TYPE Original Research
PUBLISHED 04 September 2025
DOI 10.3389/feduc.2025.1595658

Detection of cultural and linguistic differential item functioning in reading assessment

Yejin Woo and Youn-Jeng Choi*

Department of Education, Ewha Womans University, Seoul, Republic of Korea

OPEN ACCESS

EDITED BY
Ester Villalonga-Olives,
University of Maryland, United States

REVIEWED BY
Sebastian Weirich,
Institute for Educational Quality Improvement (IQB), Germany
Christopher Amissah,
University of Maryland, United States

*CORRESPONDENCE
Youn-Jeng Choi
✉ younjengchoi@ewha.ac.kr

RECEIVED 18 March 2025
ACCEPTED 14 August 2025
PUBLISHED 04 September 2025

CITATION
Woo Y and Choi Y-J (2025) Detection of cultural and linguistic differential item functioning in reading assessment.
Front. Educ. 10:1595658.
doi: 10.3389/feduc.2025.1595658

This study aims to determine whether differential item functioning (DIF) occurs in the PISA 2018 reading assessment and, if so, to explore which factors, such as linguistic elements, achievement goals, and perceived reading instructions as cultural elements, contribute most significantly to its occurrence. The United States was set as the reference group, and comparisons were made with Canada, Singapore, and South Korea. Item response theory-likelihood ratio (IRT-LR), logistic regression, and Rasch Tree analyses were utilized to identify DIF. Multiple methods consistently showed that item CR551Q06 exhibited DIF. The Rasch Tree analysis revealed that linguistic rather than cultural differences were the primary contributors to DIF. Interestingly, no DIF was detected using the Rasch Tree method in the comparisons between the United States and Canada and between the United States and Singapore, in contrast to the IRT-LR and logistic regression results. The analysis highlighted translation issues as a major source of bias, suggesting that careful adaptation of assessments is crucial to reducing DIF. These findings challenge assumptions about cultural differences in educational outcomes and emphasize the need for further research using varied DIF detection methods in different cultural contexts.

KEYWORDS
DIF, PISA, IRT-likelihood ratio, logistic regression, Rasch Tree

1 Introduction

Assessing academic achievement serves not only to measure an individual's academic abilities but also to provide a basis for offering appropriate educational support to learners. Therefore, academic achievement assessments must employ valid tools that accurately measure the intended construct. A valid assessment tool should be aligned with the purpose of the test and free from any bias towards or against specific groups. Tests that favor or disadvantage a particular group prevent making fair conclusions about the test-takers based on their scores. This bias can manifest either across the entire test or within individual items, which is referred to as differential item functioning (DIF). Scholars have long argued that DIF can occur based on variables such as gender, race, or social status (Coleman, 1968).

DIF occurs when test-takers with the same ability level perform differently on individual test items, which is closely related to test fairness and equity in large-scale assessments. The presence of DIF in an assessment can threaten its validity or lead to misinterpretation of item-level group differences. However, the presence of DIF does not necessarily undermine the validity of the entire test. Instead, it highlights how certain test items may function differently for various groups of test-takers. Thus, identifying and addressing DIF is crucial to ensuring fairness and equity in assessments.

Language differences are a key factor contributing to DIF. Test items have linguistic characteristics that make them more sensitive to translation errors compared to other types of texts (Solano-Flores et al., 2009). Moreover, translation issues are among the primary causes of DIF in international comparative studies (Yildirim and Berberoĝlu, 2009). Even when test items are translated properly, translation bias can still occur, as noted by Grisay and Monseur (2007). For example, translations may result in longer texts, altered word counts, or imprecise meaning, which can impact how test-takers comprehend the test. Consequently, even if the translated items ask the same questions as the source items, the test-takers' understanding and perceived difficulty of the items may differ.

Cultural differences can also lead to DIF. Culture, shaped through interactions with the environment, reflects regional characteristics (Jang, 2010). Eastern and Western cultures, for example, show significant differences: Eastern cultures tend to emphasize collectivism and community, while Western cultures prioritize individualism. These cultural distinctions can influence how test items are interpreted and processed by different groups, contributing to DIF. Qian and Lau (2022) identified achievement goals and perceived reading instruction as cultural variables that may significantly affect DIF between Eastern and Western students.

Achievement goal theory, largely based on Western literature, may function differently in East Asian contexts due to the competitive education systems and emphasis on achievement (Lau and Lee, 2008; Lau and Nie, 2008). East Asian students exhibit unique patterns, with a strong positive correlation between mastery and performance goals (Ho and Hau, 2008) and greater adoption of avoidance goals compared to Western students (Zusho and Clayton, 2011). Traditional reading instruction in East Asia, characterized by teacher-centered, competitive, and exam-focused methods (Watkins and Biggs, 2001), contrasts with mastery-oriented Western practices. However, recent reforms in some societies with Confucian heritage culture (CHC) have influenced traditional pedagogical practices (Lau and Ho, 2015). In particular, the reforms in China promote student autonomy and cooperation, aligning more with mastery-oriented approaches (The Ministry of Education and People's Republic of China, 2011). Examining how these aspects have changed after the educational reforms could provide meaningful insights.

Based on the cultural differences between East and West, Qian and Lau (2022) utilized PISA 2018 reading achievement data to explore the impact of achievement goals and perceived reading instruction on the academic performance of Chinese students. The study showed that achievement goals and perceived reading instruction, particularly disciplinary climate, adaptive instruction, and teacher stimulation, influenced Chinese students' reading performance. However, while the study examined variables likely to differ between Eastern and Western cultures, it focused solely on China, an Eastern country, and did not address potential differences in reading ability between students from Eastern and Western countries.

This study aimed to address this limitation in the literature by focusing on how linguistic and cultural factors contribute to DIF. It specifically examined whether cultural variables, such as achievement goals and perceived reading instruction, significantly impact differences in reading ability between students from Eastern and Western countries. It employed multiple DIF detection methods, including item response theory-likelihood ratio (IRT-LR), logistic regression (LR), and Rasch Tree (RT).

DIF detection techniques are grounded in two primary theories: Item response theory (IRT) and classical test theory. These techniques vary in their algorithms, synchronization criteria, and the cutoff points used to identify DIF. However, DIF detection methods are not entirely consistent with one another (Bakan Kalaycıoğlu and Berberoğlu, 2010). In response, simultaneously applying multiple DIF detection methods is often recommended (Hambleton, 2006).

Traditional DIF detection methods include the Mantel–Haenszel method, SIBTEST, and logistic regression, which are based on classical test theory, while methods such as the likelihood ratio test, Lord's method, and Raju's method are based on IRT. These traditional methods are statistically intuitive, relatively simple to interpret, and have been validated for reliability through several studies and practical evaluations. In this study, the likelihood ratio test was selected for its foundation in IRT (Camilli and Shepard, 1994), and logistic regression was chosen for its ability to detect both uniform and non-uniform DIF (Zumbo, 1999).

Although traditional DIF detection methods have these strengths, they often fall short in accounting for the diverse subgroups within the test-taker population because they require predefined distinctions between focal and reference groups. Several innovative approaches have been proposed to overcome this limitation. For example, the IRTree model (Böckenholt, 2012) models the test-taker's response process as a multi-step decision-making process, represented by a tree structure. Additionally, the Rasch Tree method (Strobl et al., 2015) combines logistic regression with recursive partitioning to form subgroups based on various characteristics and response patterns, while other methods, such as the item-focused tree model (Tutz and Berger, 2016) and the mixture IRT model combining latent class and IRT models, have also been introduced.

Among these, the Rasch Tree method has the advantage of using all explanatory variables in the data to define possible subgroups through recursive partitioning without the need for researchers to set arbitrary threshold values for group definitions (Jang and Lee, 2023). For this reason, the present study employed the Rasch Tree method for analysis.

This study utilized data from the PISA 2018 student questionnaire and reading assessment. The United States was designated as the reference country, with the following countries included for comparison: (1) Canada, which shares the similar written language and culture as the U.S.; (2) Singapore, which shares the similar written language but differs culturally; and (3) South Korea, which differs from the U.S. in both written language and culture. The objective is to investigate the presence of DIF within the PISA 2018 reading assessment across these countries. Using IRT-LR, logistic regression, and Rasch Tree methods, the study aims to determine whether DIF occurs and, if so, identify common DIF items and examine their characteristics. To further explore the potential causes of DIF, IRT-LR and logistic regression are used to infer the indirect influences of linguistic and cultural factors based on predefined country groupings. Building on this, the Rasch Tree method is applied to identify specific factors that directly contribute to DIF without relying on predefined group structures. The specific research questions are as follows:

1. Does DIF occur in the PISA 2018 Reading assessment when utilizing IRT-likelihood ratio (IRT-LR), logistic regression, and Rasch Tree methods across the four different countries?
2. If DIF occurs, what are the DIF items, and what characteristics do they exhibit?
3. If DIF occurs, what cultural and linguistic factors influence the DIF in the PISA 2018 Reading assessment, as identified by the Rasch Tree method?

2 Theoretical frameworks

2.1 Test translation error and its impact

The theory of test translation error was developed to tackle the difficulties of accurately assessing diverse populations across multiple languages (Solano-Flores et al., 2009). This theory characterizes translation error as the perceived discrepancies in content, structure, and meaning between the original and translated versions of test items. Accordingly, translation errors are not solely caused by poor-quality translations; even when translators perform exceptionally well, such errors remain. These errors occur because languages represent cultural experiences in distinct ways (Greenfield, 1997) and utilize different semiotic systems (Bezemer and Kress, 2008). For instance, certain languages employ classifiers with nouns modified by numbers (Aikhenvald, 2003), with different classifiers required depending on the type of noun. These classifiers convey specific characteristics (such as shape or quantity) about the noun, leading to potential differences in the information provided by translated items compared to their source material. Similarly, translating quantifiers (like "some" or "any") can be challenging when dealing with non-Indo-European languages because they do not utilize these elements in the same way as English or French (Grisay, 2007).

The theory suggests that translation involves not just the translator's work but also various factors in both the translation process and the development of assessment tools. These factors can influence aspects such as content, vocabulary frequency, and grammatical complexity in the translated tests. Translation encompasses numerous features that may vary between language versions of the same test items, ranging from visual layout to the content volume, cognitive demands, linguistic requirements, and the cultural context assumed for test-takers. For example, even a slight change in the scale of a graph can make a curve appear steeper than in the original, and contextual details meant to make an item relatable may refer to scenarios that are unfamiliar to students in the target language. Although such errors are not the direct fault of the translator, they significantly influence the translated test's overall characteristics.

The theory posits that translation errors can be categorized based on various aspects, ranging from the design and format of the items to their linguistic features and the nature of the content. A key idea in the theory is that translation errors are multidimensional (Solano-Flores et al., 2009). For instance, a punctuation mistake might not only be a style error but also affect the meaning, making it a semantic error. Similarly, word-for-word translations and the use of syntactic structures uncommon in the target language fall under grammar and syntax errors. Depending on the content being assessed, word-for-word translations may also cause errors in semantics or construct dimensions, as they can modify the intended meaning or alter the content being evaluated.

The concept of multidimensionality is further linked to the idea of a trade-off between error dimensions: correcting or minimizing an error in one dimension may inadvertently introduce errors in another. For example, avoiding the use of classifiers to prevent additional information about an object's characteristics (as mentioned in the previous example) may result in increased grammatical complexity.

The theory suggests that translation errors are unavoidable because languages encode cultural experiences differently (Greenfield, 1997) and rely on distinct semiotic systems (Bezemer and Kress, 2008). Additionally, the trade-offs between the error dimensions mentioned above make the complete elimination of translation errors impossible, even with high-quality translations. For instance, the amount of space required for printed text differs across languages, leading to variations in how much space the text occupies on a page and how much blank space is available for students to write their responses.

Although some translation errors are inevitable, many of them are insignificant—they may go unnoticed or are unlikely to affect the constructs being measured or the difficulty of the items. However, studies (Solano-Flores et al., 2005, 2013) have demonstrated a stronger correlation between translation errors and item difficulty in cases where numerous or severe errors occur—those that are likely to change the constructs or meanings of the items—compared to items with minimal or minor translation errors.

Among studies investigating cross-cultural measurement invariance in PISA assessments, Oliden and Lizaso (2013) reported that the PISA 2009 reading skills test displayed metric invariance across different languages. However, Söyler Bağdu (2020) found that the PISA 2015 reading skills test did not exhibit measurement invariance between native English-speaking and non-native English-speaking countries.

Ceyhan (2019) investigated the measurement invariance of the PISA 2012 reading skills test, focusing on comparisons involving the same language with different cultures, as well as different languages with different cultures. The study showed that structural invariance was achieved for comparisons using the same language, while only weak invariance was observed for different languages. Analyses of the PISA 2000 reading skills items showed fewer items exhibiting DIF when comparing countries using the same language, as opposed to within-country comparisons using different languages (Grisay et al., 2009; Grisay and Monseur, 2007). This suggests that language—in other words, translation—plays a critical role in DIF. Similarly, in the present study, certain items were found to display DIF when comparing different languages and cultures.

2.2 Cultural difference and its impact

Culture is shaped through interactions with the environment, reflecting regional characteristics (Jang, 2010). As a result, Eastern and Western cultures that developed during the same historical period exhibit significant differences.

One key example is creativity, which is influenced by the cultural context in which individuals are situated. According to Sung (2006), culture and creativity are inseparably linked, with individualism and liberalism in Western cultures contrasting with collectivism and Confucianism in Eastern cultures, leading to differences in the development of creative traits and environments. Research suggests that Western individuals tend to excel in divergent and creative thinking, while Eastern individuals are generally more reflective and intuitive in their approaches. Furthermore, Eastern cultures often take a holistic view, perceiving objects as interconnected entities before analyzing their structure. In contrast, Western cultures typically adopt a more analytical worldview, focusing on individual elements first and then synthesizing them into a whole. These cultural distinctions align with collectivism and community-oriented values in the East and individualism in the West, shaped by their respective regional environments.

In the realm of education, these cultural differences are also reflected in distinct educational practices. Eastern education often emphasizes a strong foundation in cultural heritage, focusing on the acquisition and deep understanding of historical knowledge. In contrast, Western education prioritizes independent thinking, encouraging students to cultivate a broad range of skills through practice and experience, with an emphasis on learning how to approach problems critically and develop individual thought processes (Ko, 2013).

This divergence is particularly pronounced in language education. For instance, in South Korea, the teaching of the Korean language is imbued with a nationalistic ideology that connects individual development with national progress. This connection is reflected in the curriculum, where grammar and literature are combined into a single comprehensive subject. By contrast, in the United States, literature is treated as one component of reading instruction, indicating a different pedagogical focus (Lee, 2013). These differences highlight how culture shapes educational goals and practices, which in turn influence students' learning experiences and outcomes.

From an educational achievement perspective, achievement goal theory, which originates from Western literature, operates differently in East Asian societies due to cultural influences (Lau and Lee, 2008; Lau and Nie, 2008). Studies have revealed unique patterns in the achievement goals of students from CHCs, where education systems are highly competitive, and achievement holds significant value. For example, East Asian students show a strong positive correlation between mastery and performance goals (Ho and Hau, 2008). Moreover, performance goals may contribute positively to adaptive learning and academic achievement (Salili and Lai, 2003). Another notable difference is that East Asian students tend to adopt higher levels of avoidance goals compared to their Western counterparts (Zusho and Clayton, 2011).

The nature of reading instruction in East Asian societies also warrants attention, as it differs markedly from Western instructional practices. Traditional reading classrooms in East Asia, shaped by CHC, are typically teacher-centered, characterized by large class sizes, a competitive atmosphere, directive teaching methods, and an emphasis on examination performance (Watkins and Biggs, 2001). These instructional methods often seem at odds with the mastery-oriented and student-centered approaches advocated in Western educational systems, which aim to foster adaptive achievement goals.

However, global educational reforms have altered some traditional teaching methods in these societies (Lau and Ho, 2015). To examine whether these reforms have influenced reading instruction and achievement goals—factors shaped by cultural contexts—this study investigates DIF based on CHC. Using the United States as the reference country, the analysis compares traditionally CHC-influenced countries, such as South Korea and Singapore, with Canada, which, like the U.S., does not have a CHC background. This approach indirectly evaluates CHC's impact on educational environments and outcomes, particularly in reading instruction and achievement goals (Qian and Lau, 2022).

3 Materials and methods

3.1 Sample

The Programme for International Student Assessment (PISA), conducted by the Organization for Economic Co-operation and Development (OECD), was selected as the research focus. Since 2000, PISA has been administered every three years to assess the mathematics, science, and reading skills of 15-year-old students receiving formal education. Additionally, PISA collects data on various educational variables through student, teacher, school, and parent questionnaires. As a large-scale international comparative study, PISA serves as a valuable tool for countries to evaluate their educational systems and environments and to inform improvements in their education policies and practices.

While the most recent PISA cycle took place in 2021, this study utilizes data from the 2018 PISA cycle, where reading was designated as the core domain. In PISA, the core domain is assessed in greater detail, comprising approximately half of the total testing time (OECD, 2019c). One core domain is assessed for all students, while the other domains are treated as minor and are not administered to all students (OECD, 2019c). This emphasis provides more comprehensive coverage and a larger sample size, enabling deeper analysis and more reliable insights into student performance across countries (OECD, 2019c). Thus, the 2018 PISA reading assessment data were selected for this study.

Given that this study aimed to examine potential bias in the reading items of PISA 2018, the analysis was conducted using published items. Specifically, the research focused on responses to seven items from the unit coded as "Rapa Nui", which were uniquely both publicly released and implemented in the PISA 2018 reading main survey simultaneously (OECD, 2019a). Table 1 presents the distribution of the items by code, item type, cognitive process being measured, text source, text organization and navigation, text format, text type, and item difficulty. As shown in Table 1, these items measure a wide range of cognitive characteristics.

The analysis included data from the United States as the reference group, with Canada (sharing the similar written language and culture), Singapore (sharing the similar written language but a different culture), and South Korea (with a different language and culture) as comparison groups. The United States was selected as the reference group due to the widespread use of English, the most commonly spoken language, and its frequent designation as a target country in prior studies (Khorramdel et al., 2020; Muench et al., 2022; Sachse et al., 2016). Canada shares Western cultural traits with the United States, with both English and French as the languages of assessment. Canada was selected as a comparison country due to its cultural and linguistic similarities with the United States. To ensure consistency in the language of assessment, only the sample of students who took the test in English was included in the analysis. Singapore, while sharing the same language of assessment (English), represents a distinctly different Eastern cultural context. South Korea, as an East Asian country, represents a different cultural context and uses Korean as the primary language, which is linguistically distinct from English. This selection of comparison groups is expected to provide clearer insights into the effects of linguistic and cultural factors on DIF.

Participants with any missing values were excluded, resulting in the removal of approximately 15% of cases from the cognitive item data. For the Rasch Tree analysis, explanatory variables were categorized into linguistic and cultural constructs and merged with the data based on student IDs (see Table 2). Specifically, the variable "Language" was used as the linguistic explanatory variable, while "Perceived Reading Instruction" and "Achievement Goals" were classified as cultural explanatory variables. All cases with missing values in these explanatory variables were removed using listwise deletion (missing rate; United States: 3.7%, Canada: 6.6%, Singapore:

1.8%, South Korea: 2.0%). In addition, for item “CR551Q06,” scores coded as “2” were recoded to “1,” while partial scores coded as “1” were recoded to “0.” The sample sizes from each country were as follows: United States (reference group) with 678 students, Canada with 2,898 students, Singapore with 1,222 students, and South Korea with 1,197 students.

Because model fit indices can be influenced by sample size (Hu and Bentler, 1995; Fan et al., 1999; Lei and Lomax, 2005; Fan and Sivo, 2007; Mahler, 2011), the study aimed to use an equal number of students from each country. Therefore, because the United States had the smallest sample size (678), the same number of students was randomly selected from the other

TABLE 1 Item characteristics of “Rapa Nui” unit.

Item code	Item type	Cognitive process subscale	Cognitive process	Source	Text organization and navigation	Text format	Text type	Difficulty
CR551Q01	Simple multiple choice	Locate information	Access and retrieve information within a text	Single	Dynamic	Continuous	Narration	Level 4
CR551Q05	Open Response	Understand	Represent literal meaning	Single	Dynamic	Continuous	Narration	Level 3
CR551Q06	Complex multiple choice	Evaluate and reflect	Reflect on content and form	Single	Static	Continuous	Argument	Level 5
CR551Q08	Simple multiple choice	Locate information	Access and retrieve information within a text	Single	Static	Continuous	Argument	Level 5
CR551Q09	Simple multiple choice	Evaluate and reflect	Detect and handle conflict	Multiple	Multiple	Continuous	Multiple	Level 4
CR551Q10	Complex multiple choice	Understand	Integrate and generate inferences across multiple sources	Multiple	Multiple	Continuous	Multiple	Level 5
CR551Q11	Open Response	Evaluate and reflect	Detect and handle conflict	Multiple	Multiple	Continuous	Multiple	Level 4

TABLE 2 Explanatory variables for Rasch Tree (OECD, 2019b).

Aspect	Variable	Explanation
Language: language of assessment	LANGTEST_COG	Language of assessment is the language utilized during the actual administration of the test. English was encoded as “313,” French as “493,” and Korean as “301,” which are categorical variable.
Culture: perceived reading instruction	DISCLIMA	The disciplinary climate in the test language classroom is assessed through five items on a four-point Likert scale with the categories “Every lesson,” “Most lessons,” “Some lessons,” and “Never or hardly ever.”
	TEACHSUP	Teacher support is measured through four items on a four-point Likert scale with the categories “Every lesson,” “Most lessons,” “Some lessons,” and “Never or hardly ever.”
	DIRINS	Teacher-directed instruction is evaluated using four reverse-coded items on a four-point Likert scale with the categories “Every lesson,” “Most lessons,” “Some lessons,” and “Never or hardly ever.”
	PERFEED	Perceived teacher feedback is assessed through three items on a four-point Likert scale with the categories “Never or almost never,” “Some lessons,” “Many lessons,” and “Every lesson or almost every lesson.”
	STIMREAD	Teacher stimulation of reading and teaching strategies is measured through four items on a four-point Likert scale with the categories “Never or hardly ever,” “In some lessons,” “In most lessons,” and “In all lessons.”
	ADAPTIVITY	Instruction adaptivity in test language lessons is evaluated through three items with a four-point Likert scale with the categories “Never or almost never,” “Some lessons,” “Many lessons,” and “Every lesson or almost every lesson.”
Culture: achievement goals	COMPETE	Competitiveness is assessed through three items on a four-point Likert scale with the categories “Strongly disagree,” “Disagree,” “Agree,” and “Strongly agree.”
	WORKMAST	Working motive and mastery achievement motive are measured with three items on a four-point Likert scale with the categories “Strongly disagree,” “Disagree,” “Agree,” and “Strongly agree.”
	GFOFAIL	General fear of failure is evaluated using three items on a four-point Likert scale with the categories “Strongly disagree,” “Disagree,” “Agree,” and “Strongly agree.”

countries. Consequently, data from a total of 2,712 students were analyzed.

For the Rasch Tree analysis, when comparing Canada to the United States, the variables “TEACHSUP,” “DIRINS,” and “PERFEED” were excluded, as these items were not administered in Canada. Additionally, the variable “language of assessment” was processed using one-hot encoding for comparisons involving different languages, such as Korea vs. the U.S. The variable “LANGTEST_COG” was dummy-coded to create new variables indicating the use of English (LANGTEST_COGEnglish), and Korean (LANGTEST_COGKorean). For example, when comparing Korea and the U.S., since the test languages are only Korean and English, the value for LANGTEST_COGEnglish is coded as 1, and LANGTEST_COGKorean is coded as 0 for those who took the test in English. However, for the United States–Canada and United States–Singapore comparisons, since all participants took the test in the same language (English), the variable was not included as it does not carry meaningful variance for analysis.

3.2 Statistical analysis and interpretation

Using the complete dataset, IRT-LR was employed with the IRTLRDIF program (Thissen, 2001) using the 3-parameter IRT model, and we used the “difR” package for logistic regression and the “psychotree” package for the Rasch Tree analyses in R.

IRT-LR, based on IRT, compares two different models: a compact model, which assumes no DIF, and an augmented model, which assumes DIF is possible in the item under study. For the IRT-LR analysis using the IRTLRDIF program, the likelihood ratio test statistic, G^2, was employed. The formula for G^2, as outlined by Thissen (2001), is as follows:

$$G^2\left(df\right) = -2\log L_c - \left(-2\log L_A\right)$$

where $df = p$, p represents the number of parameters. L_c is the compact model, and L_A is the augmented model. $G^2\left(df\right)$ follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the augmented and compact models (Thissen, 2001). Since the three-parameter model was applied, DIF is considered present when G^2 exceeds 7.81, with 3 degrees of freedom (df) (Choi et al., 2015). The formula for the three-parameter model is as follows:

$$P_i\left(\theta\right) = c_i + \left(1 - c_i\right) \cdot \frac{1}{1 + \mathrm{e}^{-\alpha_i\left(\theta - \beta_i\right)}}$$

The second DIF technique used in this research was logistic regression (LR). LR assesses the effect of multiple independent variables on a binary outcome, determining which of two categories a subject belongs to. The logistic regression models can be described as follows:

Model 1 (full model): logit(p) = $\tau_0 + \tau_1\Lambda + \tau_2\Gamma + \tau_3\left(\Lambda + \Gamma\right)$

Model 2 (1st reduced model): logit(p) = $\tau_0 + \tau_1\Lambda + \tau_2\Gamma$

Model 3 (2nd reduced model): logit(p) = $\tau_0 + \tau_1\Lambda$

where p represents the probability of answering the item correctly, Λ is a measure of an individual’s ability (e.g., IRT ability parameters (θ) or total scores), Γ is a categorical predictor variable indicating group membership for an individual (where $\Gamma = 1$ for members of the focal group and $\Gamma = 0$ for members of the reference group), $\left(\Lambda + \Gamma\right)$ is the interaction of a person’s ability and their group membership.

The term τ_1 represents the main effect of a person’s ability on their performance on the item. τ_2 reflects the difference in intercepts between the focal and reference groups, which indicates uniform DIF when statistically significant. That is, uniform DIF occurs when one group consistently performs better or worse than the other across all levels of ability. τ_3 captures the interaction between ability and group membership (i.e., whether the relationship between ability and item performance differs by group), and a significant τ_3 implies the presence of non-uniform DIF. While τ_2 and τ_3 are sometimes interpreted as approximating group-specific differences in intercepts ($\tau_2 = \beta_{0F} - \beta_{0R}$) and slopes ($\tau_3 = \beta_{1F} - \beta_{1R}$), these are regression-based estimates and should not be directly equated with IRT-based item parameters. If the null hypothesis H_0: $\tau_3 = 0$ (comparison between Model 1 and Model 2) is rejected, this suggests a significant interaction between ability and group, indicating non-uniform DIF; in this case, the DIF testing procedure concludes. If H_0: $\tau_3 = 0$ is not rejected, the subsequent comparison between Model 2 and Model 3 tests H_0: $\tau_2 = 0$, and a significant result indicates uniform DIF (Swaminathan and Rogers, 1990). This stepwise approach enables a clear distinction between the two types of DIF.

The statistical testing of LR DIF analysis was conducted by analyzing the difference in model fit between the two nested models using the chi-square statistic (x^2) (Scott et al., 2010; Sohn, 2010). The LRT (Likelihood Ratio Test) statistics evaluate DIF by comparing the fit of two nested models, while the Wald statistics assess model parameters using an appropriate contrast matrix (Johnson and Wichern, 1998). Since LRT statistics is the default setting in the R package (Magis et al., 2010), it was utilized in this study.

LR DIF analysis can also be viewed as a weighted least squares approach. The contributions of explanatory variables are reflected in the change in the coefficient of determination (R^2) between the augmented and compact models. This change is computed as:

$$\Delta R^2 = R_1^2 - R_2^2$$

where R_1^2 represents the value for the augmented model and R_2^2 represents the value for the compact model. The difference in R^2 illustrates the additional explanatory power provided by the variables in the augmented model. LR was selected for its ability to detect both uniform and non-uniform DIF (Zumbo, 1999), making it a robust method.

The DIF level was decided according to the ΔR^2 value obtained from employing the LR technique. According to Jodoin and Gierl (2001), the ΔR^2 value is interpreted as follows: $0 < \Delta R^2 < 0.035$, no or negligible DIF; $0.035 \leq \Delta R^2 < 0.07$, moderate DIF; $\Delta R^2 \geq 0.07$, high DIF. According to another source, $\Delta R^2 < 0.13$ indicates no or negligible DIF; $0.13 \leq \Delta R^2 < 0.26$, and $\Delta R^2 \geq 0.26$ indicate moderate and high DIF, respectively (Zumbo and Thomas, 1996). In this study, Jodoin and Gierl’s (2001) criterion was to determine DIF. Items with a level of effect classified as B or higher in the logistic regression results were identified as exhibiting DIF.

To account for the increased risk of Type I errors due to multiple hypothesis testings, a Bonferroni correction was applied. This method

TABLE 3 Average reading achievement among United States, Canada, Singapore, and South Korea (OECD, 2019a).

Countries	United States	Canada	Singapore	South Korea
Average reading score	505	520	549	514
Rank among 81 countries (excluding Spain)	13	6	2	9
Average reading score among OECD countries	487			

adjusts the significance threshold by dividing the desired alpha level by the number of hypotheses conducted (Holland and Thayer, 1986). In this study, seven items were analyzed, resulting in an adjusted significance level of 0.05/7 = 0.007 for each item. This correction was uniformly applied to both the IRT-LR and LR analyses to ensure consistency across methods. In the context of IRT-LR, the critical value for detecting DIF was determined as 12.11 based on the chi-square distribution with 3 degrees of freedom at the adjusted alpha level. For LR, which is based on a chi-square distribution with 1 degree of freedom, the corresponding critical value was approximately 7.27.

Nevertheless, Type I error can occur during the DIF detection process and may have significant impact on the analysis. However, according to the study by Atar and Kamata (2011), the simulation conditions considered three sample sizes (600, 1,200, 2,400) and two group sample size ratios (1:1 and 1:2). Regarding Type I error control, their findings indicated that the Type I error rates of both LRT (Likelihood Ratio Test) and LRP (Logistic Regression Procedure) were well controlled at clearly defined significance levels across all simulation conditions. However, previous studies have reported that when there is a difference in ability between groups, Type I error tends to increase (DeMars, 2009; Li et al., 2012; Narayanan and Swaminathan, 1996). Nevertheless, as shown in Table 3, all four countries belong to the highest or high-performing group in PISA reading. This suggests that the ability differences between groups are not substantial, indicating that Type I error is relatively well controlled in this study. Furthermore, using ΔR^2 along with the chi-square test is advantageous for controlling Type I error (Jodoin and Gierl, 2001). Therefore, this study employed both ΔR^2 and the chi-square test to analyze DIF to reduce the likelihood of Type I error.

Finally, the Rasch Tree method (Strobl et al., 2015) integrates logistic regression with recursive partitioning to identify subgroups based on response patterns and explanatory variables. This approach allows for the data-driven formation of subgroups without relying on arbitrary thresholds set by researchers (Jang and Lee, 2023), making it a suitable method for DIF analysis. However, this method has certain limitations, particularly in terms of interpretative complexity due to its data-driven nature. Unlike traditional DIF detection methods that rely on predefined groups, Rasch Tree iteratively identifies subgroups based on statistical splits, which may lead to challenges in result interpretation and theoretical alignment (Strobl et al., 2015). To address these limitations, this study incorporates traditional DIF detection methods, specifically the IRT-LR and LR, to enhance the robustness and interpretability of the analysis.

For interpreting the Rasch Tree results, item difficulty estimates were used. Higher item difficulty values indicate more difficult items, and groups with higher item difficulty estimates are considered disadvantaged in relation to those specific items. The Rasch Tree analysis examines differences in item difficulty parameters between groups under the null hypothesis that there is no difference in item difficulty parameters (Camilli and Shepard, 1994). The threshold for defining DIF based on item difficulty was calculated using the Mantel–Haenszel (MH) effect size from the Rasch Tree model (Henninger et al., 2023; Holland and Thayer, 1986; Roussos et al., 1999). In this case, ΔMH was used as an indicator using the MH odds ratio to evaluate whether items function differentially between groups.

TABLE 4 ETS classification scheme for the Mantel–Haenszel odds ratio in the $|\Delta \mathrm{MH}|$.

Class	Interpretation	Classification rule
A	Negligible DIF	$0 \le \lvert\Delta MH\rvert \le 1$
B	Medium (moderate) DIF	$1 < \lvert\Delta MH\rvert < 1.5$
C	Large DIF	$\lvert\Delta MH\rvert \ge 1.5$

$$\Delta MH = -2.35 \times \left(b_{iR} - b_{iF}\right)$$

b_{iR} = difficulty parameter for item i in the reference group; b_{iF} = difficulty parameter for item i in the focal group.

Item difficulty estimates for each subgroup from the Rasch Tree results were used to calculate ΔMH, and items were classified as A, B, or C based on the ETS classification system for the Mantel–Haenszel effect size (Henninger et al., 2023). Items classified as A, B, and C indicate no, medium, and large DIF, respectively. The stopping criterion for Rasch Tree analysis occurs when all items are classified as A or at least one item is classified as B. Specifically, if the absolute difference in item difficulty between groups exceeds 0.426(=1/2.35) but is less than 0.638(=1.5/2.35), the item is classified as B, indicating moderate DIF. If the absolute difference exceeds 0.638, the item is classified as C, indicating a large DIF effect. The item difficulty estimates for each comparison and the corresponding Mantel–Haenszel effect size classifications are presented in Table 4. Additionally, items with a level of effect classified as B or higher in the Rasch Tree results were identified as exhibiting DIF.

4 Results

4.1 DIF analysis with IRT-LR and LR

As shown in Table 5, the comparison between the United States and Canada revealed that item CR551Q06 exhibited a DIF effect based on the IRT-LR and LR techniques. DIF had to be identified by at least one of the DIF detection methods for an item to be classified as a DIF item. In the case of the LR analyses, only items with at least a B level of effect were considered as exhibiting DIF. Item CR551Q06 showed negligible DIF based on LR techniques but were identified as

TABLE 5 A comparison between the USA, Canada, Singapore, and South Korea using IRT-LR & logistic regression.

Item code	Canada vs. USA						Singapore vs. USA						South Korea vs. USA					
	IRT-LR			Logistic regression			IRT-LR			Logistic regression			IRT-LR			Logistic regression		
	G^2	DIF presence	Favored group	LRT χ^2	Level of effect	Favored group	G^2	DIF presence	Favored group	LRT χ^2	Level of effect	Favored group	G^2	DIF presence	Favored group	LRT χ^2	Level of effect	Favored group
CR551Q01	10.4	X	-	10.84*	A	Canada	1.7	X	-	5.07	-	-	55.0	O	Korea	8.82	A	-
CR551Q05	2.9	X	-	2.17	-	-	4.0	X	-	2.11	-	-	26.0	O	Korea	119.37*	C	Korea
CR551Q06	21.5	O	USA	23.13*	A	USA	77.1	O	USA	105.82*	B	USA	466.5	O	USA	616.56*	C	USA
CR551Q08	0.1	X	-	0.78	-	-	8.5	X	-	9.77*	A	USA	46.8	O	USA	8.49	A	-
CR551Q09	0.0	X	-	0.60	-	-	8.8	X	-	25.49*	A	Singapore	14.6	O	USA	11.61	A	-
CR551Q10	0.5	X	-	0.03	-	-	0.0	X	-	13.92*	A	Singapore	20.6	O	Korea**	9.93*	A	Korea
CR551Q11	0.1	X	-	0.47	-	-	4.7	X	-	2.66	-	-	18.0	O	USA	18.89*	A	Korea

*IRT-LR: DIF is present if G^2 exceeds 12.11(df = 3) (Bonferroni correction) (Choi et al., 2015), Logistic regression: "A": negligible DIF (ΔR^2 < 0.035), "B": moderate DIF ($0.035 \leq \Delta R^2 < .07$), "C": high DIF ($0.07 \leq \Delta R^2$) (Jodoin and Gierl, 2001). Values marked with * represent the Likelihood Ratio Test (LRT) chi-square values for uniform DIF, while unmarked values indicate non-uniform DIF. The IRTLRDIF program does not differentiate between uniform and nonuniform DIF; therefore, this study does not report them separately. Values marked with ** correspond to items for which the discrimination parameters were considerably higher in Korean group than in the U.S. group, according to the DIF analysis.

displaying DIF in the IRT-LR analysis. Item CR551Q06 favored the United States in both the IRT-LR and LR analyses.

In the comparison between the United States and Singapore, item CR551Q06 showed a DIF effect. Item CR551Q06 was detected as having DIF based on the IRT-LR technique and a moderate effect according to the LR technique. Conversely, items CR551Q08, CR551Q09, and CR551Q10 displayed negligible DIF based on the LR method and were not classified as DIF items according to the IRT-LR analysis. Item CR551Q06 favored the United States in both the IRT-LR and LR analyses and was classified as exhibiting uniform DIF in the LR analysis.

In the comparison between the United States and South Korea, all items were confirmed to exhibit a DIF effect. IRT-LR analysis classified all items as exhibiting DIF. LR analysis identified that items CR551Q05 and CR551Q06 demonstrated a C level of effect, indicating high DIF. In contrast, CR551Q01, CR551Q08, CR551Q09, CR551Q10, and CR551Q11 showed an A level of effect, suggesting negligible DIF. LR analysis identified CR551Q01, CR551Q08, and CR551Q09 as exhibiting non-uniform DIF, while CR551Q05, CR551Q06, CR551Q10, and CR551Q11 exhibited uniform DIF. According to the IRT-LR analysis, items CR551Q01 and CR551Q05 were found to favor Korea, while items CR551Q06, CR551Q08, CR551Q09, and CR551Q011 favored the United States. For item CR551Q010, the difficulty parameter (b) reported by the IRT-LR was 0.98 for both groups, making it difficult to determine which group the item favored. Based on LR analysis, among the items showing uniform DIF, items CR551Q05, CR551Q010, and CR551Q011 favored Korea, while item CR551Q06 favored the United States. In summary, both IRT-LR and LR analyses consistently indicated that item CR551Q05 favored Korea and item CR551Q06 favored the United States. Therefore, all items showed DIF when comparing these two countries with distinct written languages and cultures.

Based on the comparisons between the United States, Canada, Singapore, and South Korea, item CR551Q06 consistently displayed DIF across all comparisons. Additionally, item CR551Q06 showed moderate to high DIF (at least level B for LR analysis) in both the United States-Singapore and United States-South Korea. Furthermore, item CR551Q06 was consistently classified as a uniform DIF item favoring the United States across all three comparisons: United States–Canada, United States–Singapore, and United States–South Korea. In contrast, item CR551Q05 was not identified as exhibiting DIF in either the IRT-LR or LR analyses for the United States–Canada and United States–Singapore comparisons. However, in the comparison with South Korea, it was detected as a DIF item in both methods. Notably, the LR analysis indicated a C level of effect, suggesting high DIF, and both IRT-LR and LR analyses consistently classified it as a uniform DIF item favoring Korea. The release of the Rapa Nui unit highlights the need to investigate which specific components of these items contribute to DIF.

Upon reviewing the characteristics of the items, no significant common features were immediately apparent. However, when the items were analyzed in terms of the cognitive processes required to solve them in each language, notable differences emerged. Specifically, item CR551Q06 requires "reflecting on content and form," a relatively complex cognitive process. In contrast, item CR551Q05 involves "representing literal meaning," which is considered a lower-level cognitive process. These findings highlight the need for a thorough review to determine whether differences in required cognitive processes contribute to the occurrence of DIF, and whether the importance or difficulty of these cognitive skills varies across the countries examined.

Based on the frequency and severity of DIF identified, it appears most significant in South Korea then Singapore, followed by Canada. This pattern corresponds to the degree of linguistic and cultural differences from the United States, which was used as the reference country. These findings suggest that linguistic and cultural disparities are positively associated with DIF.

4.2 DIF analysis with Rasch Tree

The results of the DIF analysis for the seven items from the Rapa Nui unit of the PISA 2018 Reading Assessment, represented in a tree structure using Rasch Tree analysis, are shown in Figure 1. All linguistic and cultural variables listed in Table 2 were incorporated into the Rasch Tree model as candidate splitting variables. However, for the United States–Canada and United States–Singapore comparisons, language-related variables were excluded from the model because all participants took the test in English, and thus the language of assessment lacked variability, while cultural variables were retained.

The nodes at the bottom of the figure represent the partitions within the decision tree model, each corresponding to a specific subgroup in the data. Initially, a Rasch Tree analysis was conducted using the academic performance data of students from the United States and Canada. All cultural variables listed in Table 2 were included in the Rasch Tree model as candidate splitting variables. However, no splits were observed, which is an expected outcome given the cultural similarities between the two countries. Moreover, this finding aligns with the IRT-LR and LR results, in which only item CR551Q06 was identified as exhibiting DIF.

Next, a Rasch Tree analysis was conducted for the United States and Singapore. The results showed no branches split, indicating the absence of significant DIF items and any meaningful influence from cultural factors. Although Singapore is culturally distinct from the United States, the lack of subgroup splits in the Rasch Tree analysis suggests that cultural differences are less likely to influence DIF in the context of international test translation.

Subsequently, we analyzed the reading assessment data of students from the United States and South Korea, which was the only comparison that resulted in a node split in the Rasch Tree analysis. Two subgroups were formed based on the language of assessment, as indicated by the "LANGTEST_COGEnglish" dummy variable: those using English and those using Korean. The left subgroup (node 2), where the "LANGTEST_COGEnglish" variable has a value of less than or equal to 0, represents the group assessed in Korean, while the right subgroup (node 3), where the variable has a value greater than 0, represents the group assessed in English. This reflects the fact that, while the language of assessment in the United States is exclusively English, in South Korea, the test is administered in Korean. Even though the subgroups were not predefined by country, the analysis revealed a clear division between the United States and South Korea, indicating a particularly strong DIF effect when comparing these two countries. Furthermore, despite including various cultural variables as background factors, the analysis confirmed that linguistic differences, rather than cultural differences, contribute to the presence of DIF in these items.

As shown in Table 6, the effect size of the test was analyzed using ΔMH, and item CR551Q09 was classified as level B, indicating moderate DIF, while items CR551Q01, CR551Q05, CR551Q06, and CR551Q08 were classified as level C, indicating a large DIF. In particular, item CR551Q06 had an absolute ΔMH value of 7.8, and items CR551Q01 and CR551Q05 had absolute values of 3.3 ~ 3.8, showing a significant difference that led to their classification as DIF items. The fact that such items with large DIF values were identified in the subgroups divided by Korean and English suggests that a thorough review of the translation process from English to Korean is necessary. In this case, items CR551Q01, CR551Q05, CR551Q09, CR551Q10, and CR551Q11 favored the group that took the test in Korean (node 2) compared to the group that took the test in English (node 3). Conversely, items CR551Q06 and CR551Q08 favored the group that took the test in English (node 3) compared to the group that took the test in Korean (node 2).

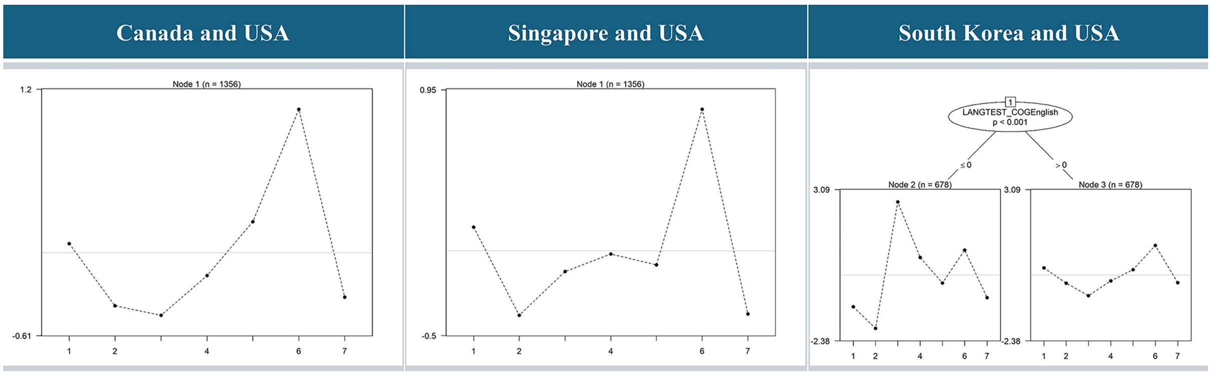


FIGURE 1
Rasch Tree by comparison between the USA, Canada, Singapore, and South Korea. For the United States–Canada and United States–Singapore comparisons, the "language of assessment" variable was excluded due to the lack of variance, as all participants completed the assessment in English. Accordingly, only the two cultural variables—Perceived Reading Instruction and Achievement Goals—were included in these analyses, as shown in Table 2. In contrast, the United States–South Korea comparison included all linguistic and cultural variables listed in Table 2.

TABLE 6 A comparison between the USA, Canada, Singapore, and South Korea using Rasch Tree.

Comparison	Canada and USA	Singapore and USA	South Korea and USA					
Item code	Item difficulty estimates	Item difficulty estimates	Item difficulty estimates					
	Node 1	Node 1	Node 2	Node 3	Favored Group	Difference	$\lvert\Delta MH\rvert$	Level of effect
CR551Q01	0.1	0.1	−1.1	0.3	**Korea**	**1.4**	**3.3**	**C**
CR551Q05	−0.4	−0.4	−1.9	−0.3	**Korea**	**1.6**	**3.8**	**C**
CR551Q06	−0.5	−0.1	2.6	−0.7	**USA**	**3.3**	**7.8**	**C**
CR551Q08	−0.2	0.0	0.6	−0.2	**USA**	**0.8**	**1.9**	**C**
CR551Q09	0.2	−0.1	−0.3	0.2	**Korea**	**0.5**	**1.2**	**B**
CR551Q10	1.1	0.8	0.9	1.1	Korea	0.2	0.5	A
CR551Q11	−0.3	−0.4	−0.8	−0.3	**Korea**	**0.5**	**1.2**	**B**

Items classified as B or C level, based on differences in item difficulty estimates between subgroups, are highlighted in bold. Rasch Tree: "A": negligible DIF ($0 \le \lvert \Delta MH \rvert \le 1$), "B": medium DIF ($1 < \lvert \Delta MH \rvert < 1.5$), "C": large DIF ($\lvert \Delta MH \rvert \ge 1.5$) (Henninger et al., 2023). For the United States–Canada and United States–Singapore comparisons, the "language of assessment" variable was excluded due to the lack of variance, as all participants completed the assessment in English. Accordingly, only the two cultural variables—Perceived Reading Instruction and Achievement Goals—were included in these analyses, as shown in Table 2. In contrast, the United States–South Korea comparison included all linguistic and cultural variables listed in Table 2.

In addition to the DIF results shown in Table 6, we observed a notable pattern in the item difficulty estimates that warranted further exploration. While the main Rasch Tree analyses, as presented in Table 6, did not include the country variable "CNT" and focused solely on linguistic and cultural covariates, no node splits were observed for the United States–Canada and United States–Singapore comparisons. As a result, item difficulty estimates in both cases were derived from the entire combined sample without country-level differentiation. Interestingly, the resulting item difficulty estimates from these two comparisons were similar. To supplement these findings and better understand the basis of this similarity, we conducted an additional Rasch Tree analysis that included the country variable "CNT" as a covariate. When the CNT variable was included, both the U.S.–Canada and U.S.–Singapore comparisons resulted in a binary split based solely on country, allowing for the estimation of item difficulties separately for each group. Although these estimates were not used for the main analysis, they are presented in Supplementary Table 1 to support a more comprehensive understanding of group-level patterns.

4.3 Comparative analysis of IRT-LR, LR, and Rasch Tree results

The results of comparing and analyzing the outcomes of the IRT-LR, LR, and Rasch Tree analyses are as follows.

First, the comparison between the United States and Canada yielded divergent results across the three DIF detection methods. While both IRT-LR and LR identified item CR551Q06 as exhibiting DIF, the Rasch Tree analysis revealed no node splits, indicating no evidence of DIF. Given the strong cultural and linguistic alignment between the two countries, it is plausible that this discrepancy arises from the nature of subgroup specification. Traditional approaches like IRT-LR and LR rely on predefined groups—typically based on nationality—which may amplify even minor differences. In contrast, the Rasch Tree method identifies subgroups through a data-driven process without imposing prior group definitions. These findings imply that the DIF observed in item CR551Q06 may reflect the analytical structure of the traditional methods rather than genuine linguistic or cultural sources. Further investigation is warranted to clarify the origins of this DIF.

In addition, the supplementary analysis presented in Supplementary Table 1—conducted by including the country variable "CNT" in the Rasch Tree model—produced results consistent with those in Table 5, again identifying item CR551Q06 as exhibiting DIF. This finding confirms that when DIF is analyzed based on predefined country groups such as the United States and Canada, the same item tends to be detected as DIF regardless of the analytical method employed.

Next, a similar discrepancy was observed in the United States–Singapore comparison. Item CR551Q06 was again identified as a DIF item by IRT-LR and LR, whereas the Rasch Tree analysis showed no evidence of subgroup splits. Unlike Canada, Singapore has a markedly different cultural background, despite sharing English as the test language. Although differences between Eastern and Western cultures exist, these differences are unlikely to be substantial enough to induce DIF. Traditional methods may have captured differences based on predefined national groups rather than cultural factors, such as perceived reading instruction and achievement goals, which were specifically examined in this study. This pattern reinforces the notion that DIF detection may be sensitive to the method's reliance on pre-established group structures.

In addition, the supplementary analysis presented in Supplementary Table 1—conducted by including the country variable "CNT" in the Rasch Tree model—yielded results consistent with those in Table 5, identifying items CR551Q06, CR551Q09, and CR551Q10 as exhibiting DIF. This consistency suggests that when DIF is analyzed based on predefined national groups such as the United States and Singapore, the same items tend to be flagged as DIF regardless of the analytical method employed.

Lastly, the comparison between the United States and South Korea shows partially consistent results. While all items were classified as DIF items using IRT-LR and LR, the Rasch Tree method identified CR551Q01, CR551Q05, CR551Q06, CR551Q08, CR551Q09, and CR551Q11 as DIF items. In the case of the Rasch Tree analysis, only items with at least a B level of effect were considered as exhibiting DIF. In the comparison between the United States and South Korea, all items except CR551Q10 were consistently identified as DIF items in at least two of the three analyses.

An analysis of the characteristics of the six items revealed the following: three items were simple multiple-choice, two was open-response, and one was complex multiple-choice item. In terms of cognitive processes, two items assessed students' ability to access and retrieve information within a text, two items focused on detecting and handling conflict, one item assessed representing literal meaning, and one item assessed reflecting on content and form. Regarding the cognitive process subscale, three items focused on evaluating and reflecting, two items on locating information, and one on understanding. For text organization and navigation, two items were classified as dynamic, two as static, and two as multiple. All items had a continuous text format. The text types included two narrative items, two argumentative items, and two multiple type items. In terms of item difficulty, three items were classified as level 4, two as level 5, and one as level 3 according to the PISA proficiency scale, which ranges from below level 1 to level 6. Based on this classification, the items primarily represent high difficulty levels, with level 3 indicating moderate difficulty and levels 4 and 5 representing more challenging tasks. Although the six items were classified as DIF items in the analyses comparing the United States and South Korea, no distinct commonalities were revealed after analyzing their characteristics.

At least two of the three methods consistently identified item CR551Q06 (Figure 2) as DIF item when comparing the United States with two or more countries. All analyses indicated that item CR551Q06 favored the United States. In the comparison with South Korea, both items CR551Q05 and CR551Q06 exhibited large DIF effect sizes across all methods, with item 5 consistently favoring Korea (Figures 2, 3). CR551Q06 is a complex multiple-choice item that targets the cognitive process of reflecting on content and form, categorized under the "evaluate and reflect" subscale. It relies on a single text source, features static organization and navigation, uses a continuous text format, and is of the argumentative type with a difficulty level of 5, which indicates a high level of complexity. Additionally, CR551Q05 is an open-response item that targets the cognitive process of representing literal meaning, categorized under the "understand" subscale. It relies on a single text source, features dynamic organization and navigation, uses a continuous text format, and is of the narrative type. The item has a difficulty level of 3, which indicates a moderate level of complexity. Further investigation, including a detailed review of the released items, is necessary to understand why this item consistently exhibited DIF across all analyses, and particularly in the United States–South Korea comparison. Such analysis could help minimize DIF in future updates of international academic assessments.

Overall, there are notable similarities between the Rasch Tree results and those obtained using the IRT-LR and LR methods. Specifically, when considering both the presence and effect size of DIF in the IRT-LR and LR analyses, the frequency of DIF follows the order: South Korea > Singapore > Canada. This finding suggests that greater linguistic and cultural differences between countries are associated with a higher likelihood of DIF.

The Rasch Tree results further confirmed that among linguistic and cultural variables, the only significant factor influencing DIF occurrence was linguistic differences. This finding challenges the preconceived notion that perceived reading instruction and achievement goals differ significantly between Eastern and Western cultures. It may also imply that recent educational reforms in CHC cultures have substantially impacted educational culture, leading to outcomes that resemble those in the West.

5 Conclusion

5.1 Discussion

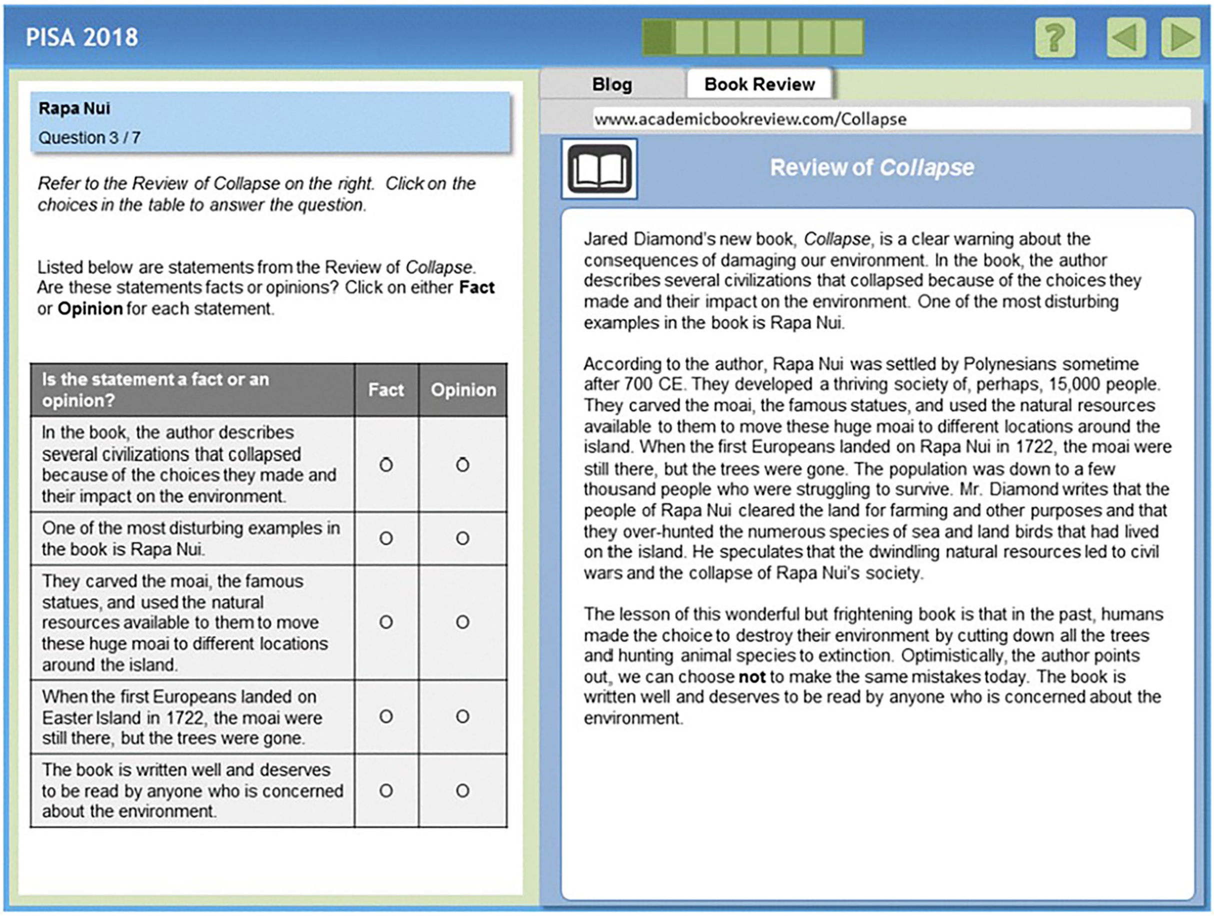
This study applied traditional DIF detection methods, including IRT-LR and LR, as well as the newly emerging Rasch Tree method, to explore DIF and analyze its contributing factors. The analysis focused on the Rapa Nui Unit, consisting of seven items from the reading domain of PISA 2018. The reference group in the analysis was the United States, while Canada, Singapore, and South Korea were comparison groups.

This study applied traditional DIF detection methods, IRT-LR and LR, alongside the Rasch Tree method, to explore DIF across comparisons between the United States and Canada, Singapore, and South Korea. In the comparison between the United States and Canada, item CR551Q06 was identified as DIF item in the IRT-LR and LR analyses. However, the Rasch Tree analysis showed that no node splits, suggesting that the absence of substantial cultural differences between the two countries contributes to the low likelihood of DIF occurrence.

In the comparison between the United States and Singapore, a notable discrepancy was observed across methods. While IRT-LR and LR identified item CR551Q06 as exhibiting DIF, the Rasch Tree method found no significant subgroup splits or DIF items. The absence of node splits—despite the inclusion of cultural background variables—suggests that cultural differences alone are unlikely to lead to the occurrence of DIF.

This absence of node splits in the Rasch Tree analysis highlights the potential dependence of traditional methods on predefined subgroups. As supplementary evidence, when the Rasch Tree analysis was conducted with the country variable—representing predefined national groups—explicitly included, node splits emerged in both the U.S.–Canada and U.S.–Singapore comparisons. The resulting DIF patterns closely aligned with those identified by the IRT-LR and LR analyses. These findings suggest that when DIF is analyzed based on predefined country groups, the same items are flagged for DIF across methods.

When comparing the United States and South Korea, all items under consideration were classified as DIF by IRT-LR and LR, while the Rasch Tree method identified six items—CR551Q01, CR551Q05, CR551Q06, CR551Q08, CR551Q09, and CR551Q11—as exhibiting DIF. This was the only comparison where the Rasch Tree analysis



Rapa Nui
Question 3 / 7

Refer to the Review of Collapse on the right. Click on the choices in the table to answer the question.

Listed below are statements from the Review of *Collapse*. Are these statements facts or opinions? Click on either **Fact** or **Opinion** for each statement.

Is the statement a fact or an opinion?	Fact	Opinion
In the book, the author describes several civilizations that collapsed because of the choices they made and their impact on the environment.	○	○
One of the most disturbing examples in the book is Rapa Nui.	○	○
They carved the moai, the famous statues, and used the natural resources available to them to move these huge moai to different locations around the island.	○	○
When the first Europeans landed on Easter Island in 1722, the moai were still there, but the trees were gone.	○	○
The book is written well and deserves to be read by anyone who is concerned about the environment.	○	○

Review of *Collapse*

Jared Diamond's new book, *Collapse*, is a clear warning about the consequences of damaging our environment. In the book, the author describes several civilizations that collapsed because of the choices they made and their impact on the environment. One of the most disturbing examples in the book is Rapa Nui.

According to the author, Rapa Nui was settled by Polynesians sometime after 700 CE. They developed a thriving society of, perhaps, 15,000 people. They carved the moai, the famous statues, and used the natural resources available to them to move these huge moai to different locations around the island. When the first Europeans landed on Rapa Nui in 1722, the moai were still there, but the trees were gone. The population was down to a few thousand people who were struggling to survive. Mr. Diamond writes that the people of Rapa Nui cleared the land for farming and other purposes and that they over-hunted the numerous species of sea and land birds that had lived on the island. He speculates that the dwindling natural resources led to civil wars and the collapse of Rapa Nui's society.

The lesson of this wonderful but frightening book is that in the past, humans made the choice to destroy their environment by cutting down all the trees and hunting animal species to extinction. Optimistically, the author points out, we can choose **not** to make the same mistakes today. The book is written well and deserves to be read by anyone who is concerned about the environment.

FIGURE 2
Released item "CR551Q06" from Rapa Nui unit (Reproduced from OECD, 2019a, © OECD 2019).

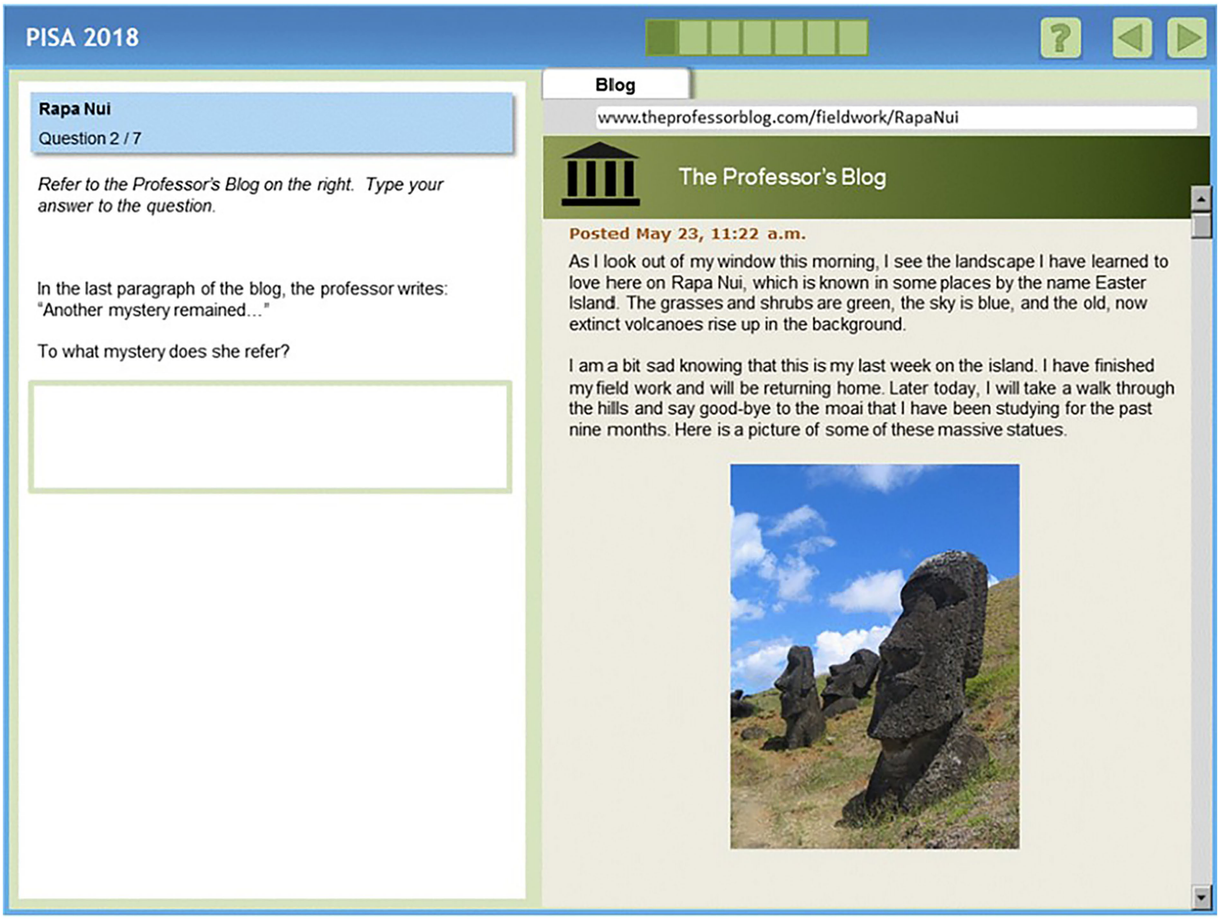
revealed node splits, corresponding to differences in the language of assessment. These findings indicate that linguistic factors had a substantial impact on the observed DIF, partially consistent with the results from traditional methods.

Given the significant role of language in inducing DIF, greater attention should be paid to the accurate and culturally sensitive translation of test items into each country's target language. Translation effects have also been identified as a major source of bias, as prior studies have demonstrated their significant impact on DIF (Grisay and Monseur, 2007; Grisay et al., 2009; Oliden and Lizaso, 2013; Solano-Flores et al., 2005, 2013). These findings underscore the importance of careful and systematic adaptation procedures in cross-cultural assessments to reduce DIF. Potential sources of item bias must be considered during the item writing and adaptation phases. If necessary, specialized training for item writers and translators should be provided for this purpose. Based on the findings of this study, future research should be conducted using data from diverse cultural and linguistic contexts, employing multiple DIF detection techniques to validate and extend the current results.

These results also challenge the commonly held assumption that substantial differences in reading instruction and achievement goals exist between Eastern and Western cultures. They suggest that recent educational reforms in East Asia may have aligned instructional practices more closely with Western standards.

In addition, certain items repeatedly exhibited DIF across multiple comparisons. In particular, item CR551Q06 consistently showed DIF, with all analyses indicating that it favored the United States. In the comparison with South Korea, both items CR551Q05 and CR551Q06 demonstrated large DIF effect sizes across all detection methods, with CR551Q05 consistently favoring Korea. The consistent advantage of CR551Q06 for U.S. students across all comparisons warrants further investigation to verify and understand the underlying cause. Furthermore, in the U.S.–South Korea comparison, the emergence of a relatively large number of DIF items and the magnitude of their effect sizes call for careful review and interpretation.

Moreover, future DIF research should incorporate both methods that require predefined groups and those that do not. In this study, IRT-LR and LR—which necessitate prior group definitions—were used alongside the Rasch Tree method, which operates without such assumptions. When IRT-LR and LR were applied using country-based groupings, DIF was detected in the comparisons between the United States and Canada, as well as between the United States and Singapore. In contrast, using the same dataset without country-based grouping, the Rasch Tree method did not identify any DIF. This contrast suggests that predefined groupings by country may lead to attributing DIF to prominent factors like language or culture, even though the true source of DIF may stem from other factors. Therefore, future research should incorporate methods like the Rasch Tree to



Rapa Nui
Question 2 / 7

Refer to the Professor's Blog on the right. Type your answer to the question.

In the last paragraph of the blog, the professor writes: "Another mystery remained…"

To what mystery does she refer?

Posted May 23, 11:22 a.m.

As I look out of my window this morning, I see the landscape I have learned to love here on Rapa Nui, which is known in some places by the name Easter Island. The grasses and shrubs are green, the sky is blue, and the old, now extinct volcanoes rise up in the background.

I am a bit sad knowing that this is my last week on the island. I have finished my field work and will be returning home. Later today, I will take a walk through the hills and say good-bye to the moai that I have been studying for the past nine months. Here is a picture of some of these massive statues.

FIGURE 3
Released item "CR551Q05" from Rapa Nui unit (Reproduced from OECD, 2019a, © OECD 2019).

more accurately uncover the underlying mechanisms that contribute to DIF, beyond predefined group structures.

Beyond methodological considerations, it is also important to reflect on the broader educational implications of how DIF is handled in international assessments. A related consideration concerns the inherent tension between cultural neutrality and task engagement. In efforts to develop culturally comparable assessments, item developers may be inclined to minimize or eliminate cultural references from tasks. However, doing so risks producing tasks that are overly generic, potentially diminishing students' motivation and engagement. Tasks that completely avoid cultural context may fail to reflect authentic language use or meaningful scenarios, which are essential for assessing higher-order reading skills. This presents a dilemma for international assessment developers, as it may be more practical to accept a manageable level of DIF than to sacrifice the relevance and richness of the tasks. Acknowledging this trade-off is important when designing items that aim to be both culturally fair and pedagogically valuable.

Taken together, the findings from this study not only offer methodological insights but also highlight important practical considerations for international assessments. Finally, the present study contributes valuable insights into the potential cultural and linguistic sources of DIF, particularly by identifying items that may systematically favor or disadvantage specific groups. These insights can inform item development processes for future assessments. Specifically, the findings suggest that certain item features—such as cultural references and linguistic complexity—may introduce differential functioning that undermines cross-national comparability. Hence, test developers should carefully consider such factors during item construction, translation, and adaptation phases to ensure greater fairness and construct validity.

5.2 Future research and limitations

Future research should explore how linguistic and cultural factors influence the occurrence of DIF items in the PISA 2022 Reading Core stage in comparison to the 2018 assessment. This study selected the 2018 data based on the fact that the core domain in PISA 2018 was reading. However, as linguistic and cultural influences evolve, it is crucial to examine which specific factors contribute to DIF detection in the modern context. Investigating these temporal changes will provide valuable insights into the shifting dynamics of language and culture and their impact on item functioning, helping to refine future assessments and reduce potential biases. However, because an individualized test application was used in PISA 2018, no students took identical tests, allowing for analysis across different items. For this reason, DIF and item bias studies should also be conducted across different item clusters.

While this study focused on a single reading unit (Rapa Nui) due to the constraints of PISA's multiple matrix sampling design, this narrow scope may limit the generalizability of the findings. In practice, large-scale assessments such as PISA aim to capture broad constructs using diverse item sets across multiple units. As such, results derived from a single unit should be interpreted with caution, particularly regarding their applicability to the entire reading construct. To address this limitation and enhance the practical utility of DIF analyses, future research should replicate this study's approach across a wider range of units and domains. Doing so will help validate the observed patterns and provide more robust evidence for improving the fairness and interpretability of international large-scale assessments.

While the Rasch Tree analysis found no cultural variables influencing DIF, this may be due to the exclusion of practical cultural factors as explanatory variables. This study did include all available background variables; rather, it focused on only achievement goals and perceived reading instruction, which have been reported to differ significantly between Eastern and Western cultures and are believed to influence reading achievement (Qian and Lau, 2022). However, other cultural factors could also significantly impact DIF occurrence. Therefore, further research is needed to identify and incorporate additional practical cultural factors as explanatory variables to better assess their impact on DIF detection.

Finally, further analysis is needed for DIF items commonly identified across the comparisons of the United States, Canada, Singapore, and South Korea using IRT-LR, LR, and Rasch Tree analyses. In this study, DIF items were briefly analyzed based on item characteristics provided by the OECD. Beyond these basic characteristics, a detailed analysis of the released items is required. If the primary factor influencing DIF, as suggested by the Rasch Tree analysis, is linguistic, it is essential to assess whether the translation process for each country's language was appropriate and how the items were actually translated. Furthermore, consideration must be given to the characteristics of items that may introduce bias during translation.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.oecd.org/en/data/datasets/pisa-2022-database.html#data.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

YW: Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. Y-JC: Conceptualization, Data curation, Methodology, Project administration, Software, Supervision, Validation, Writing – review & editing.

Funding

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

Acknowledgments

This paper is an expanded version of the poster presented at the International Meeting of the Psychometric Society 2024.

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.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The authors acknowledge the use of GPT-4o (OpenAI API, 2025 version) for language refinement and manuscript editing. The AI was not utilized for generating original scientific content, conducting data analysis, or formulating research interpretations. All AI-assisted edits were meticulously reviewed and manually validated by the authors to ensure factual accuracy, coherence, and adherence to ethical standards. Any necessary modifications were made accordingly. The initial and final prompts used in AI-assisted editing have been included in the supplementary materials for full transparency.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1595658/full#supplementary-material

References

Aikhenvald, A. (2003). Classifiers: A typology of noun categorization devices. New York, NY: Oxford University Press.

Atar, B., and Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *H. U. J. Educ.* 41, 36–47.

Bakan Kalaycıoğlu, D., and Berberoğlu, G. (2010). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *J. Psychoeduc. Assess.* 20, 1–12. doi: 10.1177/0734282910391623

Bezemer, J., and Kress, G. (2008). Writing in multimodal texts: a social semiotic account of designs for learning. *Written Commun.* 25, 166–195. doi: 10.1177/0741088307313177

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychol. Methods* 17, 665–678. doi: 10.1037/a0028111

Camilli, G., and Shepard, L. A. (1994). Methods for identifying biased test items, vol. *4*: Sage.

Ceyhan, E. (2019). Assessing measurement invariance of PISA 2012 reading literacy scale among the countries determined in accordance with the language of application (master's thesis). Antalya: Akdeniz University.

Choi, Y.-J., Alexeev, N., and Cohen, A. S. (2015). DIF analysis using a mixture 3PL model with a covariate on the TIMSS 2007 mathematics test. *Int. J. Test.* 15, 239–253. doi: 10.1080/15305058.2015.1007241

Coleman, J. S. (1968). Equality of educational opportunity. *Equity Excell. Educ.* 6, 19–28. doi: 10.1080/0020486680060504

DeMars, C. E. (2009). Modification of the mantel-haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *J. Educ. Behav. Stat.* 34, 149–170. doi: 10.3102/1076998608329515

Fan, X., and Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivar. Behav. Res.* 42, 509–529. doi: 10.1080/00273170701382864

Fan, X., Thompson, B., and Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Struct. Equ. Modeling* 6, 56–83. doi: 10.1080/10705519909540119

Greenfield, P. M. (1997). You can't take it with you: why ability assessments don't cross cultures. *Am. Psychol.* 52, 1115–1124. doi: 10.1037/0003-066X.52.10.1115

Grisay, A. (2007). The challenge of adapting PISA materials into non Indo-European languages: Some evidence from a brief of exploration of language issues in Chinese and Arabic. Available online at: http://www.aspe.ulg.ac.be/grisay/fichiers/PISA07.pdf

Grisay, A., Gonzalez, E., and Monseur, C. (2009). "Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments," in *IERI monograph series: Issues and methodologies in large-scale assessments*. eds. F. Scheuermann and J. Björnsson (Hamburg, Germany: IEA-ETS Research Institute), 2, 63–83.

Grisay, A., and Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Stud. Educ. Eval.* 33, 69–86. doi: 10.1016/j.stueduc.2007.01.006

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Med. Care* 44, S182–S188. doi: 10.1097/01.mlr.0000245443.86671.c4

Henninger, M., Debelak, R., and Strobl, C. (2023). A new stopping criterion for Rasch trees based on the mantel-Haenszel effect size measure for differential item functioning. *Educ. Psychol. Meas.* 83, 181–212. doi: 10.1177/00131644221077135

Ho, I. T., and Hau, K. T. (2008). Academic achievement in the Chinese context: the role of goals, strategies, and effort. *Int. J. Psychol.* 43, 892–897. doi: 10.1080/00207590701836323

Holland, P. W., and Thayer, D. T. (1986). Differential item functioning and the mantel-Haenszel procedure. *ETS Res. Rep. Series* 1986, i–24. doi: 10.1002/j.2330-8516.1986.tb00186.x

Hu, L., and Bentler, P. (1995). "Evaluating model fit" in Structural equation modeling: concepts, issues and application. ed. R. Hoyle (Thousand Oasks: Sage Publications), 76–99.

Jang, K. B. (2010). A guideline for contents and method of cultural education based on the definition and characteristics of culture. *Korean J. Cult. Arts Educ. Stud.* 5, 19–37. doi: 10.15815/kjcaes.2010.5.2.19

Jang, Y. S., and Lee, J. Y. (2023). Exploring differential item functioning in PISA 2015 science test with the Rasch-tree. *Korean Soc. Educ. Eval.* 36, 83–110. doi: 10.31158/JEEV.2023.36.1.83

Jodoin, M. G., and Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Appl. Meas. Educ.* 14, 329–349. doi: 10.1207/S15324818AME1404_2

Johnson, R. A., and Wichern, D. W. (1998). Applied multivariate statistical analysis (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Khorramdel, L., Pokropek, A., Joo, S.-H., Kirsch, I., and Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: a partial invariance approach. *Psychol. Test Assess. Model.* 62, 179–231.

Ko, Y. C. (2013). *A study on cultural differences between the East and the West (Master's thesis)*. Chungju, South Korea: Korea National University of Transportation, Graduate School of Humanities.

Lau, K. L., and Ho, E. S. C. (2015). Reading performance and self-regulated learning of Hong Kong students: what we learnt from PISA 2009. *Asia-Pac. Educ. Res.* 25, 159–171. doi: 10.1007/s40299-015-0246-1

Lau, K. L., and Lee, J. (2008). Examining Hong Kong students' achievement goals and their relations with students' perceived classroom environment and strategy use. *Educ. Psychol.* 28, 357–372. doi: 10.1080/01443410701612008

Lau, S., and Nie, Y. (2008). Interplay between personal goals and classroom goal structures in predicting student outcomes: a multilevel analysis of person-context interactions. *J. Educ. Psychol.* 100, 15–29. doi: 10.1037/0022-0663.100.1.15

Lee, K. K. (2013). A comparative study on national language curricula of Korea, China, and USA. *Han-Geul* 300, 183–212. doi: 10.22557/HG.2016.06.300.183

Lei, M., and Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Struct. Equ. Model.* 12, 1–27. doi: 10.1207/s15328007sem1201_1

Li, Y., Brooks, G. P., and Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educ. Psychol. Meas.* 72, 847–861. doi: 10.1177/0013164411426157

Magis, D., Beland, S., Tuerlinckx, F., and De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behav. Res. Methods* 42, 847–862. doi: 10.3758/BRM.42.3.847

Mahler, C. (2011). The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling (Doctoral dissertation). Vancouver, Canada: The University of British Columbia.

Muench, R., Wieczorek, O., and Gerl, R. (2022). Education regime and creativity: the eastern Confucian and the Western enlightenment types of learning in the PISA test. *Cogent Educ.* 9:2144025. doi: 10.1080/2331186X.2022.2144025

Narayanan, P., and Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Appl. Psychol. Meas.* 20, 257–274. doi: 10.1177/014662169602000306

OECD (2019a). PISA 2018 Results (Volume I): What Students Know and Can Do, PISA. Paris: OECD Publishing. doi: 10.1787/5f07c754-en

OECD. (2019b). PISA 2018 technical report. Available online at: https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2019c). PISA 2018 assessment and analytical framework. Paris, France: OECD Publishing. doi: 10.1787/b25efab8-en

Oliden, P. E., and Lizaso, J. M. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema* 25, 390–395. doi: 10.7334/psicothema2013.46

Qian, Q., and Lau, K.-l. (2022). The effects of achievement goals and perceived reading instruction on Chinese student reading performance: evidence from PISA 2018. *J. Res. Read.* 45, 137–156. doi: 10.1111/1467-9817.12388

Roussos, L. A., Schnipke, D. L., and Pashley, P. J. (1999). A generalized formula for the mantel-Haenszel differential item functioning parameter. *J. Educ. Behav. Stat.* 24, 293–322. doi: 10.3102/10769986024003293

Sachse, K. A., Roppelt, A., and Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *J. Educ. Meas.* 53, 152–171. doi: 10.1111/jedm.12106

Salili, F., and Lai, M. K. (2003). Learning and motivation of Chinese students in Hong Kong: a longitudinal study of contextual influences on students' achievement orientation and performance. *Psychol. Schs.* 40, 51–70. doi: 10.1002/pits.10069

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual. Life Outcomes* 8:81. doi: 10.1186/1477-7525-8-81

Sohn, W. (2010). Exploring potential sources of DIF for PISA 2006 mathematics literacy items: application of logistic regression analysis. *J. Educ. Eval.* 23, 371–390.

Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. Á. (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. (2005).The Mexican translation of TIMSS-95: test translation lessons from a post-mortem study. Paper presented at the annual meeting of the National Council on measurement in education, Montreal, Quebec, Canada.

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 PISA. eds. M. Prenzel, M. Kobarg, K. Schöps and S. Rönnebeck (Dordrecht, The Netherlands: Springer), 71–85.

Söyler Bağdu, P. (2020). Investigation of the measurement variability of PISA 2015 reading skills test according to the language variability (master's thesis). İzmir: Ege University, Institute of Educational Sciences.

Strobl, C., Kopf, J., and Zeileis, A. (2015). Rasch trees: a new method for detecting differential item functioning in the Rasch model. *Psychometrika* 80, 289–316. doi: 10.1007/s11336-013-9388-3

Sung, E. H. (2006). Creativity in the Est and the west. *J. Korean Soc. Gift. Talent.* 5, 67–93.

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *J. Educ. Meas.* 27, 361–370. doi: 10.1111/j.1745-3984.1990.tb00754.x

The Ministry of Education and People's Republic of China (2011). Curriculum standards of basic Chinese language education. Beijing, China: People's Education Press.

Thissen, D. (2001). IRTLRDIF v2.0b—Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [computer software documentation]. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Tutz, G., and Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika* 81, 727–750. doi: 10.1007/s11336-015-9488-3

Watkins, D. A., and Biggs, J. B. (2001). Teaching the Chinese learner: Psychological and pedagogical perspectives. Hong Kong: Hong Kong University Press.

Yildirim, H. H., and Berberoĝlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *Int. J. Test.* 9, 108–121. doi: 10.1080/15305050902880736

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework forbinary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., and Thomas, D. R. (1996). A measure of DIF effect size using logistic regression procedures. Philadelphia, PA: Paper presented at National Board of Medical Examiners.

Zusho, A., and Clayton, K. (2011). Culturalizing achievement goal theory and research. *Educ. Psychol.* 46, 239–260. doi: 10.1080/00461520.2011.614526