GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2025) 37(3) 1-11

Psicothema

https://www.psicothema.com/es • ISSN 0214-9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Article

Navigating Multi-Language Assessments: Best Practices for Test Development, Linking, and Evaluation

Louise Badham[1], María Elena Oliveri[2] and Stephan G. Sireci[3]

[1]International Baccalaureate (United Kingdom)
[2]Purdue University (USA)
[3]University of Massachusetts Amherst (USA)

ARTICLE INFO

Received: 15/10/2024
Accepted: 20/03/2025

Keywords:
Comparability
Cross-lingual assessment
Linking tests
Test translation
Validity

ABSTRACT

Background: Developing assessments in multiple languages is hugely complex, impacting every stage from test development to scoring, and evaluating scores. Different approaches are needed to examine comparability and enhance validity in cross-lingual assessments. **Method:** A review of literature and practices relating to different methods used in cross-lingual assessment is presented. **Results**: There has been a shift from source-to-target language translation to developing items in multiple languages simultaneously. Quantitative and qualitative methods are used to link and evaluate assessments across languages and provide validity evidence. **Conclusions:** This article provides practitioners with an overview and research-based recommendations relating to test development, linking, and validation of assessments produced in multiple languages.

Evaluaciones Multilingües: Mejores Prácticas para su Desarrollo, Vinculación y Evaluación

Palabras clave:
Comparabilidad
Evaluación interlingüística
Pruebas de vinculación
Traducción de exámenes
Validez

RESUMEN

Antecedentes: El desarrollo de evaluaciones en varios idiomas es enormemente complejo y afecta a todas las etapas, desde el desarrollo de las pruebas hasta la puntuación y la evaluación de las puntuaciones. Se necesitan diferentes enfoques para examinar la comparabilidad y mejorar la validez de las evaluaciones interlingües. **Método:** Se presenta una revisión de la literatura y las prácticas relacionadas con los métodos utilizados en diferentes áreas de la evaluación interlingüística. **Resultados:** Se ha pasado de la traducción del idioma de origen al idioma de destino al desarrollo simultáneo de items en varios idiomas. Se utilizan métodos cuantitativos y cualitativos para vincular y evaluar las evaluaciones en varios idiomas y proporcionar pruebas de validez. **Conclusiones:** Este artículo proporciona a los profesionales una visión general y recomendaciones de la literatura relacionada con el desarrollo de pruebas, la vinculación y la validación de evaluaciones producidas en varios idiomas.

Cite as: Badham, L., Oliveri, M. E., & Sireci, S. G. (2025). Navigating multi-language assessments: Best practices for test development, linking, and evaluation. *Psicothema, 37*(3), 1-11. https://doi.org/10.70478/psicothema.2025.37.19

Corresponding author: Louise Badham, International Baccalaureate, louise.badham@ibo.org

Introduction

Developing assessments in multiple languages is complex, impacting test development, scoring, and evaluating results. Ensuring fairness and validity across languages requires considering linguistic structures, sociolinguistic factors, and educational policies that shape assessment outcomes in diverse global settings. For instance, as the second most spoken language in the United States, Spanish versions of large-scale assessments are designed to accommodate emergent bilingual students. However, linguistic differences between English and Spanish, including verb conjugation complexity and sentence structure require careful adaptation to ensure construct equivalence. Additionally, regional dialectal differences among Spanish speakers from Spain, Mexico, the Caribbean, and Central and South America must be addressed to avoid cultural bias.

Multilingual assessment practices in other global contexts further highlight the need for tailored approaches. In Canada, where English and French are official languages, assessments must ensure validity across linguistic groups while accounting for language-specific conceptual distinctions. In sub-Saharan Africa, where indigenous languages coexist with colonial languages (English, French, and Portuguese), educational policies influence assessments in local languages versus global languages. Test developers must navigate complex decisions about language prioritization, given variations in literacy levels and educational access. Considerations are particularly complex in international large-scale assessments (ILSAs) spanning diverse linguistic and cultural contexts.

This article reviews cross-lingual assessment methods and practices, providing guidance in identifying and mitigating biases against linguistic or cultural groups. Bias in assessments can disadvantage certain groups, making it crucial to consider global and local linguistic variations in translations (van de Vijver & Poortinga, 2005). Such bias can have far-reaching social consequences, such as limiting educational outcomes or career progression. Therefore, assessment development and evaluations must be robust and rigorous to minimize bias and allow valid score-based inferences (Ercikan & Lyons-Thomas, 2013).

Our review covers: (a) multi-language test development; (b) "linking" different language versions of assessments; and (c) evaluating results of cross-lingual assessments. Our review extends previous work by Sireci et al. (2016)—which reviews some of the same sources—by connecting linking and evaluation methods with test development approaches and exploring new techniques from emerging technologies. Whilst the scope of the study was limited to cognitive skills in multi-language educational assessments, the issues addressed apply across multiple settings, such as personnel selection in multinational corporations.

Adapting Tests Across Languages

As highlighted in the *Standards for Educational and Psychological Testing*, "simply translating a test from one language to another does not ensure that the translation produces a version of the test that is comparable in content and difficulty" (AERA et al., 2014, p.60). Therefore, whilst *translation* is often used to describe the process of adjusting tests into other languages, the term *adaptation* is preferable (Hambleton, 2005; ITC, 2017). Adaptation reflects that the process accounts for cultural relevance, aiming to maximize validity in target language assessments (Ercikan & Por, 2020). *Transadaptation* is also used, but is redundant, having the same definition as adaptation. Thus, "adaptation" is used here, although "translation" is sometimes used interchangeably, or to describe part of the adaptation process.

Approaches to Developing Multi-Language Tests

Approaches include: (a) *adapting* tests from one (*source*) language to another (*target*) language(s); (b) *simultaneous development,* where multilingual teams create and adapt items together; and (c) *parallel development,* where each language version is developed separately.

(Successive) adaptation involves developing a monolingual, source-language test version, which is translated into one or more target languages (Rogers et al., 2003). The process may include "back-translation" (Brislin, 1970), where tests are translated from source to target language and back again, then source versions are compared to verify whether the original meaning has been retained. *Simultaneous development* is a form of adaptation, but rather than developing a source language first, multilingual committees develop and immediately adapt items across languages (Tanzer, 2005). In *parallel development,* content is developed independently in each language according to common specifications. Rather than using common items, the approach to defining and representing constructs on multi-language assessments is designed to be comparable. Some items may also be adapted to maximize construct comparability (Ercikan & Lyons-Thomas, 2013).

Linking and Comparing Tests Across Languages

Cross-lingual assessment literature discusses different levels of "equivalence" or score comparability. Examples include "structural," "metric," and "scalar" equivalence, which sit within the broader areas of *linking* or *equating* test scores (Sireci et al., 2016). In equating, scores from different test forms are adjusted and placed onto a common scale and can theoretically be considered interchangeable (Lord, 1980). Equating requires measurement of the same construct, and that tests are developed from common content specifications. Adapted tests typically involve the same construct and content specifications. However, translated content cannot be considered "common", so strict equating of translated tests is impossible (Dorans & Middleton, 2012; Sireci, 1997). "Weaker" forms of equating known as "linking" have fewer assumptions and are usually sought for multi-language assessments (Sireci, 1997; Sireci et al., 2016).

Sireci (1997) identified three cross-lingual linking designs: (a) *separate monolingual groups* where each group takes the language version it was developed for; (b) *matched monolingual groups* where examinees from different languages are matched on external criteria (e.g., socioeconomic status) rather than using anchor items; or (c) *bilingual groups* where bilingual examinees either take both language versions, or are randomly assigned one language. Related to linking, are methods that *evaluate comparability* across multi-language assessments (Sireci et al., 2016). Rather than seeking to establish a relationship between assessments, these approaches examine whether tests measure the same construct in the same way across language groups. The most common method is differential item functioning (DIF) (Zumbo, 2015), which examines how different groups with similar abilities respond to the same items.

DIF can, for example, be used to examine cross-language item comparability, or cross-country variations arising from cultural differences (Ercikan, 2002).

International Guidelines

The International Test Commission (ITC) provides internationally recognized guidelines on assessment practices, such as *Guidelines for Translating and Adapting Tests* (ITC, 2017). For emerging technology-based approaches, *Guidelines for Technology-Based Assessments* also aim "to ensure fair and valid assessment in a digital environment" (ITC & Association of Test Publishers, 2022, p.1). Informed by these guidelines, our review explored how cross-lingual assessment theory and methods have been applied in practice.

Method

Search Parameters

To identify relevant literature, we used the keywords: *"translation," "adaptation," "transadaptation," "cross-lingual,"* and *"dual-language,"* combined with *"test,"* and *"assessment."* Literature citing "ITC" was also included. Citation histories for key articles were reviewed to identify influential research and emerging themes. Grey literature from international assessment organizations (e.g., adaptation guidelines) was included to illustrate multi-language assessment practices. Key publications are provided in the References.

Search Process

Our multi-step search strategy combined database searches with manual reviews of high-impact journals. We searched four major academic databases: Educational Resource Information Center (ERIC), PsycInfo, Web of Science, and EBSCO to ensure broad coverage of peer-reviewed literature on test adaptation, translation, and cross-lingual validity. ERIC was particularly relevant for educational research, while PsycInfo captured studies on psychological and linguistic aspects of assessment. Web of Science and EBSCO broadened disciplinary scope, ensuring that emerging research trends in related fields were considered.

We compiled a vetted bibliography of studies that met our inclusion criteria, then conducted a secondary manual review of top-ranked journals in educational measurement, assessment, and cross-cultural psychology (e.g., *International Journal of Testing*). Journals were selected based on impact factor, citation influence, and reputations for publishing high-quality research in test adaptation and multilingual assessment. Combining systematic database searches with a targeted review of high-impact journals allowed us to capture broad trends and in-depth discussions from specialized sources. By integrating diverse methodological perspectives, our review reflected both empirical research and theoretical insights valuable for practitioners and researchers in multilingual assessments.

Screening and Filtering

An initial 618,761 records were filtered by publication type (journal), field of study (education and related disciplines), and language (English), to 51,744. Selections were further streamlined by reviewing citation counts to prioritize highly cited and influential studies, while recognizing that recent publications (2022–2024) may have lower citation counts. Publications that did not contribute new information were excluded due to saturation. Selected publications prioritized studies that contributed new theoretical, methodological, or empirical knowledge relevant to multilingual assessment adaptation. These were categorized according to key cross-lingual assessment areas (Table 1).

We read, analyzed, and synthesized selected publications to extract key themes, methodologies, and best practices related to multi-language assessments.

Table 1
Sources Reviewed

Methods used to…	# Publications
(a) create exams for use in multiple languages	16
(b) adapt exams from one language into other languages	17
(c) link different language versions of exams	23
(d) evaluate comparability of scores in multi-language exams	24

Results

Selected publications were stratified by their focus with respect to: (a) "Test Development," involving methods for creating or adapting multi-language assessments; (b) "Test Score Linking," comprising cross-lingual linking methods; and (c) "Evaluating Comparability" including measurement invariance studies at test and item-levels, and computational linguistic techniques.

Test Development

The test adaptation literature spans over 50 years and includes discussions of the pros and cons of different models for developing multi-language tests (e.g., Hambleton, 1994; van der Vijver & Tanzer, 1997). "Test development" encompasses *adapting* and *creating* multi-language tests, as both refer to processes of constructing assessments for use in different languages. Different test development models are presented here, with examples of ILSAs that illustrate them in practice.

Adaptation

Adaptation is the most common approach to developing multi-language assessments (Ercikan & Por, 2020), and is used for virtually all ILSAs (Ebbs & Korsňáková, 2016). Multi-language ILSAs using adaptation include the Programme for International Student Assessment (PISA) (OECD, 2016, 2024), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Ebbs et al., 2021; Ebbs & Korsňáková, 2016; Martin et al., 2020).

TIMSS and PIRLS are delivered in over 50 languages around the world (Ebbs et al., 2021, Martin et al., 2020), using a decentralized translation approach. National research or study centers adapt assessments into national language(s) following agreed procedures. Translator(s) and reviewer(s) must have experience of the cultural context and working with students in the target demographic, which helps mitigate risks of (dis)advantaging respondents from using direct translations. Korsňáková et al. (2020) illustrated this approach in an Arabic version of TIMSS for Middle East and North African countries, which accommodated cultural and regional

variations in language. Translators from different Arabic-speaking countries each produced initial translations, and a reviewer cross-checked translations to select the best version. Finally, an expert panel reviewed and refined the translation for the target audience. This approach helped ensure cross-linguistic, cross-national, and cross-cultural equivalence, without which one cannot achieve a quantitative cross-cultural comparison (Dept et al., 2017; ITC, 2017; Koršňáková et al., 2020).

PISA is also delivered globally in over 50 languages (OECD, 2024). Assessments are first developed in English and French source versions (Grisay et al., 2009; OECD, 2016, 2024), which are both translated into the target language, before being reconciled into a final, target-language version. This method helps identify linguistic discrepancies during adaptation. Some assessments are cross-checked against other verified language versions to increase cultural relevance (e.g., Catalan, Galician and Basque compared to Spanish versions (OECD, 2024)). Back-translation is also used (ibid), which is a useful quality control check (ITC, 2017). However, idiosyncratic features of source languages translated into target languages can go unnoticed (El Masri et al., 2016), or high-quality back-translations may mask issues from poorer quality initial translations (Koršňáková et al., 2020). Grisay (2003) demonstrated double-translation's advantages over back-translation in a PISA reading passage, where irony was lost in literal translations—an issue identified in double-translation reconciliation, but would likely have been missed in back-translation.

Challenges in Adaptation

Zhao et al. (2018) created a typology of language translation errors in PISA items to examine characteristics of specific source-target language combinations. In reviewing error types from English-to-Spanish translations, they found one required modification, 14 were eliminated, and 11 new error types were identified in English-to-Chinese translations. Different error types also occurred when translating science versus mathematics items, indicating different content areas create different translation challenges. Thus, different translation approaches may be needed for different content areas or language combinations.

When investigating cross-lingual comparability in PISA science items, El Masri et al. (2016) observed different word frequencies can make "common" words more challenging in some languages than others. They exemplified this with "crescent moon." "Crescent" is a high-frequency word in French (due to the famous pastry) and Arabic (being a symbol of Islam, the dominant religion in Arabic-speaking countries), but low-frequency—so more cognitively challenging—in English. Additionally, they observed biases can arise from inherent linguistic complexities and "untranslatable language idiosyncrasies" (p. 440), rather than any fault in adaptation processes. For example, "abbreviation incongruence" (p. 444) can occur where universal Latin script abbreviations (e.g., chemical elements) are used in non-Latin script assessments (e.g., Arabic), placing an additional cognitive burden on students in those languages. Similarly, Lu and Sireci (2007) identified "differential speededness" in translated assessments, where some languages require more words than others to express the same meaning. Consequently, examinees may require more time to take assessments in some languages than others.

Simultaneous Development

To mitigate issues arising from linguistic and cultural differences, checklists (e.g., Hambleton & Zenisky, 2011) can support quality assurance in adaptation. Additionally, linguistic and cultural considerations can be integrated directly into adaptation. *Simultaneous development* involves developers from different languages and cultures throughout test development, thereby ensuring "maximum linguistic and cultural decentering" in the process (Tanzer, 2005, p.238). Linguistic and cultural nuances, such as dialectical differences, can be identified during adaptation. For example, "*aula*" (classroom) is used in U.S. Spanish and Spain, but "*salón*" is preferred in Mexico. Such subtle differences could impact comprehension and familiarity, especially for students in specific educational settings. In addition to improving cultural relevance, Rogers et al. (2010) suggested integrating such information from different linguistic and cultural groups can make test development more efficient, by minimizing review time later in the process.

Parallel Development

Parallel development is a relatively rare approach. One example is International Baccalaureate (IB) exams, which are offered in up to 75 languages (IBO, 2024). Whilst many IB assessments (e.g., Sciences) are adapted across languages, others including Literature are created in parallel. Cross-lingual comparability is supported through common test specifications providing guidance on content, cognitive areas to be measured, quantity and format of items (IBO, 2018). Comparability is further enhanced through translation templates and cross-language "assessment editing meetings", where authors from different languages review and discuss exam drafts together to align standards (Sireci & Oliveri, 2023). Parallel development naturally removes many challenges inherent in adaptation—including translation errors and untranslatable language idiosyncrasies—since each language is developed independently, rather than representing a source version.

The *ITC Guidelines for Translating and Adapting Tests* (2017) specify the legitimacy of assessing constructs across cultural/linguistic groups must be established in multi-language assessments. In their checklist for operationalizing ITC guidelines, Hernández et al. (2020) suggested different test versions are preferable where constructs are not generalizable across populations. Parallel approaches may be more suitable in these situations. Yet, with different content in each language, fewer statistical methods are available to investigate comparability (Badham & Furlong, 2023), thus parallel development represents "a compromise between comparability and cultural authenticity" (Ercikan & Lyons-Thomas, 2013, p.552).

GenAI and Multi-Language Test Development

Recently, dramatic surges in generative AI (GenAI) have presented innovative opportunities for developing multi-language assessments. Duolingo, with over 40 languages offered through its language learning app (Blanco, 2024) offers an example. Goodwin et al. (2023) demonstrated how GenAI could support multi-language item development simultaneously and at scale. Expert judgement and open-source corpora were used to train multilingual large language models (LLMs) on extensive word-sets to create prototype Duolingo listening and reading items in French

and Spanish. GenAI has the potential to transform multilingual test development practices, by improving efficiencies for large-scale assessments, decreasing costs, and reducing workload (Hao et al., 2024). AI-supported automated item generation can "mitigate security risks and avoid overexposure of test content" (ITC, 2022, p.135), but also faces challenges including copyright and intellectual ownership (Hao et al., 2024). Additionally, there are validity concerns, as LLMs can reflect bias inherent in internet-scraped data, and models may be "biased for or against particular groups and/or produce poor outputs in under-represented languages" (ibid, p.26). Despite current limitations, GenAI offers enormous potential for supporting multi-language test development.

Test Score Linking

In many cases, score scales from multilingual assessments are linked in some fashion to facilitate interpretations and comparisons. There are several appropriate cross-lingual linking and data collection designs (Table 2) that are helpful for specific assessment contexts (ITC, 2017). We briefly describe these designs next.

In *separate monolingual groups* designs, links are formed using anchor items assumed to be comparable across languages. This assumption is generally verified via statistical analyses of DIF across languages (i.e., items not flagged for DIF are used as anchor items). However, "this justification is somewhat circular, because DIF analyses assume the variable on which examinees are matched is free of construct and method bias" (Sireci et al., 2016, p.189). Thus, it does not rule out the possibility of unidirectional bias (e.g., all items are more difficult in one language). Nevertheless, it is a helpful validation check, and has been used to link scores, or evaluate cross-lingual comparability using item response theory (IRT) (e.g., OECD, 2024).

Separate monolingual groups are used to link Psychometric Entrance Tests (PET), high-stakes college admissions exams in Israel. PET verbal and quantitative reasoning tests are adapted from Hebrew into five other languages, using adaptation and parallel development. Quantitative reasoning items are translated and DIF procedures conducted for each language using Hebrew as the reference group. However, vocabulary and analogy items are constructed uniquely in each language (Allalouf et al., 2009), since they are too different across languages to be translated (Allalouf et al., 1999). Cross-lingual DIF analyses are conducted to select items to link scales across languages. Items comprising linking anchors must demonstrate a correlation of >0.80 with respect to their item difficulty parameters across languages. This is less stringent than most equating studies (e.g., Kolen & Brennan, 2004), illustrating the lower level of linking being conducted (Sireci et al., 2016). The PET example demonstrates tests using parallel development may be statistically linked using adapted items as linking anchors.

A qualitative variation on separate monolingual groups is *social moderation*. Social moderation is the lowest level of linking, where expert judgement is used to form a link or common standard of achievement across languages. IB "cross-language standardization meetings" are an example, where examiners from parallel language versions discuss and align marking standards together (Sireci & Oliveri, 2023). Similarly, Davis et al. (2008) convened separate language panels of experts to set pass/fail standards on English and French high school reading and writing tests. The standards set on each exam resulted in about 1–6% differences in pass rates, illustrating that parallel standard setting processes using social moderation may be used to set credible standards on parallel language versions. Whilst parallel cross-lingual assessments can be linked using social moderation, this offers a "weaker" level of linking (Sireci et al., 2016) compared to methods such as IRT.

Matched monolingual groups designs have been used rarely (e.g., Milman et al., 2018), due to challenges identifying valid external criteria that can be considered equivalent across language groups, and exhibit sufficient overlap for matching purposes. Bilingual groups have been used in small-scale contexts, but have limited practical applicability in large-scale assessments due to limited availability of bilingual examinees. Ong and Sireci (2008) used a bilingual design where examinees took English and Malay 9th grade math tests in counterbalanced order. Overall, 7 of 40 items were flagged for DIF. Next, they performed linking using several methods including linear, equipercentile, and IRT, both with and without using DIF items as part of the equating anchor. The equating resulted in a 2-point adjustment across languages with DIF items included, and a 1-point difference without them. Such results underscore the need to screen items for DIF prior to linking (ibid; Sireci et al., 2016).

Evaluating Comparability

Considerable literature focuses on evaluating comparability (measurement invariance) across adapted tests, both at item-(using DIF) and test-score levels (using dimensionality procedures). Many focus on ILSAs such as PISA, TIMSS and PIRLS. Comparability studies have different purposes, including establishing measurement equivalence to justify linking procedures, or evaluating adaptation processes by identifying translation issues.

Evaluating Invariance Across Languages

Rapp and Allalouf (2003) used a *double-linking plan*—where a test form is equated to two other forms—to evaluate whether

Table 2
Cross-Lingual Linking Designs (adapted from Sireci, 1997)

Design	Assumptions	Examples
Separate monolingual groups	No systematic method bias exists across all items, which justifies DIF analyses	Allalouf et al. (2009); Angoff & Cook (1988); Hulin et al. (1982); Hulin & Mayer (1986); OECD (2024); Woodcock & Muñoz-Sandoval (1993)
Matched monolingual groups	Valid matching criteria sufficiently account for group differences. Overlap of distributions on these criteria are sufficient for matching	Milman et al. (2018)
Bilingual groups	Bilingual examinees are sufficiently representative of monolingual groups, with roughly equal proficiency across languages	Boldt (1969); Cascallar & Dorans (2005); CTB-McGraw Hill (1988); Ong & Sireci (2008); Sireci & Berberoglu (2000); Sukin et al. (2015)

the PET linking process introduced equating error. When double-linking studies are conducted in a single language, typically the two separate equating results are averaged. However, Rapp and Allaouf used within-language equating to establish a baseline equating error, before comparing it to the cross-lingual equating error. The verbal test contained a pair of parallel sections, which could be equated within each target language using a common person design (same-language equating), and to their Hebrew (source language) counterparts using linking items. Rapp and Allalouf assumed differences between the within-language and across-language equating results would reflect the instability associated with their cross-lingual linking. The average equating difference across test forms in the first target language was about ten times that observed for within-language equating forms. They concluded the within- and across-language double-linking design was useful for evaluating cross-lingual linking stability, hypothesizing various reasons for instability, including translation differences, cultural familiarity, item position effects, and different anchor test lengths.

As Sireci et al. (2016) highlight, the PET research illustrates how cross-lingual linking has been evaluated on high-stakes tests. Lower-stakes tests, including TIMSS and PISA use similar approaches (i.e., translated items, DIF-screening, and common-item linking). The approach has limitations, "as the viability of the linking anchor cannot be unequivocally established. The linking anchor may have items that differ across languages but escape DIF detection, or it may underrepresent the construct the test is designed to measure" (ibid, p.191). Allalouf et al. (2009) questioned whether "a superior, no-DIF link with an inferior representation of content" was preferable, "or an inferior link (that includes some DIF items) with a superior representation of content" (p.105).

Assessing Invariance of Dimensionality

Multi-group confirmatory factor analysis (MGCFA) has been widely used to evaluate construct bias and score comparability in cross-lingual assessment (Davidov, 2011; van de Vijver et al., 2019), partly because it can handle multiple groups in a single analysis. Multidimensional Scaling (MDS) has also frequently been used to explore comparability from a dimensionality perspective in cross-lingual research (e.g., Robin et al., 2003; Wolff et al., 2011). MDS is useful as data from multiple groups can be analyzed concurrently to determine the structural similarities (and differences) across groups by using an individual differences MDS analysis and evaluating the group weights to modify the common structure for each group (Sireci, 2005; Sireci & Wells, 2010; Sireci et al., 2016). Asparouhov and Muthén's (2014) alignment procedure can also accommodate multiple groups (e.g. countries and languages), and is more flexible than MGCFA, as it does not require parameters to be exactly equal across groups (van de Vijver et al., 2019). It accommodates partial invariance, by seeking patterns of parameter estimates that allow small variations between parameters, but only minimal large differences. The estimation stops when the overall amount of non-invariant parameters is minimised, providing "the best possible comparability that can be achieved with the given data" (ibid, p.16).

As discussed in Sireci et al. (2016), an advantage of MDS over MGCFA is its exploratory nature, so dimensionality of the assessment does not need to be specified in advance. This is helpful when the dimensionality is "unknown, or the hypothesized dimensionality is not widely supported" (ibid, p.193). The disadvantage is that MDS is solely descriptive—it provides no statistical test to evaluate structural differences across groups (Fischer & Fontaine, 2011)—necessitating reliance primarily on visual interpretations and descriptive indices (Sireci et al., 2016). Dimensionality structure across cultures is not sufficient to ensure score comparability. Therefore, MDS is useful for exploratory purposes, but on its own, is insufficient for establishing measurement equivalence to justify linking procedures. The strictness of MGCFA—requiring exact equality of parameters across groups—is also a drawback, since this requirement is rarely met in practice (van de Vijver et al., 2019). The flexibility of the alignment procedure makes it more practical than MGCFA for real-life data analysis. Whilst a useful investigative technique, as it allows partial invariance, it is insufficient for establishing measurement equivalence. Therefore, such methods are typically combined with procedures like DIF to justify linking.

Evaluating Invariance of Adapted Items

DIF procedures are commonly used to evaluate cross-lingual comparability at item-level, often combined with structural equivalence or qualitative analyses to help interpret cross-lingual differences. Grisay et al. (2009) evaluated deviations of item difficulty parameters for countries participating in PISA and PIRLS reading assessments, from "international" item parameters using the global population. Despite a large commonality across the global and country-specific item difficulty parameters and only a modest level of DIF on average, higher magnitudes of DIF were noted for non-Indo-European languages (e.g., Arabic and Chinese). Gökçe et al. (2021) also investigated whether DIF on TIMSS math was associated with differences between language families and cultures. They compared DIF across three language-country combinations: (a) same language, but different countries, (b) same countries, but different languages, and (c) different languages and countries. With more distant cultures and language families, the presence of DIF increased. The magnitude of DIF was greatest when both language and country differed, and smallest when languages were same, but countries were different.

Ercikan and Koh (2005) investigated English and French TIMSS math and science assessments across countries using DIF and structural equivalence using MGCFA. There was a lack of equivalence at both structural and item-levels, with substantial DIF found in some comparisons (e.g., 79% of science items flagged for DIF across France and the U.S.). The global fit indices associated with the MGCFA illustrated relatively worse fit of the models to the data where the greatest amount of DIF was observed. Ercikan and Koh cautioned against making cross-lingual comparisons when substantial DIF and inconsistencies in test structure are observed across translated assessments. Similarly, Oliveri et al. (2012) evaluated item- and test-level comparability of English and French PISA mathematics problem-solving subtests. Although 3 of 10 items functioned differentially across languages, when aggregating these results to evaluate differential test functioning, they found comparable test characteristic curves, suggesting comparability overall. Their study illustrates the importance of considering invariance at both test and item-levels, as

item-level differences may balance out, causing no apparent effect at test-level (Wainer et al., 1991, Sireci et al., 2016).

Allalouf et al. (1999) followed DIF analyses with qualitative investigations. Hebrew-Russian bilingual content specialists and translators investigated items flagged for DIF in Russian translations of PET verbal reasoning items. They identified four potential causes of DIF: word familiarity and frequency across languages; content changes due to translation; item format; and cultural relevance. Similarly, Gierl and Khaliq (2001) examined English and French, 6th and 9th-grade math and social tests, where bilingual content specialists hypothesized potential sources of DIF on flagged items. Translators then categorized items flagged for DIF on a subsequent assessment into the hypothesized categories, illustrating how previously identified sources of DIF could be used to explain subsequently flagged items. The identified sources of DIF also aligned with Allalouf et al. (1999) study, although different languages were involved (Sireci et al., 2016).

Computational Linguistics

Computational linguistics is increasingly used to investigate cross-lingual differences. El Masri et al. (2016) used computational linguistics to identify linguistic intricacies across languages in PISA science items. They noted idiosyncrasies may be overlooked in expert review-based quality assurance processes, and recommended computational linguistics tools (e.g., Educational Testing Service's *Text-Evaluator Tool*) for evaluating text complexity and identifying differences across translated assessments. Similarly, McGrane et al. (2022) used computational linguistics to examine linguistic complexity across languages in IB science exams. Natural Language Processing (NLP) techniques using a multilingual text processing framework were used to analyze large DIF items across languages. Differences in linguistic complexity explained up to 11% of DIF results. They recommended that text analysis tools be used during item development to examine item complexity across languages. AI-based NLP techniques can be particularly useful in test development contexts where piloting may be infeasible (e.g. due to reduced timelines) (ITC, 2022).

Discussion

Our review illustrated different approaches to develop, link, and evaluate cross-language comparability. Adaptation is most common, with iterative, team-based approaches preferred over back-translation. Simultaneous item development helps prevent language prioritization, and identifies cross-lingual and cross-cultural issues during adaptation processes. Parallel development, though rare, is useful when adaptation cannot adequately capture constructs. Emerging GenAI tools show promise but raise concerns over intellectual ownership and potential biases in LLMs.

Test development balances comparability and cultural authenticity (Ercikan & Lyons-Thomas, 2013). Adapted tests enhance comparability through anchor items, but face challenges in translation and ensuring cultural relevance. Parallel development largely removes challenges with translation and language differences, thereby maximizing cultural authenticity. However, with fewer statistical techniques available, comparability and linking are inherently weaker. Hybrid approaches—such as adapting items in parallel tests—offer a compromise between comparability and cultural authenticity, as stronger linking can be established with adapted items as anchors across languages. (e.g. Allalouf, 2009).

Empirical studies have evaluated comparability of dimensionality, items, and achievement level standards from cross-lingual tests (Table 3).

Table 3
Selected Summary of Comparability Studies

Citation	Context	Validity Evidence	Statistical Analyses	Findings
Alatli (2020, 2022)	PISA science & reading	Internal structure	DIF, MGCFA	Only structural invariance held. Approx. 35% of science items exhibited DIF due to translation issues; 5 of 7 reading items displayed DIF across China and Turkey.
Allalouf et al. (1999)	PET verbal tests	Internal structure, Test content	DIF	DIF explained by differential difficulty caused by translation, item format, or cultural relevance.
Cascallar & Dorans (2005)	SAT, PAA, & ESLAT	Relations to other variables	Multiple regression	Bilinguals used to compute predicted scores on SAT from PAA and ESLAT.
Davis et al. (2008)	High school reading & writing	Test content	n/a	Setting standards on each test simultaneously using bilingual translators and facilitators to ensure consistent processes across languages.
Ercikan & Koh (2005)	TIMSS math and science	Internal structure	DIF, MGCFA	Structure of assessments was inconsistent across languages in some countries and substantial DIF was found.
Gierl & Khaliq (2001)	Math and social studies tests	Internal structure, Test content	DIF	Bilingual translators and content specialists identified causes of DIF, confirmed by content and statistical analyses on a similar test.
Grisay et al. (2009)	PISA & PIRLS reading	Internal structure	DIF	Greater DIF for non-Indo-European languages.
Gökçe et al. (2021)	TIMSS math	Internal structure	DIF	As differences between language families and cultures increased, observed DIF increased.
McGrane et al. (2022)	IB sciences	Internal structure, Test content	DIF, NLP	Linguistic complexity accounted for up to 11% of variance of DIF.
Oliveri et al. (2012)	TIMSS math	Internal structure	DIF, MGCFA	Whilst 3 of 10 items functioned differentially across languages, DIF did not manifest at test score level.
Rapp & Allalouf (2003)	PET verbal test	Internal structure	Equating analyses	Equating error across language versions was 10x larger than within-language equating error.

Studies focusing on internal structure as sources of validity evidence were most common, using DIF procedures to evaluate item invariance and MGCFA to evaluate structural (dimensional) equivalence. Computational linguistics techniques including text analysis tools offer opportunities for evaluating cross-lingual comparability post-hoc and during test development. Most cross-lingual assessment research indicates many items are differentially difficult across languages, but also that differences are not in one systematic direction, and sufficient comparability likely exists. Some degree of non-invariance must be expected in cross-lingual assessment, as it is unrealistic to assume all items will function equally across all subpopulations (Oliveri & von Davier, 2011, 2014, 2017). Having most, but not all, items from different languages on the same scale is more realistic, and likely sufficient for most comparability needs (ibid). No studies focusing on validity evidence based on testing consequences were found, which is an area recommended for future research.

Adaptation/development approaches have different benefits and drawbacks, including different analyses being available for linking and evaluating comparability (Table 4).

Selection of appropriate multi-language assessment methods depends on the specific context of assessments (e.g. content area, language combinations, or large-scale versus small-scale). The importance of score comparability will always depend on the test purpose, and the decisions and actions taken based on scores. The advantages and challenges for different multi-language development approaches presented here may guide practitioners to choose the most appropriate approach for their contexts. We hope this review, and the many studies referenced, help test developers and evaluators build more valid cross-lingual assessments.

Table 4
Benefits and Challenges of Multi-Language Development Approaches

Development approach	Benefits	Challenges
(Successive) adaptation	Stronger link across languages, Established statistical methods to investigate equivalence (e.g. DIF).	Cultural relevance & authenticity, Translation errors, Language differences (e.g. language idiosyncrasies, word frequencies, differential speededness).
Simultaneous development	Stronger link across languages, Established statistical methods to investigate equivalence (e.g. DIF), Linguistic and cultural decentering, Reduced review time.	Language differences (e.g. language idiosyncrasies, word frequencies, differential speededness).
Parallel development	Cultural relevance & authenticity, Removes risk of translation errors, Reduces impact of language differences.	Weaker link across languages, Labour intensive, Harder to investigate comparability statistically.
GenAI	Time efficient, Cost effective, Reduced labour, Reduced security risk, Lowers exposure of test content.	Copyright/intellectual ownership, Risk of bias, Not sufficiently developed in all languages.

Author Contributions

Louise Badham: Conceptualization, Methodology, Project administration, Formal analysis, Writing_Review and Editing. **María Elena Oliveri:** Methodology, Investigation, Formal analysis, Writing_Review and Editing. **Stephen G. Sireci:** Funding acquisition, Methodology, Investigation, Formal analysis, Writing – original draft.

Funding

This research was funded by the International Baccalaureate (IB). The first author is an employee of the IB, and participated throughout the study.

Declaration of Interests

The authors declare there are no conflicts of interest. The views expressed are those of the authors and not to be taken as views of the IB.

Data Availability Statement

The data supporting this review are available within the cited references.

References

Alatli, B. (2020). Cross-cultural measurement invariance of the items in the Science Literacy Test in the Programme for International Student Assessment (PISA-2015). *International Journal of Education and Literacy Studies*, *8*(2), 16–27.

Alatli, B. (2022). An investigation of cross-cultural measurement invariance and item bias of PISA 2018 reading skills items. *International Online Journal of Education and Teaching*, *9*(3), 1047–1073.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185–198. https://doi.org/10.1111/j.1745-3984.1999.tb00553.x

Allalouf, A., Rapp, J., & Stoller, R. (2009). Which item types are better suited to the linking of verbal adapted tests? *International Journal of Testing*, *9*(2), 92–107. https://doi.org/10.1080/15305050902880686

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Académica and the Scholastic Aptitude Test (Report No. 88-2). *ETS Research Report Series*. https://doi.org/10.1002/j.2330-8516.1988.tb00259.x

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Badham, L., & Furlong, A. (2023). Summative assessments in a multilingual context: What comparative judgment reveals about comparability across different languages in Literature. *International Journal of Testing*, *23*(2), 111-134. https://doi.org/10.1080/15305058.2022.2149536

Blanco, C. (2024). *2024 duolingo language report.* Duolingo. https://blog.duolingo.com/2024-duolingo-language-report/

Boldt, R. F. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers* (*College Entrance*

Examination Board Research and Development Report 68-69, No. 3). Educational Testing Service.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration of methodological alternatives. *International Journal of Testing*, *5*(4), 337–356. https://doi.org/10.1207/s15327574ijt0504_1

CTB/McGraw-Hill (1988). *Spanish assessment of basic education: Technical report*. McGraw Hill.

Davidov, E. (2011). Nationalism and constructive patriotism: A longitudinal test of comparability in 22 countries with the ISSP. *International Journal of Public Opinion Research*, *23*(1), 88–103. https://doi.org/10.1093/ijpor/edq031

Davis, S. L., Buckendahl, C. W., & Plake, B. S. (2008). When adaptation is not an option: An application of multilingual standard setting. *Journal of Educational Measurement*, *45*(3), 287–304. https://doi.org/10.1111/j.1745-3984.2008.00065.x

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust and R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 168–191). Wiley. https://doi.org/10.1002/9781118762462.ch6

Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, *49*(1), 1–18. https://doi.org/10.1111/j.1745-3984.2011.00157.x

Ebbs, D., & Korsňáková, P. (2016). Translation and translation verification for TIMSS 2015. In Martin, M. O., Mullis I. V. & Martin H. (Eds.), *Methods and procedures in TIMSS 2015* (pp. 7.1–7.16). TIMSS & PIRLS International Study Center, Boston College.

Ebbs, D., Flicop, S., Hidalgo, M. M., & Netten, A. (2021). Systems and instrument verification in PIRLS 2021. In *Methods and procedures: PIRLS 2021 technical report* (pp. 5.1–5.24). TIMSS & PIRLS International Study Center, Boston College. https://doi.org/10.6017/lse.tpisc.tr2103.kb2485

El Masri, Y. H., Baird, J-A., & Graesser, A. (2016) Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3–4), 199–215. https://doi.org/10.1080/15305058.2002.9669493

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23–35. https://doi.org/10.1207/s15327574ijt0501_3

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K.Geisinger, B. Bracken, J. Carlson, J-I. Hansen, N. Kuncel, S. Reise, & M. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 545–569). American Psychological Association. https://doi.org/10.1037/14049-026

Ercikan, K., & Por, H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). National Academy of Education Press. https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto and F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 179–215). Cambridge University Press. https://doi.org/10.1017/CBO9780511779381.010

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164–187. https://doi.org/10.1111/j.1745-3984.2001.tb01121.x

Gökçe, S., Berberoglu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on Trends in International Mathematics and Science Study mathematics assessment items. *Journal of Psychoeducational Assessment*, *39*(6), 728–745. https://doi.org/10.1177/07342829211010537

Goodwin, S., Bilsky, L., Mulcaire, P., & Settles, B. (2023, 26–28 April). *Machine learning applications to develop tests in multiple languages simultaneously and at scale* [Conference presentation]. Association of Language Testers in Europe 8th International Conference, Madrid, Spain.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, *20*(2), 225–240. https://doi.org/10.1191/0265532203lt254oa

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 63–83.

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

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence Erlbaum Publishers.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). Cambridge University Press.

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16–29. https://doi.org/10.1111/emip.12602

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *32*(3), 390–398. https://doi.org/10.7334/psicothema2019.306

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*(6), 818–825. https://doi.org/10.1037/0021-9010.67.6.818

Hulin, C.L., & Mayer, L.J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, *71*(1), 83–94. https://doi.org/10.1037/0021-9010.71.1.83

International Baccalaureate Organization (2018). *Assessment principles and practices—Quality assessments in a digital age*. International Baccalaureate Organization. https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/dp-final-statistical-bulletin-may-2024_en.pdf

International Baccalaureate Organization (2024). *The IB Diploma Programme and Career-Related Programme: May 2024 assessment*

session final statistical bulletin. International Baccalaureate Organization. https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/the-ib-dp-and-cp-statistical-bulletin_en.pdf

International Test Commission. (2017). *ITC Guidelines for translating and adapting tests (2nd edition)*. International Test Commission. https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

International Test Commission and Association of Test Publishers (2022). *Guidelines for technology-based assessments*. International Test Commission and Association of Test Publishers. https://www.intestcom.org/upload/media-library/tba-guidelines-final-2-23-2023-v4-1677851446 42TgY.pdf

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices (2nd edition)*. Springer-Verlag.

Koršňáková, P., Dept, S., & Ebbs, D. (2020). Translation: The preparation of national language versions of assessment instruments. In H. Wagemaker (Ed.), *Reliability and validity of international large- scale assessment: Understanding IEA's comparative studies of student achievement, volume 10* (pp. 85–111). IEA Research for Education, Springer, Cham. https://doi.org/10.1007/978-3-030-53081-5_6

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Publishers.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and procedures: TIMSS 2019 technical report. *International Association for the Evaluation of Educational Achievement*. https://timssandpirls.bc.edu/timss2019/methods/

McGrane, J., Kayton, H., Double, K., Woore, R., & El Masri, Y. (2022). *Is science lost in translation? Language effects in the International Baccalaureate Diploma Programme science assessments*. Oxford University Centre for Educational Assessment. https://ibo.org/globalassets/new-structure/research/pdfs/ib-dp-sciene-translation-final-report.pdf

Milman, L. H., Faroqi-Shah, Y., Corcoran, C. D., & Damele, D. M. (2018). Interpreting mini-mental state examination performance in highly proficient bilingual Spanish–English and Asian Indian–English speakers: Demographic adjustments, item analyses, and supplemental measures. *Journal of Speech, Language, and Hearing Research*, *61*(4), 847–856.

OECD. (2016). PISA 2018 translation and adaptation guidelines. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/pisa-database/survey-implementation-tools/pisa-2018/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf

OECD. (2024). *PISA 2022 Technical Report.* OECD Publishing. https://www.oecd.org/en/publications/pisa-2022-technical-report_01820d6d-en.html

Oliveri, M. E., Olson, B., 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. https://doi.org/10.1080/15305058.2011.617475

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. https://doi.org/10.1080/15305058.2013.825265

Oliveri, M. E., & von Davier, M. (2017). Analyzing invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R.W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp.121–146). Information Age Publishing.

Ong, S. L., & Sireci, S. G. (2008). Using bilingual students to link and evaluate different language versions of an exam. *US-China Education Review*, *5*(11), 37–46.

Rapp, J., & Allalouf, A. (2003). Evaluating cross-lingual equating. *International Journal of Testing*, *3*(2), 101–117. https://doi.org/10.1207/S15327574IJT0302_1

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*(1), 1–20. https://doi.org/10.1207/S15327574IJT0301_1

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *Alberta Journal of Educational Research*, *49*(3), 290–304. https://doi.org/10.11575/ajer.v49i3.54986

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2010). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education*, *24*(1), 39–70. https://doi.org/10.1080/08957347.2011.532416

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19. https://doi.org/10.1111/j.1745-3992.1997.tb00581.x

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

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated- adapted items. *Applied Measurement in Education*, *13*(3), 229–248. https://doi.org/10.1207/S15324818AME1303_1

Sireci, S. G., & Oliveri, M. E. (2023). *A Critical Review of the International Baccalaureate Organization's Multilingual Assessment Processes and Best Practices' Recommendations [Report for the IB]*. International Baccalaureate Organization.

Sireci, S. G., Rios, J. A., & Powers, S. (2016). Comparing test scores from tests administered in different languages. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 181–202). Routledge.

Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Council of Chief State School Officers.

Sukin, T., Sireci, S. G., & Ong, S. L. (2015). Using bilingual examinees to evaluate the comparability of test structure across different language versions of a mathematics exam. *Actualidades en Psicología*, *29*(119), 131–139. http://doi.org/10.15517/ap.v29i119.19244

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

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 263–279. https://doi.org/10.1016/j.erap.2003.12.004

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–64). Lawrence Erlbaum Publishers.

van de Vijver, F., Avvisati, F., Davidov, E., Eid, M., Fox, J. P., Le Donné, N., Lek, K., Meuleman, B., Paccagnella, M., & Van de Schoot, R. (2019). *Invariance analyses in large-scale studies*. OECD Publishing.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*(3), 197–219. https://doi.org/10.1111/j.1745-3984.1991.tb00354.x

Wolff, H. G., Schneider-Rahm, C. I., & Forret, M. L. (2011). Adaptation of a German multidimensional networking scale into English. *European Journal of Psychological Assessment*, *27*(4), 244–250. https://doi.org/10.1027/1015-5759/a000070

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment*, *9*(3), 233– 241.

Zhao, X., Solano-Flores, G., & Qian, M. (2018). International test comparisons: Reviewing translation error in different source language-target language combinations. *International Multilingual Research Journal*, *12*(1), 17–27. https://doi.org/10.1080/19313152.2017.1349527

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136–151. https://doi.org/10.1080/15434303.2014.972559