics of people by language spoken at home.* Retrieved May 30, 2019, from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_S1603&prodType=table.

Ward, M. E., Shelley, K., Kaase, K., & Pane, J. F. (2008). Hurricane Katrina: A longitudinal study of the achievement and behavior of displaced students. *Journal of Education for Students Placed at Risk, 13*(2–3), 297–317.

Wu, A., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*, 287–300.

Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for twenty-first-century skills: Theoretical and practical implications from modern research* (pp. 147–169). Charlotte, NC: Information Age Publishing.

Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). Using DIF dissection method to assess effects of item deletion. *ETS Research Report Series, 2005*(2), i–11.

Zirkel, P. A., & Weathers, J. M. (2016). K–12 students eligible solely under section 504: Updated national incidence data. *Journal of Disability Policy Studies, 27*(2), 67–75.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136–147.