frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 14 December 2023
DOI 10.3389/feduc.2023.1303617

Check for updates

OPEN ACCESS

EDITED BY
Raman Grover,
Consultant, Vancouver, BC, Canada

REVIEWED BY
Attiya Nourdine,
Moulay Ismail University, Morocco
Yi-Hsin Chen,
University of South Florida, United States

*CORRESPONDENCE
Xueyu Zhao
✉ xueyu.zhao@dukekunshan.edu.cn

RECEIVED 28 September 2023
ACCEPTED 29 November 2023
PUBLISHED 14 December 2023

CITATION
Zhao X and Solano-Flores G (2023) Test translation review: a study on discussion processes and translation error detection in consensus-based review panels.
Front. Educ. 8:1303617.
doi: 10.3389/feduc.2023.1303617

Test translation review: a study on discussion processes and translation error detection in consensus-based review panels

Xueyu Zhao[1]* and Guillermo Solano-Flores[2]

[1]Language and Culture Center, Duke Kunshan University, Kunshan, China, [2]Graduate School of Education, Stanford University, Stanford, CA, United States

We examined the discussion processes through which two independent consensus-based review panels detected errors in the same sample of items from an international test translated from English to Chinese. The discussion processes were defined according to four events: (1) *identifying* a potential error; and (2) *agreeing with*, (3) *disagreeing with*, and (4) *elaborating* an opinion expressed by other panelists. We found that, while the two panels had similar error detection rates, only half of the errors detected by the two panels altogether were detected by both panels. In addition, of the errors detected by the two panels, more than half were detected by the panels through different discussion processes. No discussion process occurred substantially more frequently or less frequently for any translation error dimension. We conclude that the unique combination of backgrounds, skills, and communication styles of panel members and the unique combination of textual features in each item shape which errors each panel is capable of detecting. While panels can be highly effective in detecting errors, one single panel may not be sufficient to detect all possible errors in a given set of translated items. Consensus-based translation error review panels should not be assumed to be exchangeable.

KEYWORDS

test translation, test translation review, consensus-based procedures, discussion process, international comparisons

Introduction

As part of the process of globalization in education, international tests are routinely administered in multiple languages to different populations with the intent to assess student knowledge in various disciplinary areas and forms of competencies across countries (Kamens and McNeely, 2010; Suter, 2019). Tests such as PISA (Programme for International Student Assessment) and TIMSS (Trends in Mathematics and Science Study) have become powerful forces that influence education reform and policy in many participating countries (Teltemann and Klieme, 2017; Solano-Flores, 2019; Bray et al., 2020).

This increasing influence of international tests on the lives of countries speaks to the importance of ensuring construct equivalence and creating culturally-responsive assessment tools across languages (Hambleton et al., 2005; Sireci et al., 2005; Trumbull and Nelson-Barber, 2019; Berman et al., 2020; Kūkea Shultz and Englert, 2021). Specifically, translation may alter the difficulty of tests and the knowledge, skills, or competencies these tests are intended to assess (Cook and Schmitt-Cascallar, 2005). This may be especially the case in international test comparisons: While many items contain situations and characters intended to make problems meaningful to students (Ruiz-Primo and Li, 2015), these situations and characters may not be equally familiar to all test takers and their insertion in the text of items may increase reading load.

A wide variety of quantitative procedures for examining bias have been available for a long time (e.g., Camilli and Shepard, 1994). Such procedures can be used to detect linguistic bias resulting from translating items. Yet unfortunately, such procedures are not used routinely in large-scale assessment systems, due to cost and the fact that they need large samples of pilot student responses (Allalouf, 2003). As long as these procedures are not used systematically and exhaustively, there will be a serious need for effective item review procedures that allow low-cost, effective detection of the multiple sources of error that may adversely impact the constructs assessed when items are translated.

This paper addresses the need for improved judgmental procedures that ensure item equivalence across languages (Hambleton, 1994, 2001, 2005; Sireci et al., 2006; International Test Commission, 2017). It focuses on consensus-based procedures as critical to detecting translation error. In these procedures, panels of reviewers with different areas of expertise (e.g., linguists, certified translators, curriculum experts, assessment specialists, classroom teachers, content area experts) examine test items, discuss features that may constitute translation error, and classify those features according to typologies of translation error (Solano-Flores et al., 2009, 2013; Zhao et al., 2018).

The investigation here reported is part of a research agenda that addresses two seemingly conflicting facts: Differences due to socio-economic or professional status may influence how people interact in group decision-making (Strohschneider, 2002; Weber et al., 2005); yet consensus-based procedures assume the occurrence of rich, constructive discussions in which all members' opinions are valued equally (Fink et al., 1984). A previous study (Zhao and Solano-Flores, 2021) examined person-to-person interactions in two translation review panels. That study found that, while the interactions between members may be influenced by social status differences, with proper facilitation, panels with different cultural makeups can be comparably effective in detecting translation errors.

While that previous study focused on person-to-person interactions, the present study focused on the group discussions that lead review panels to detect translation error on different dimensions. We examined how any differences in the discussion processes through which two independent consensus-based review panels detected translation errors in a sample of translated test items were associated to differences in their translation error detection rates. Findings from this investigation contribute to the improvement of test translation review and, ultimately, to promoting more valid and fair testing in international test comparisons. Findings from this investigation also contribute to a better understanding of group processes involved in problem-solving with a probabilistic view: While each panel is unique due to the personal backgrounds, histories, personalities, and communication styles of its members consensus-review panels are implicitly assumed to be exchangeable.

Theoretical perspective

This paper builds on the theory of test translation error (Solano-Flores et al., 2009), according to which translation error occurs not only due to flaws in the translator's job but mainly because languages do not encode meaning in the same ways. The theory is consistent with a probabilistic view according to which, due to complexity, random factors, and the limited information available about linguistic groups, there is always a level of uncertainty in our understanding of language-related phenomena (Bod et al., 2003; Solano-Flores, 2014; Oliveri, 2019). This probabilistic view is in contrast with conventional, deterministic approaches to test translation, which implicitly assume linguistic homogeneity in the populations tested with translated tests.

A probabilistic view recognizes the existence of random events that shape the extent to which translation preserves meaning and the level of difficulty across test items. Accordingly, an optimal translation minimizes error but cannot be error-free. In translating many forms of text, translators have at least some leeway to use multiple resources (e.g., elaborating sentences, using multiple words, using different but culturally equivalent contextual information) to ensure that meaning is preserved across languages. In contrast, when tests are translated, the use of these resources is restricted by the linguistic properties of test items (e.g., compact language style, short sentences, pre-established format, content load of terms). Because the content of items is intimately related to the characteristics of the language in which they are administered (AERA/APA/NCME, 2014), translation may alter the constructs that items are intended to measure.

Multidimensionality is a fundamental notion in the theory of test translation error and the main justification in support of using consensus-based procedures in test translation review. Because textual features act in combination to convey meaning (Halliday, 1978; Kress, 2010), the same given translation error may belong to several translation error dimensions (Table 1). For example, in addition to being a grammar error, the literal translation of a sentence in an item can also be an error related to construct when, as a result of that literal translation, the item is likely to end up assessing different forms of knowledge or skills. Multiple reviewers with different formal backgrounds are assumed to be more effective than individual reviewers in addressing the multiple facets of language and language use involved in test translation. Available evidence shows that this approach allows identification of translation error with a high level of precision (Solano-Flores et al., 2009, 2013; Zhao, 2018).

Research question

We asked: *How are discussion processes different across consensus-based review panels and how are any differences related to translation error detection effectiveness in the Chinese context?*

Methods

Item sample

A total of 19 English test items from the Programme for International Student Assessment (PISA) administered in 2009 and 2012, together with their Chinese translations, were examined in this investigation. The English language versions of the test items were released by the Organization for Economic Co-operation and Development (OECD) and were retrieved on their official website

(OECD, 2015). The translated PISA items in Chinese were not retrieved from the Chinese office of PISA; they were downloaded from some Chinese testing websites that were designed specifically for secondary education by using a Chinese search engine.

Due to the fact that the Chinese PISA items are not officially released, the translated items reviewed by the panelists in this investigation may not be the same translated versions used with students in Shanghai in 2009 and 2012. However, this fact does not affect the integrity of our findings, since we did not attempt to relate the characteristics of the items to student performance data.

Participants

Two translation review panels participated in this study. Panel 1 comprised seven individuals: three university professors, one senior teacher with expertise in assessment, two translators, and one classroom teacher. Panel 2 comprised five individuals: two university professors, one senior teacher with experience in assessment, and two classroom teachers. Within each panel, important differences in social status were assumed to exist due to factors such as the prestige of the panelists' professions, their academic degrees, and their salaries. In Chinese culture, university professors, assessment specialists, and teachers have, respectively, a higher, medium, and lower social status (Burnaby and Sun, 1989).

The two panels had different multiple socio-demographic makeups. Reviewers in Panel 1 were located in Northeast China, whereas reviewers in Panel 2 were located in Southeast China—regions which differ substantially on dialects, local subcultures, and levels of economic development. The review panels also varied considerably in average age and years of experience in the field (Table 2). All members in Panel 1 were born after the establishment of the one-child policy, were relatively young, and tended to have fewer years of teaching experience (only one professor on this panel had more than 10 years of teaching experience, while the others had less than 6 years of experience). In contrast, all members on Panel 2 were born during the Cultural Revolution before the establishment of the one-child policy and had more than 18 years of teaching experience. Because their members differed extremely in their geographical regions of origin, ages, generational cohorts, and years of experience, the panels were deemed likely to have different sets of generational values, local or regional identities, cultural backgrounds, and communication styles.

Translation review sessions and error detection

Two full-day translation review sessions were staged with each panel in its local region. Panelists were trained in the use of a typology of test translation error whose structure is shown in Table 1. This classification system identifies 91 types of errors grouped in ten translation error dimensions. While many test translation errors can be regarded as universal (e.g., possible alteration of the cognitive demands of the item), some errors are specific to a given combination of source language and target language (Zhao et al., 2018). The set of types of errors identified within each dimension (only a few of which are shown in the Examples column of the table) are relevant to detecting English-to-Chinese test translation error (Zhao et al., 2018). This typology was created by adapting a typology originally created for English-to-Spanish test translation review (Solano-Flores et al., 2009, 2013) by eliminating errors that are not relevant to English-to-Chinese translation review (e.g., *inappropriate use of tenses*) and adapting some errors to their equivalent in Chinese (e.g., *wrong*

TABLE 1 Translation error dimensions and examples of error types: English-to-Chinese translation.

Dimension	Definition	Examples
Style	The item is written in a style that is not consistent with the style used in academic contexts in the target language/culture.	Wrong character • Missing marks
Format	The item contains format features that are different from the format features used in the original version.	*Change in the position of graphic components* • *Use of boldface not in the original* • *Undue capitalization*
Conventions	The item is written using test writing conventions not used in target language/culture.	*Uncommon use of punctuation to denote continuity between stem and options* • *Grammatical inconsistency between options*
Grammar and Syntax	The item contains grammatic errors or its syntactical structure is excessively complex or uncommon in the target language/culture.	*Inappropriate use of prepositions* • *Inappropriate use of classifiers* • *Unnatural syntactic structure*
Semantics	The meaning of ideas in the original item have been altered.	*Use of terms with multiple meanings* • *Possible alteration of modal verbs*
Register	The item contains terms or expressions that are uncommon or unfamiliar in the cultural contexts in the target language/culture.	*Use of terms in ways that differ from use in curriculum* • *Translation of a technical term in a way that is not used in the target culture*
Information	The amount or kind of information provided in the original item have been altered.	*Omission of a sentence or explanation* • *Emphasis of information changes because parentheses are eliminated*
Construct	The knowledge and skills assessed by the item may have been altered.	*Possible alteration of the cognitive demands of the item* • *Translation of a technical term as a non-technical term*
Curriculum	The content assessed by the item is not taught in the target language/culture.	*The concept assessed is not taught at the corresponding grade level* • *Discursive style not used in the curriculum*
Origin	The item contains errors in the original version that are carried over to the translation.	*More than one correct option* • *None of the options is entirely correct*

TABLE 2 Reviewers' demographic information, specialty, social status, and years of professional experience.

Gender	Age	Specialty, social status, years of professional experience
Panel 1		
F	36	High social status, associate professor in English linguistics with a Ph.D. and 14 years of teaching
M	31	High social status, assistant professor in physics with a Ph.D. and 3 years of teaching
F	29	High social status, assistant professor in French literature with a Master of Arts and 5 years of teaching
F	29	Medium social status, assessment specialist and language arts teacher with a Master of Arts degree and 6 years of teaching
F	24	Low social status, math teacher with a Master of Science degree and 3 years of teaching
F	24	Low social status, certified translator with a Master of Translation degree and 2 years of teaching
F	23	Low social status, certified translator with a Master of Translation degree and 1 years of teaching
Panel 2		
F	45	High social status, associate professor in English linguistics with a Ph.D. and 22 years of teaching
F	42	High social status, associate professor in educational psychology with a Ph.D. and 19 years of teaching
F	46	Medium social status, assessment specialist and English teacher with a Bachelor of Science degree and 23 years of teaching
M	41	Low social status, math teacher with a Bachelor of Science degree and 18 years of teaching
M	42	Low social status, science teacher with a Bachelor of Science degree and 19 years of teaching

Adapted from "Test Translation Review Procedures in International Large-Scale Assessment: Sensitivity to Culture and Society," by Zhao (2018), Doctoral dissertation. University of Colorado Boulder.

TABLE 3 Discussion processes in the identification of translation errors.

Event	Definition
Identification	A reviewer proposes a language feature as translation error and the translation error dimensions on which it should be coded (initial activity).
Disagreement	A reviewer disputes other reviewers' ideas concerning the identification of an error or the proposed translation error dimensions on which it be coded. The reviewer may propose alternative translation error dimensions.
Agreement	A reviewer agrees with other reviewers' ideas concerning the identification of an error or the proposed translation error dimensions on which to code it.
Elaboration	A reviewer builds an argument in support of other reviewers' ideas concerning the identification of an error or the proposed translation error dimensions on which to code it.

spelling was adapted into *wrong character* because Chinese characters are composed of strokes).

In both panels, reviewers were first asked to examine individually the original version of each item and its translation, identify possible translation errors, and propose the translation error dimensions on which each error should be coded. With the first author's facilitation, each panel discussed whether each feature originally identified by each panelist was truly an error and, if so, the translation error dimensions on which that error should be coded. Final coding decisions for each error were made by consensus after all the discrepancies were discussed and addressed. The sequence in which panelists shared with each other the features that they originally identified as potential errors was pre-determined; it was created in a way intended to ensure that all panelists had equal opportunities to express their initial thoughts.

Coding of discussion processes

The review sessions were video- and audio-recorded and then coded. Through an iterative process of review, coding, and revision of the translation review sessions, we developed a system of four categories of events or forms of participation that took place during the discussions held by the panels, and which contributed to the detection of translation errors (Table 3). The first event, *identification* (ID) initiated all the discussions. Following this event, any combination of three other events would follow: *agreement* (AG), *disagreement* (DI), and *elaboration* (EL). For simplicity, we defined a discussion process according to a specific combination of events, regardless of the order in which they occurred and regardless of whether they occurred several times during the same discussion. Accordingly, we identified the discussion processes shown in Table 4.

Data analysis

To respond to our research question (*How are discussion processes different across consensus-based review panels and how are any differences related to translation error detection effectiveness in the Chinese context?*), we examined the number of translation errors detected only by each panel and the translation errors detected by both panels across translation error dimensions. We also examined the frequency with which each panel identified different types of translation errors through different combinations of discussion processes.

Results

In total, 172 translation errors were detected by the two panels altogether. Of those 172 errors, 46 and 40 were identified, respectively, only by Panel 1 and only by Panel 2, and 86 were identified by both panels. Thus, only half of the total of errors detected were detected by both panels.

Table 5 shows these totals in a Venn diagram along with their breakdowns in percentages by translation error dimension. The table shows that, while the panels had similar detection rates, the structure of relative frequencies of translation errors detected across dimensions was very different across panels. Indeed, the only commonality observed was for errors in the Semantics dimension, which were the most frequently detected by each panel and by both panels.

An examination of the discussion processes through which each panel detected its own set of errors also revealed different trends across panels (Tables 6, 7). First, while the most frequent discussion process on Panel 1 was ID (60.88%), the most frequent discussion process in Panel 2 was ID-AG (45%). Second, the patterns of relative frequencies with which errors belonging to different dimensions were detected varied across panels. Errors belonging to the Semantics (39.13%), Construct (17.39%), and Grammar Syntax (15.22%) dimensions, were, in that order, the most frequently detected by Panel 1. In contrast, errors belonging to the Semantics (35%), Information (15%), and Construct (10%) dimensions, were, in that order, the most frequently detected by Panel 2.

TABLE 4 Discussion process identified.

Abbreviation	Description
ID	Identification
ID-DI	Identification-Disagreement
ID-AG	Identification-Agreement
ID-DI-AG	Identification-Disagreement-Agreement
ID-EL	Identification-Elaboration
ID-DI-EL	Identification-Disagreement-Elaboration
ID-AG-EL	Identification-Agreement-Elaboration
ID-DI-AG-EL	Identification-Disagreement-Agreement-Elaboration

TABLE 5 Translation errors detected by each panel and by both panels by translation error dimension: percentages.

Translation error dimension	Panel 1 (*n* = 46)	Panel 1 and Panel 2 (*n* = 86)	Panel 2 (*n* = 40)
Construct	17.39	3.48	10.00
Conventions	4.35	2.32	0
Curriculum	6.52	1.16	5.00
Format	4.35	16.28	7.50
Grammar and Syntax	15.22	17.43	10.00
Information	4.35	11.63	15.00
Origin	2.17	2.32	7.50
Register	0	2.32	2.50
Semantics	39.13	37.20	35.00
Style	6.51	5.81	7.50

An examination of the discussion processes involved in the commonly detected errors also revealed different trends across panels. A matrix of correspondence (Table 8) shows the percentages of errors detected by each combination of discussion processes observed in the two panels. For example, of the 86 commonly detected errors, about 33.72% were detected through ID by both Panel 1 and Panel 2; and about 26.74% were detected through ID by Panel 1 and through ID-AG by Panel 2. Two facts stand out. First, less than a half (45.34%) of the commonly detected errors were detected through the same given discussion process (as shown by the addition of the cells of the main diagonal). Second, ID was the most frequent discussion process used by the two panels in the detection of commonly detected errors.

In sum, while they were comparably effective in detecting errors, the two panels were able to detect different sets of translation errors and only half of all the errors detected by the two panels were detected by both panels. In turn, less than a half of the commonly detected errors were detected by the two panels through the same discussion process.

Summary and concluding remarks

This paper addresses the process of consensus-based test translation review. We analyzed the discussions held by translation review panels charged with examining the language features of translated test items and deciding by consensus which features should be regarded as translation errors and on which translation error dimensions they should be coded.

Our analysis revealed that the two translation review panels were comparably effective, given the fact that the total number of translation errors identified by the two panels were very similar (132 and 126 errors, respectively, for Panel 1 and Panel 2 if the 86 errors detected by both panels are included in each panel's count; and 46 and 40 if those commonly detected errors are excluded in each panel's count). Of the 172 translation errors identified during the translation review sessions, only 86 (50%) were identified by both panels. Only a bit over half of the 86 errors identified by both panels were identified through the same discussion process.

As discussed in a previous related investigation (see Zhao and Solano-Flores, 2021), with proper facilitation, translation review panels with different cultural makeups may be comparably effective in detecting translation error. However, while equally effective, the panels should not be assumed to be exchangeable in terms of the errors they are able to detect. If, in a real-life situation, we had relied on the work of one review panel to review the translation of the items in an international comparison, at least 21 to 23% of translation errors would have gone undetected. Due to factors such as local culture, communication styles, and the uniqueness of the sets of skills of each reviewer, each panel is sensitive to a unique set of translation errors and different discussion processes lead translation review panels to detect those errors.

One limitation to the generalizability of our study stems from the fact that the translation review was conducted in the Chinese cultural contexts and the discussion processes vary tremendously across cultures and societies. Clearly, more studies are needed that examine how the lessons learned from this study are held across different source language-target language combinations. Yet evidence from related research (Zhao et al., 2018) indicates that, while each source language-target language combination poses a unique set of translation challenges, it is possible to develop a good understanding of the cultural and social factors that are relevant to properly implementing consensus-based test translation review.

TABLE 6 Percentages of translation errors detected only by Panel 1 by discussion process and translation error dimension (*n* = 46).

Discussion process	Construct	Conventions	Curriculum	Format	Grammar and syntax	Information	Origin	Register	Semantics	Style	Total
ID	8.7	4.35	6.52	4.35	10.87	4.35	0	0	19.57	2.17	60.88
ID-DI	2.17	0	0	0	0	0	0	0	6.52	0	8.69
ID-AG	6.52	0	0	0	4.35	0	0	0	6.52	2.17	19.56
ID-DI-AG	0	0	0	0	0	0	2.17	0	0	0	2.17
ID-EL	0	0	0	0	0	0	0	0	2.17	2.17	4.34
ID-DI-EL	0	0	0	0	0	0	0	0	0	0	0
ID-AG-EL	0	0	0	0	0	0	0	0	4.35	0	4.35
ID-DI-AG-EL	0	0	0	0	0	0	0	0	0	0	0
TOTAL	17.39	4.35	6.52	4.35	15.22	4.35	2.17	0	39.13	6.51	99.99

TABLE 7 Percentages of translation errors detected only by Panel 2 by discussion process and translation error dimension (*n* = 40 errors).

Discussion process	Construct	Conventions	Curriculum	Format	Grammar and syntax	Information	Origin	Register	Semantics	Style	Total
ID	0	0	2.5	2.5	5	5	0	0	22.5	0	37.5
ID-DI	0	0	0	0	0	0	0	0	0	0	0
ID-AG	10	0	2.5	5	5	5	5	0	7.5	5	45
ID-DI-AG	0	0	0	0	0	0	0	2.5	0	0	2.5
ID-EL	0	0	0	0	0	0	0	0	2.5	2.5	5
ID-DI-EL	0	0	0	0	0	0	0	0	0	0	0
ID-AG-EL	0	0	0	0	0	2.5	2.5	0	2.5	0	7.5
ID-DI-AG-EL	0	0	0	0	0	2.5	0	0	0	0	2.5
TOTAL	10	0	5	7.5	10	15	7.5	2.5	35	7.5	100

TABLE 8 Matrix of correspondence of discussion processes: percentage of errors detected by both panels through each combination of discussion processes (*n* = 86).

	Panel 2							
Panel 1	ID	ID-DI	ID-AG	ID-DI-AG	ID-EL	ID-DI-EL	ID-AG-EL	ID-DI-AG-EL
ID	33.72	1.16	26.74	1.16	0	0	3.48	0
ID-DI	1.16	0	2.32	0	0	0	0	0
ID-AG	5.81	1.16	10.46	0	0	0	8.13	0
ID-DI-AG	1.16	0	0	0	0	0	0	0
ID-EL	0	0	0	0	0	0	0	0
ID-DI-EL	1.16	0	0	0	0	0	0	0
ID-AG-EL	0	0	1.16	0	0	0	1.16	0
ID-DI-AG-EL	0	0	0	0	0	0	0	0

Another limitation of our study is a reflection of the complexities of conducting applied research in real-life contexts: The disparity across panels in the number of experts and areas of expertise, which could have produced various group dynamics. Fortunately, the fact that the two panels had comparable detection rates mitigates concerns about the external validity of the study. In addition, nothing in our findings suggests that the panels differed on the depth of their discussions—hence the importance of effective, adaptive facilitation. We can speculate that, while the panel with more individuals had a wider representation of various professional profiles, the panel with fewer individuals allowed more individual participation. In the end, potential differences in effectiveness due to unequal numbers of panelists may have canceled each other. The reality is that little is known about the ways in which the different professional backgrounds represented in panels influence the overall process of evaluation (Abma-Schouten et al., 2023). But even if the number of panel members is the same, there is a limit to which review panels can be expected to be comparable. Even if the specialty, professional interests, years of experience, etc., of the panel members are similar, multiple idiosyncratic factors and circumstances shape their discussions. While review panels have been a familiar part of the testing scene, their use in test translation review has been scant. We are just beginning to study their advantages and disadvantages systematically.

For now, our findings show that different translation review panels can be equally effective, but the discussion processes that lead them to detect errors tend to be different even for errors on the same dimensions. Next steps in future research on consensus-based translation review should focus on devising cost-effective approaches to staging multiple review panels focused on the same set of test items. Also, future research should examine how the social dynamics within translation review panels shape error detection when the source or the target language in test translation is different. For now, we can conclude that no single panel is likely to be able to detect all the possible translation errors in a given set of items.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by University of Colorado Boulder. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

XZ: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. GS-F: Methodology, Supervision, Writing – review & editing, Validation.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abma-Schouten, R., Gijbels, J., Reijmerink, W., and Meijer, I. (2023). Evaluation of research proposals by peer review panels: broader panels for broader assessments? *Sci. Public Policy* 50, 619–632. doi: 10.1093/scipol/scad009

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Appl. Meas. Educ.* 16, 55–73. doi: 10.1207/S15324818AME1601_3

Berman, A. I., Haertel, E. H., and Pellegrino, J. W. (2020). *Comparability of large-scale educational assessments: issues and recommendations*. Washington, DC: National Academy of Education

Bod, R., Hay, J., and Jannedy, S. (2003). *Probabilistic linguistics*. Cambridge, MA: MIT

Bray, M., Kobakhidze, M. N., and Suter, L. E. (2020). The challenges of measuring outside-school-time educational activities: experiences and lessons from the Programme for international student assessment (PISA). *Comp. Educ. Rev.* 64, 87–106. doi: 10.1086/706776

Burnaby, S., and Sun, Y. (1989). Chinese teachers' views of western language teaching: context informs paradigms. *TESOL Q.* 23, 219–238. doi: 10.2307/3587334

Camilli, G., and Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cook, L. I., and Schmitt-Cascallar, A. P. (2005). "Establishing score comparability for tests given in different languages" in *Adapting educational and psychological tests for cross-cultural assessment*. eds. R. K. Hambleton, P. F. Merenda and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers), 139–160.

Fink, A., Kosecoff, J., Chassin, M., and Brook, R. H. (1984). Consensus methods: characteristics and guidelines for use. *Am. J. Public Health* 74, 979–983. doi: 10.2105/ajph.74.9.979

Halliday, M.A.K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London, UK: Edward Arnold (Publishers), Ltd

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. *Eur. J. Psychol. Assess.* 10, 229–244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *Eur. J. Psychol. Assess.* 17, 164–172. doi: 10.1027/1015-5759.17.3.164

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures" in *Adapting educational and psychological tests for cross-cultural assessment*. eds. R. K. Hambleton, P. Merenda and C. Spielberger (Mahwah, NJ: Lawrence Erlbaum), 3–38.

Hambleton, R. K., Merenda, P., and Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence S. Erlbaum Publishers.

International Test Commission (2017). The ITC guidelines for translating and adapting tests. Available at: www.InTestCom.org (Accessed February 20, 2021)

Kamens, D. H., and McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comp. Educ. Rev.* 54, 5–25. doi: 10.1086/648471

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

Kūkea Shultz, P., and Englert, K. (2021). Cultural validity as foundational to assessment development: an indigenous example. *Front. Educ.* 6:701973. doi: 10.3389/feduc.2021.701973

OECD. (2015). *PISA test questions*. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa-test-questions.htm

Oliveri, M. E. (2019). Considerations for designing accessible educational scenario-based assessments for multiple populations: a focus on linguistic complexity. *Front. Educ.* 4:457932. doi: 10.3389/feduc.2019.00088

Ruiz-Primo, M. A., and Li, M. (2015). The relationship between item context characteristics and student performance: the case of the 2006 and 2009 PISA. Teachers College record, 117, 1–36. Available at: https://www. tcrecord.org (Accessed March 21, 2021).

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

Sireci, S. G., Yang, Y., Harter, J., and Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations: a methodological analysis of translation quality. *J. Cross-Cult. Psychol.* 37, 557–567. doi: 10.1177/002202210629047

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Appl. Meas. Educ.* 27, 236–247. doi: 10.1080/08957347.2014.944308

Solano-Flores, G. (2019). "The participation of Latin American countries in international assessments: assessment capacity, validity, and fairness" in *Sage handbook on comparative studies in education: Practices and experiences in student schooling and learning*. eds. L. E. Suter, E. Smith and B. D. Denman (Thousand Oaks, CA: Sage), 139–161.

Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. A. (2009). Theory of test translation error. *Int. J. Test.* 9, 78–91. doi: 10.1080/15305050902880835

Solano-Flores, G., Contreras-Niño, L. A., and Backhoff, E. (2013). "The measurement of translation error in PISA-2006 items: an application of the theory of test translation error" in *Research on the PISA research conference 2009*. eds. M. Prenzel, M. Kobarg, K. Schöps and S. Rönnebeck (Heidelberg: Springer Verlag), 71–85.

Strohschneider, S. (2002). Cultural factors in complex decision making. *Online Readings in Psychology and Culture* 4, 1–14. doi: 10.9707/2307-0919.1030

Suter, L. (2019). "Changes in the world-wide distribution of large-scale international assessments" in *Sage handbook on comparative studies in education: Practices and experiences in student schooling and learning*. eds. L. E. In, E. S. Suter and B. D. Denman (Thousand Oaks, CA: Sage), 553–568.

Teltemann, J., and Klieme, E. (2017). "The impact of international testing projects on policy and practice" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge), 369–386.

Trumbull, E., and Nelson-Barber, S. (2019). The ongoing quest for culturally-responsive assessment for indigenous students in the U.S. *Front. Educ.* 4:436758. doi: 10.3389/feduc.2019.00040

Weber, E. U., Ames, D. R., and Blais, A. (2005). 'How do I choose thee? Let me count the ways': a textual analysis of similarities and differences in modes of decision-making in China and the United States. *Manag. Organ. Rev.* 1, 87–118. doi: 10.1111/j.1740-8784.2004.00005.x

Zhao, X. (2018). Test translation review procedures in international large-scale assessment: Sensitivity to culture and society. Doctoral dissertation. University of Colorado Boulder.

Zhao, X., and Solano-Flores, G. (2021). Testing across languages in international comparisons: cultural adaptation of consensus-based test translation review procedures. *J. Multiling. Multicult. Dev.* 42, 677–691. doi: 10.1080/01434632.2020.1852242

Zhao, X., Solano-Flores, G., and Qian, M. (2018). International test comparisons: reviewing translation error in different source language-target language combinations. *Int. Multilingual Res. J.* 12, 17–27. doi: 10.1080/19313152.2017.1349527