ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(4);202-215

Cross-Lingual and Cross-Cultural Validity of the TIMSS 2019 Computer Use Questionnaire: Using Measurement Invariance and DIF Methods*

Zeynep Neveser KIZILÇİM** Sevda ÇETİN***

Abstract

This study investigates whether the TIMSS 2019 Computer Use Questionnaire functions equivalently across languages and cultures. Using responses from 8th-grade students in Türkiye, England, and Qatar, we evaluated cross-group comparability with Multiple-Group Confirmatory Factor Analysis (MGCFA) and examined Differential Item Functioning (DIF) via Ordinal Logistic Regression (OLR) and Poly-SIBTEST. The instrument comprises 11 Likert-type items organized into two factors—Computer Usage Frequency and Computer Usage Self-Efficacy—supported by exploratory and confirmatory factor analyses. For the same-culture/different-language comparison (Qatar Arabic vs. English), configural and metric invariance were supported, whereas scalar invariance was not. For the different-culture/different-language comparison (England vs. Türkiye), only configural invariance was obtained, indicating that factor loadings and intercepts were not fully comparable across these countries. DIF findings varied by method: OLR flagged mostly negligible DIF in the frequency items for the same-culture comparison, while Poly-SIBTEST identified several items with moderate to large DIF; in the cross-culture comparison, both methods indicated DIF for most items, particularly within the self-efficacy factor. The pattern of results suggests that linguistic adaptation, access to technology, and differences in technology-related experiences contribute to nonequivalence. We propose revising culture-sensitive terms, clarifying item contexts, and incorporating qualitative evidence to strengthen score comparability in future administrations.

Keywords: eTIMSS, computer use, measurement invariance, differential item function

Introduction

Rapid advances in digital technology have reshaped teaching, learning, and assessment. International monitoring programs —such as those coordinated by the International Organization for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Cooperation and Development (OECD)— offer periodic snapshots of how education systems perform and how students engage with technology. Within this landscape, TIMSS-Trends in International Mathematics and Science Study provides trend data for mathematics and science at the 4th and 8th grades, and in 2019, many education systems delivered assessments electronically (eTIMSS). Student and contextual questionnaires accompanying these assessments include items about the availability and use of computers and tablets, as well as students' confidence in performing common digital tasks.

When comparing the scores of individuals, it should be ensured that the differences in the scores of individuals are not due to the measurement tool. In other words, it is expected that the different scores of individuals are due to individual characteristics. Meaningful cross-group comparisons require more than identical translations; they require evidence that a questionnaire measures the same underlying constructs in the same way across groups. Validity, in this context, is supported when intended interpretations of scores are justified by theory and empirical evidence. A core source of evidence is measurement invariance (MI)—that is, testing whether the dimensional structure, factor loadings, and

* This study is a part of the master thesis prepared by the first author and conducted under the supervision of the second author.
** Research Assistant, Kırşehir Ahi Evran University, Faculty of Education, Kırşehir-Türkiye, zeynep.kizilcim@ahievran.edu.tr, ORCID ID: 0000-0002-0164-5682
*** Assoc. Professor., Hacettepe University, Faculty of Education, Ankara-Türkiye, tsevda@hacettepe.edu.tr, ORCID ID: 0 0000-0001-5483-595X

To cite this article:
Kızılçim, Z.N.,& Çetin, S. (2025). Cross-lingual and cross-cultural validity of the TIMSS 2019 computer use questionnaire: Using measurement ınvariance and DIF methods. . *Journal of Measurement and Evaluation in Education and Psychology, 16*(4), 202-215. https://doi.org/10.21031/epod.1772214

Received: 26.08.2025
Accepted: 2.11.2025

intercepts are comparable across groups. When MI does not hold, observed differences may reflect translation choices, response styles, or contextual factors rather than true differences in the latent trait.

At the item level, the relevant concept is Differential Item Functioning (DIF). DIF occurs when individuals with the same level of the underlying construct have different probabilities of endorsing an item solely because they belong to different groups (e.g., by language or culture) (Belzak & Bauer, 2020; Molenaar & Feskens, 2024). Detecting DIF helps identify items whose content, wording, or context may advantage one group over another and, thus, threaten fairness and comparability. It is especially crucial to be aware of such biases in high-stakes assessments, as they can have major repercussions (Opesemowo, 2025). Ensuring the validity of tools like the TIMSS 2019 Computer Use Questionnaire is essential, as these instruments inform educational policy and practice by revealing how students' technological engagement shapes learning environments and influences outcomes in mathematics and science.

This study focuses on the TIMSS 2019 Computer Use Questionnaire for 8th graders. We examined: (a) whether the instrument shows configural, metric, scalar, and strict invariance across (i) a same-culture/different-language setting (Qatar Arabic vs. English) and (ii) a different-culture/different-language setting (England vs. Türkiye); and (b) whether items display DIF across these comparisons using two complementary methods (OLR and Poly-SIBTEST). The choice of these settings enables us to disentangle language adaptation issues from broader cultural and contextual differences related to digital access, terminology, and instructional practices.

By combining MI tests with DIF analyses, we provide a detailed account of where comparability holds and where it breaks down. Our findings inform both instrument refinement (e.g., rephrasing technology-specific terms, clarifying usage contexts) and the interpretation of cross-system comparisons that rely on questionnaire-based indicators of students' digital exposure and self-efficacy.

Background

In this part of the study, we focused on computer use, measurement invariance, and differential item functioning, which are important concepts that provided a basis for the study. We defined the concepts, explained their effects and the relationship between them, and depicted the methods used.

Computer Use

The integration of digital technologies such as computers, tablets, and other smart devices into education has transformed the way students interact with learning materials and perform academic tasks. Empirical evidence suggests that technology-assisted learning enhances student engagement and academic achievement when implemented effectively (Keengwe & Hussein, 2014; Valverde-Berrocoso et al., 2022). However, cross-national studies highlight substantial variation in students' access to and familiarity with technology, which may have implications for educational equity and assessment validity (Mullis et al., 2020).

In the TIMSS 2019 cycle, students were asked about their frequency of computer or tablet use for various educational purposes, such as completing assignments, taking tests, or conducting research. These questionnaire items are particularly relevant given the shift to computer-based assessment (eTIMSS) in many countries. Still, digital access and exposure are not uniformly distributed across educational systems. For instance, while over 80% of students in some countries report access to computers, this figure is below 10% in others (Mullis et al., 2020). The type and quality of access (e.g., personal device vs. shared classroom computer) also differ significantly, potentially influencing students' responses to technology-related items.

Beyond access, the instructional use of technology varies as well. TIMSS 2019 data show that computers are used more frequently in science instruction than in mathematics, yet a significant proportion of students report rarely using digital tools for academic tasks. These patterns raise concerns about construct validity when comparing responses to computer use items across diverse educational contexts.

Given these disparities, it is essential to evaluate the measurement invariance and potential item bias in questionnaire items related to computer use. Without such an analysis, differences observed between groups may be due to contextual factors that are unrelated to the construct being measured. This would compromise the validity of cross-cultural comparisons.

Measurement Invariance

Measurement invariance (MI) can be defined as the psychometric property whereby a measurement instrument assesses the same latent construct in an equivalent manner across different groups (e.g., cultures, languages, or genders) (Cheung & Rensvold, 2002; Şekercioğlu, 2018). Ensuring MI is essential in cross-cultural or multilingual research, particularly when interpreting observed score differences as true differences in the underlying trait, rather than artifacts of the measurement process (Meredith, 1993; Kline, 2016).

MI is not a statistical method perse, but a condition that is assessed through structured, multi-step testing using techniques such as Multiple Group Confirmatory Factor Analysis (MGCFA). It is commonly used to examine whether the factor structures, factor loadings, item intercepts, and residual variances are comparable among groups (Brown, 2006). Without adequate evidence of invariance, any between-group comparisons may be misleading or invalid.

The evaluation of measurement invariance typically follows a hierarchical, stepwise framework:

Configural Invariance: As the baseline model, this determines whether there is a common factorial structure (i.e., fixed and free loadings) across groups. When configural invariance is established, the construct has the same conceptual meaning across groups, regardless of the measurement parameters used (Vandenberg &Lance, 2000).

Metric Invariance (Weak Invariance): This level tests whether factor loadings (λ) are equivalent across groups. Valid comparisons between construct associations will be possible when latent constructs are loaded equally across groups. A failure to achieve metric invariance may indicate DIF at the factor loading level.

Scalar Invariance (Strong Invariance): This stage constrains item intercepts and factor loadings (i.e., thresholds) to be equal across groups. A meaningful comparison of latent mean scores requires scalar invariance. Lack of scalar invariance may indicate systematic response style differences or cultural biases.

Strict Invariance: This most stringent level adds the constraint of equal item residual variances (Ɛ) across groups. Although rarely achieved in practice, it provides the strongest evidence that the observed differences in group scores reflect differences in an underlying latent trait rather than measurement error.

Establishing MI is especially important in large-scale international assessments, where instruments are translated and adapted for different populations. In the present study, we tested MI for the computer use scale using MGCFA, adhering to the four-step procedure proposed by Meredith (1993) and Cheung and Rensvold (2002). We evaluated MI across both same-culture/different-language (AR–ENG) and different-culture/different-language (TUR-ENG) group comparisons. Model comparisons were conducted using fit indices such as CFI, TLI, and RMSEA, and changes in CFI (ΔCFI $\leq .01$) were used as the primary criterion to evaluate invariance at each step.

This rigorous approach allows us to determine whether the scale captures the same constructs equivalently across diverse linguistic and cultural contexts, providing the foundation for valid cross-group comparisons and further analysis of potential item bias through DIF methods.

Differential Item Functioning

Differential Item Functioning (DIF) is a psychometric condition whereby individuals from different groups (e.g., cultural, linguistic, or gender groups) who have equivalent levels of an underlying latent trait (e.g., digital literacy or self-efficacy) are more or less likely to endorse or correctly respond to a test item (Angoff, 1993; Zumbo, 1999). DIF indicates potential item bias, which is a violation of MI at the item level. This calls into question the fairness and validity of assessments, particularly in cross-cultural contexts (Zwick, Donoghue & Grima, 1993). Two primary forms of DIF are recognized in the literature:

Uniform DIF: In this case, the differences in item response probabilities between groups are consistent across all levels of the latent trait. The uniform DIF occurs when students in Group A consistently endorse an item more than students in Group B, regardless of their ability level. (Swaminathan & Rogers, 1990).

Non-uniform DIF: This occurs when the magnitude or direction of group difference varies with ability level. This type of DIF implies an interaction between group membership and the latent trait, such that an item might favor one group at lower ability levels and the other group at higher levels. (Shepard, Camilli, & Williams, 1985).

Since no single method provides a definitive diagnosis of DIF, best practices recommend triangulating multiple analytical techniques and supplementing them with expert judgment (Ackerman, 1992; Dorans & Holland, 1993). In this study, we used two well-established methods for detecting DIF: Poly-SIBTEST and Ordinal Logistic Regression (OLR).

Poly-SIBTEST

Poly-SIBTEST (Polytomous Simultaneous Item Bias Test) is a nonparametric DIF detection method developed for both dichotomous and polytomous (e.g., Likert-type) items. It compares item performance between reference and focal groups while conditioning on a matching criterion (e.g., total score or estimated latent ability). The analysis yields a DIF effect size coefficient, β, which quantifies the extent and direction of bias (Roussos & Stout, 1996). Poly-SIBTEST requires defining an anchor and a studied item set, where the β coefficient quantifies the extent of DIF. This β value is derived by comparing item performance conditional on estimated ability and reflects the degree of item bias.

•A positive and statistically significant β indicates DIF against the focal group (i.e., the item favors the reference group).

•A negative and significant β indicates DIF against the reference group.

•The size of DIF is categorized as follows:

$|\beta| < 0.059$ → Level A: Negligible DIF

$0.059 \leq |\beta| < 0.088$ → Level B: Moderate DIF

$|\beta| \geq 0.088$ → Level C: Large DIF

Poly-SIBTEST is advantageous for handling ordinal data and does not require strict assumptions of parametric models. However, it is sensitive to the choice of matching variable and requires careful item anchoring procedures to avoid contamination effects.

Ordinal Logistic Regression (OLR)

Ordinal Logistic Regression (OLR) is a widely used parametric method for DIF detection, especially suitable for polytomously scored items such as those on Likert scales (Zumbo, 1999). OLR compares nested models to determine whether group membership significantly predicts item responses after controlling for the latent trait.

The method tests DIF through three hierarchical models:

1.Model 1: Includes only the latent trait (θ or total score).

2.Model 2: Adds group membership (reference vs. focal group) to detect uniform DIF.

3.Model 3: Adds an interaction term between the latent trait and group to detect non-uniform DIF.

The statistical significance of the comparison between Model 1 and Model 2 (or Model 3) indicates the presence of DIF. Additionally, the change in pseudo-R^2 (e.g., McFadden's R^2) is used as an effect size measure:

$R^2 < 0.13$ → Level A: Negligible DIF

$0.13 \leq R^2 < 0.26$ → Level B: Moderate DIF

$R^2 \geq 0.26$ → Level C: Large DIF (Zumbo & Thomas, 1996).

OLR allows for the detection of both uniform and non-uniform DIF and is particularly effective when used with large sample sizes and well-fitting models. However, it requires the correct specification of the latent trait and assumes proportional odds, which may not always hold in practice.

In the present study, we examined DIF in computer use items from the TIMSS 2019 student questionnaire across two group comparisons: Same culture–different language: Arabic-English (Qatar) and Different culture–different language: Türkiye-English (Türkiye vs. England).

We applied both Poly-SIBTEST and OLR to each item, using total test scores as the matching criterion. The consistency and divergence between methods were noted, and items flagged by either method were further discussed in terms of effect size and directionality of bias. This dual-method approach enhanced the robustness of our DIF analysis and provided converging evidence for interpreting cross-group differences in item functioning.

Method

The Research Sample

The study drew on 8th-grade student data from Türkiye, England, and Qatar collected during eTIMSS 2019. These three education systems were selected to capture both cross-language comparisons within a shared cultural context (AR-ENG) and cross-cultural comparisons with different languages (TUR-ENG). The Turkish sample included 4,077 students, the English sample 3,365 students, and the Qatari sample 3,745 students (1,891 tested in Arabic, 1,854 in English). Treating the Qatar subgroups separately allowed for direct examination of the influence of language within a single national context.

Data Analysis

We conducted the data analysis in four steps. In the first step, assumptions were examined before starting the analyses. For this purpose, missing values and outliers were examined, and normality-linearity and multicollinearity were checked. In the second step, Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were conducted to determine the items' dimensions. In the third step, measurement invariance, and in the fourth step, DIF analyses were performed. SPSS 22.0, R Studio, and SIBTEST software were used for measurement invariance and DIF analyses.

Data Collection Tools

Data were obtained from the publicly available TIMSS 2019 database provided by the IEA. We focused on the student background questionnaire items related to computer use. Eleven items, organized into two hypothesized dimensions—Computer Usage Frequency and Computer Usage Self-Efficacy—were analyzed. Items previously flagged for issues of dimensionality (e.g., BSBE03) or dichotomous response format (e.g., BSBE01) were excluded from the main analyses. Reliability analyses were also conducted to evaluate internal consistency within each factor for each national sample.

Data Screening and Preparation

Prior to statistical modeling, data quality checks were conducted. Missing values exceeded 5% overall; therefore, missing responses were imputed using the Expectation–Maximization (EM) algorithm to preserve sample size. Outliers were identified at both the univariate level (z-scores outside ±3) and the multivariate level (based on Mahalanobis distance, $p<.01$) and removed. This resulted in the exclusion of 875 Turkish cases, 351 English cases, and 141 Qatari cases. Tests of multivariate normality using the Henze–Zirkler procedure indicated significant departures from normality for all groups (England Hz: 7.527, $p<0.05$; Türkiye Hz: 32.341, $p<0.05$; Qatar Hz: 32.341, $p<0.05$). Given the ordinal scale of items and non-normal distribution, analyses employed the Diagonally Weighted Least Squares (DWLS) estimation method.

Results

Factor Analyses

The items coded "BSBE03" were not included in the analyses due to the lack of unidimensionality in the EFA result comparisons across countries, and the items coded "BSBE01" were not included in the analyses because they were dichotomously scored items. When the results of EFA analysis were examined, the results of KMO and Barlett Sphericity tests for all three country samples England sample; KMO=0.798 and Barlett Sphericity (χ^2= 7438.218; $p<0.001$); Türkiye sample; KMO=0.808 and Barlett Sphericity (χ^2= 13070.391; $p<0.001$); Qatar sample; KMO=0.780 and Barlett Sphericity (χ^2= 11251.176; $p<0.001$) showed that the data sets were suitable for EFA analysis (Kaiser, 1974; Sharma, 1996).

In EFA, factor loadings must be above 0.30 (Floyd & Widaman, 1995; Tabachnick & Fidell, 2007). As seen in the Table 1, factor loadings for all three country samples ranged from 0.55 to 0.83 for one dimension and from 0.39 to 0.69 for the other dimension. An 11-item, two-factor structure was confirmed for England, Türkiye and Qatar. Based on these results, the dimensions were labeled "Computer Usage Frequency" and "Computer Usage Self-Efficacy".

Table 1
EFA results for computer use items

		Computer Usage Frequency			Computer Usage Self-Efficacy		
Code	**Item**	**ENG**	**TUR**	**QAT**	**ENG**	**TUR**	**QAT**
BSBE02A	At school this year how often did you use a computer or tablet to do work on a school assignment such as a paper, report, or presentation?	0.55	0.70	0.58			
BSBE02B	At school this year how often did you use a computer or tablet to do Mathematics schoolwork?	0.78	0.82	0.78			
BSBE02C	At school this year how often did you use a computer or tablet to do Science schoolwork?	0.81	0.83	0.77			
BSBE02D	At school this year how often did you use a computer or tablet to take a test or quiz?	0.60	0.71	0.60			
BSBE04A	How well do you know the term WIFI?				0.57	0.58	0.39
BSBE04C	How well do you know the term instant messaging?				0.65	0.61	0.59
BSBE04D	How well do you know the term cut and paste?				0.69	0.67	0.53
BSBE04E	How well do you know the term spreadsheet?				0.65	0.64	0.58
BSBE04F	How well do you know the term icon?				0.70	0.69	0.69
BSBE04G	How well do you know the term drag and drop?				0.75	0.60	0.69
BSBE04H	How well do you know the term scroll?				0.66	0.62	0.64

Subsequently, Confirmatory Factor Analysis (CFA) was conducted. The reported goodness-of-fit indices in Table 2 indicate that the model demonstrates a good fit, suggesting that the 11-item, two-factor structure has been validated through CFA for the datasets from England, Türkiye, and Qatar.

Table 2
Goodness of Fit Indices

Sample	χ^2	χ^2 / sd	CFI	RMSEA	AGFI	TLI
England	452.297*	8.533	0.958	0.050	0.979	0.948
Türkiye	607.696*	1.146	0.963	0.053	0.978	0.953
Qatar	696.205*	1.313	0.949	0.058	0.976	0.937

*p<0,05

Reliability analyses for the England Sample showed that Cronbach's α value was 0.84 for the first factor, "Computer Usage Frequency," and 0.78 for the second factor, "Computer Usage Self-Efficacy." For the Türkiye Sample, α value was 0.66 for the first factor, "Computer Usage Frequency," and 0.78 for the second factor, "Computer Usage Self-Efficacy", for the Qatar sample, α value was 0.78 for the first factor, "Computer Usage Frequency" and 0.76 for the second factor "Computer Usage Self-Efficacy"

Invariance Analyses

According to the results given in Table 3, the configural invariance model established for the same culture-different languages (AR-ENG); Qatar was found to have acceptable fit indices for Arabic and English languages (CFI > .90; GFI > .95; RMSEA < .08; TLI > .90). It was concluded that the GFI indicated good fit, while the CFI, TLI, and RMSEA values indicated acceptable levels of fit. Considering these results, it can be concluded that the fit index values are at an acceptable level. This means that the measurement tool used by 8th-grade students in Qatar, who take the items in different languages (English and Arabic), has the same structure for both groups. In other words, configural invariance according to language is provided.

After the configural invariance, which is a prerequisite for metric invariance, metric invariance was tested. When the goodness of fit indices of metric invariance was examined, it was seen that they were at an acceptable level (CFI>.90; GFI>.95; RMSEA <.08; TLI>.90). When these results are taken into consideration, it is seen that the measurement tool has the same structure and the item factor loadings are the same for students answering the questions in different languages. Since the chi-square statistic can give biased results in large data sets, the CFI difference value (ΔCFI) was calculated to compare the two measurement invariance stages (Cheung & Rensvold, 2002). The calculated ΔCFI=0.008 value was found to be within ±0.01, which is the accepted limit for fit. This result shows that the items related to 8th-grade students' computer use provide metric invariance according to language.

After ensuring metric invariance, scalar invariance, the next stage of measurement invariance, was tested. It was observed that the goodness of fit index values in the scalar invariance stage, except for the RMSEA and GFI coefficients, were not at an acceptable level (CFI>.90; GFI>.95; RMSEA <.08; TLI>.90). At the same time, the ΔCFI value between metric invariance and scalar invariance was not within the acceptable range (±0.01). The fact that scalar invariance was not achieved indicates that the item constants of the items related to 8th-grade students' computer use in TIMSS 2019 are not similar. This situation indicates that there may be linguistic bias in the items subject to the study.

Table 3

Measurement Invariance Fit Statistics Results

	Model	χ^2	χ^2 / sd	RMSEA	CFI	ΔCFI	GFI	TLI
Same culture-Different language AR-ENG	Configural Invariance	750.560	1.10	.064	.94	…	.98	.92
	Metric Invariance	842.732	1.68	.064	.93	.001	.97	.92
	Scalar Invariance	1294.309	1.24	.077	.89	.04	.96	.88
Different language-Different culture ENG- TUR	Configural Invariance	1635.784	1.90	.073	.92	…	.99	.90
	Metric Invariance	2276.424	2.39	.083	.89	.03	.98	.87

When the measurement invariance of the established model was examined for different languages-different cultures (TUR-ENG), fit indices for configural invariance were assessed in the initial stage. As seen in Table 3, the fit indices indicated an acceptable level (CFI>.90; GFI>.95; RMSEA<.08; TLI>.90). According to the results, it can be said that the measurement tool measures the same construct for 8th-grade students taking the items in the context of England and Türkiye, in different languages-different cultures (TUR-ENG). Since the establishment of configural invariance is a prerequisite for testing metric invariance, the analysis proceeded to test metric invariance. The goodness-of-fit indices for metric invariance, however, were found to be at unacceptable levels (CFI>.90; GFI>.95; RMSEA<.08; TLI>.90). Among the fit indices, only the GFI was found to indicate a good fit. These results indicate that item factor loadings of computer use items are not similar according to different languages-different cultures (TUR-ENG) and that there may be item bias.

DIF Analyses

Different Language-Same Culture (AR-ENG) DIF Analysis Results: DIF results within the ORL method were evaluated based on the McFadden Model3-Model1 R^2. Table 4 shows that only the items in the "Computer Usage Frequency" dimension exhibited DIF at Level A. For items with DIF at this level, DIF is either absent or negligible.

Table 4

Different Language-Same Culture (AR-ENG) DIF Analysis Results

	OLR Results of Analysis					SIBTEST Results of Analysis			
Item	β_1	R^2_{1v2}	R^2_{1v3}	R^2_{2v3}	**Level**	p	$\beta_{Estimate}$	**Level**	**Favor**
BSBE02A	.000	.000	.001	.001	A	.002	.093	C	ENG
BSBE02B	.023	.000	.010	.009	A	.000	.242	C	ENG
BSBE02C	.017	.000	.009	.009	A	.000	.121	C	ENG
BSBE02D	.019	.000	.008	.008	A	.000	.210	C	ENG
BSBE04A	.007	.000	.000	.000	-	.000	-.064	B	AR
BSBE04C	.000	.000	.000	.000	-	.009	-.067	B	AR
BSBE04D	.010	.000	.000	.000	-	.000	-.115	C	AR
BSBE04E	.003	.000	.000	.000	-	.008	-.074	B	AR
BSBE04F	.015	.000	.000	.000	-	.034	-.059	B	AR
BSBE04G	.009	.000	.000	.000	-	.000	-.131	C	AR
BSBE04H	.012	.000	.000	.004	-	.000	-.125	C	AR

According to the analysis results of the Poly-SIBTEST method, four items exhibited DIF at Level B and seven items at Level C. It was observed that items BSBE02A, BSBE02B, BSBE02C, and BSBE02D showed DIF in favor of English, while other items showed DIF in favor of Arabic.

The results of analyses conducted using the OLR and Poly-SIBTEST methods differed when comparing the same culture-different language (AR-ENG). However, for items related to the "Computer Usage Frequency" dimension, both methods concluded that the items exhibited DIF in favor of English, even though the levels of DIF varied between the methods.

Different Language-Different Culture (TUR-ENG) DIF Analysis Results: DIF results within the ORL method were evaluated based on the McFadden Model3-Model1 R^2. When the DIF analysis results seen in Table 5 were examined, the items in the "Computer Usage Frequency" dimension exhibited DIF at Level A. In the “Computer Usage Self-Efficacy” dimension, one item was at Level A, three were at Level B, and three were at Level C in terms of DIF. This indicates that while one item exhibited negligible DIF, the other six displayed moderate or high levels of DIF.

Table 5

Different Language-Different Culture (TUR-ENG) DIF Analysis Results

	OLR Results of Analysis					SIBTEST Results of Analysis			
Item	β_1	R^2_{1v2}	R^2_{1v3}	R^2_{2v3}	**Level**	p	$\beta_{Estimate}$	**Level**	**Favor**
BSBE02A	.023	.002	.004	.001	A	.229	-.028	A	TUR
BSBE02B	.000	.000	.001	.001	A	.000	.233	C	ENG
BSBE02C	.027	.003	.004	.001	A	.201	.028	A	ENG
BSBE02D	.130	.000	.004	.004	A	.134	-.032	A	TUR
BSBE04A	.055	.019	.019	.000	B	.000	.221	C	ENG
BSBE04C	.003	.002	.002	.000	A	.031	-.041	A	TUR
BSBE04D	.110	.041	.042	.000	C	.000	.254	C	ENG
BSBE04E	.010	.017	.017	.000	B	.000	-.243	C	TUR
BSBE04F	.051	.036	.037	.000	C	.000	-.368	C	TUR
BSBE04G	.077	.047	.047	.000	C	.000	-.264	C	TUR
BSBE04H	.042	.016	.020	.004	B	.000	-.113	C	TUR

According to the Poly-SIBTEST method analysis results in Table 5, four items exhibited DIF at Level A, while seven items were at Level C. Furthermore, it was observed that items BSBE02B, BSBE02C, BSBE04A, and BSBE04D favored those from the English group, whereas other items favored test-takers from the Turkish group.

When the results were examined for different languages-different cultures (TUR-ENG), it was found that all items exhibited DIF according to the results of both analysis methods.

Discussion

The transition to eTIMSS in the TIMSS cycle in 2019 and the widespread use of computers and tablets, together with the fact that some countries have joined the assessment cycle in this way, may have affected the assessment results. When comparing the scores of countries, it should be ensured that the differences in scores are not due to the measurement tool, so it is very crucial to investigate and verify measurement invariance in studies comparing different groups. This is because a lack of evidence of MI can make it difficult to interpret differences between groups clearly. For these purposes, it was important to conduct measurement invariance studies for the items related to computer use in the eTIMSS student questionnaires and then to conduct a DIF analysis by considering the measurement invariance results.

In the study, it was found that in the same culture-different language (AR-ENG) comparison, configural invariance and metric invariance were achieved, but scalar and strict invariance were not achieved, and only configural invariance was provided in the different culture-different language (TUR-ENG) comparison. These and other studies have shown that measurement invariance is compromised as cultural differences increase (Asil & Gelbal, 2012; Yandı, Köse & Uysal, 2017). Configural invariance indicates that the construct's structure is consistent across different groups and is generally established in studies that test for full measurement invariance. The lack of metric invariance may be due to method bias. This bias occurs when most or all test items are similarly affected by a factor that is independent of the construct under study. For instance, if acquiescence is a general trend in certain cultures, it will affect the responses to all items (Van Herk et al., 2004). Some studies have interpreted the lack of scalar invariance as an indication of culturally determined response styles. (e.g., Acceptance Response Style, ARS). Lack of scalar invariance can result from scale drift over groups, which is caused by a type of ARS. (Mullen, 1995; Smith, 2004; Van Herk et al., 2004). Van Herk et al.'s study found that the ratings on the scale did not correspond to the actual behaviour seen across different countries. This suggests that if researchers ignore the way in which people from different countries respond, it could potentially result in incorrect conclusions being drawn in cross-cultural research.

Beyond cultural and linguistic factors, educational policy differences such as curriculum emphasis on technology use, access to digital devices, and levels of digital literacy can all contribute to differential responses across countries. For example, students from countries with higher availability of computers at home or school may have more familiarity, leading to an advantage on certain items. Socioeconomic status may also influence access and experience with digital tools, introducing further variability in item responses. For practitioners, these findings underscore the need to revise questionnaire items to minimize cultural and linguistic biases. For instance, items like "*BSBE02A: At school this year, how often did you use a computer or tablet to do work on a school assignment such as a paper, report, or presentation?*" and "*BSBE04A: How well do you know the term WIFI?*" showed variability in factor loadings across countries, potentially due to differences in computer access or familiarity with technical terms. Psychoeducational consultants can collaborate with test developers to rephrase such items to use more universally understood language or to include context-specific examples that resonate across cultures (e.g., replacing "WIFI" with "internet connection"). Additionally, items should be piloted with diverse groups to ensure equivalence in meaning and difficulty, reducing the likelihood of DIF.

The difference in the computer access rates of the countries compared may be a reason why students did not respond to the questionnaire items similarly. Students' access to computers in math and science lessons are as follows, respectively, according to the countries (Mullis et al., 2020): In Qatar, 38 percent for math and 50 percent for science; in England, it is 24 percent for math and 32 percent for science and in Türkiye, it is 15 percent for math and 18 percent for science. The percentages of each student who has a computer in these countries are also 17-24 percent in Qatar, 13-14 percent in England, and 1-3 percent in Türkiye, respectively. The percentages in Türkiye are considerably lower than in the other two countries. Also, in the established model, it was observed that the results of exploratory factor analysis varied across countries. Specifically, items such as "BSBE02A," "BSBE02D," "BSBE04A," and "BSBE04D" showed variability in country factor loadings. Differences in access to computers across countries and variability in factor loadings may be the reasons why students interpreted the items differently and, as a result, invariance failed to be achieved. Educational diagnosticians can use these findings to interpret TIMSS questionnaire results cautiously, recognizing that lower computer access in some countries (e.g., Türkiye) may lead to underreporting of computer use frequency or self-efficacy, which could skew diagnostic conclusions about students' technological readiness. To address this, diagnosticians should supplement questionnaire data with qualitative assessments, such as interviews or observations, to capture contextual factors influencing responses.

The fact that measurement invariance was not achieved indicated that there might be bias in the items in the model. In the different language-same culture (AR-ENG) comparisons, four items were found to be DIF at level A in the findings obtained with the OLR method, but when the same comparison was made with the Poly-SIBTEST method, each item was found to be DIF at levels C or B. In previous bias studies on international exams, it was seen that there may be bias according to translation and culture, according to both DIF analysis methods and expert opinions on these items (Atalay, 2010; Asil & Gelbal, 2012; Gök et al., 2014; Uzun & Gelbal, 2017; Gören et al., 2024). It has been concluded that linguistic differences are one of the possible sources of DIF in international exams (Oliveri et al., 2012; Liu, 2019; Bialo & Li, 2022). In the model established according to different language-different cultures (TUR-ENG), all of the items were found to have DIF, even though their levels were different according to the results of two different methods. When the findings obtained were compared with the study conducted by Oliveri et al. (2012), it was seen that different DIF methods detected items with DIF similarly, but the DIF levels of the detected items were different. For practitioners, these results highlight the importance of training assessors to recognize and account for cultural biases in assessment administration. Training programs should include modules on cultural competence, teaching assessors to identify items prone to DIF (e.g., those with culture-specific terminology like “cut and paste”) and to adjust administration procedures, such as providing additional explanations or context for students from less technology-rich environments. This training can enhance the fairness of assessments and ensure more accurate diagnostic outcomes.

While some items had negligible DIF, some items had high levels of DIF. In the comparison in this study (AR-ENG and TUR-ENG), there was no consistency between the DIF levels calculated according to the OLR and poly-SIBTEST methods. These results are similar to previous studies (Atar & Kamata, 2011; Choi et al., 2011). In the different language-same culture (AR-ENG) comparison, DIF was observed in fewer items. The different language-different culture (TUR-ENG) comparison was the group with more DIF. This is similar to the study conducted by Gök et al. (2014). Their study stated that the common characteristic of countries with more than 90 percent of items with DIF is that they have different cultures. In addition, they stated that the effect of countries having the same or different languages on the number of items with DIF is negligible. The result of this study is that cultural differences, rather than language differences, are effective in the DIF of items. This finding is also supported by similar studies (Asil & Gelbal, 2012; Huang, 2010; Mohorić & Takšić, 2016), in which the number of items with DIF increased in groups with more differentiation. Psychoeducational consultants can leverage these findings to advocate for culturally responsive assessment practices. For example, when consulting with schools or educational systems, they can recommend adapting assessments for diverse populations by incorporating items that reflect universal experiences or by using adaptive testing formats that adjust item difficulty based on students’ cultural and technological backgrounds. Such adaptations can improve the validity of diagnostic assessments and support equitable educational planning.

Recommendations and Limitations

This study examined the findings of measurement invariance and DIF across language and culture by selecting some of the educational systems that received the items in English, taking as a reference point the language in which the questionnaire on computer use was developed. The findings suggest the need for studies to consider language and cultural differences in comparisons of educational systems and to investigate strong evidence of validity. Future research should also investigate whether the characteristics of computer use in different cultures and languages are similar in different international applications.

For practitioners, the study offers actionable recommendations to enhance the fairness and validity of psychoeducational assessments. First, test developers and psychoeducational consultants should revise questionnaire items to reduce cultural and linguistic biases. For example, items like “BSBE04A” (knowledge of “WIFI”) and “BSBE04D” (knowledge of “cut and paste”) showed DIF, likely due to varying familiarity with technical terms across cultures. These items could be rephrased to use more neutral language (e.g., “internet access” instead of “WIFI”) or supplemented with examples relevant to diverse educational contexts. Second, training programs for assessors should emphasize cultural competence, equipping them to recognize items susceptible to DIF and to provide contextual support during administration, such as clarifying terminology for students with limited technology exposure. Third, educational diagnosticians should integrate findings from this study into their practice by using supplementary qualitative methods (e.g., student interviews or teacher reports) to contextualize questionnaire responses, particularly in countries with low computer access, like Türkiye. These strategies can improve the accuracy of diagnostic conclusions about students' technological readiness and inform targeted interventions to enhance digital literacy. Finally, psychoeducational consultants can advocate for adaptive assessment models that account for cultural and technological disparities, ensuring that assessments are equitable and relevant for diverse student populations. By implementing these practices, assessment specialists can contribute to more valid and fair educational diagnostics, supporting equitable educational outcomes.

This study is limited to items related to computer use in Qatar, England, and Türkiye, TIMSS 2019. Comparisons can be made with education systems in different cultures and languages. In addition, testing DIF analyses with different methods will allow for more valid results and more information about the items to be obtained. Different methods can be used in future studies. The quantitative focus of this study limits insights into the qualitative aspects of cultural and linguistic biases. Future research could incorporate qualitative methods, such as expert reviews of item content or focus groups with students, to provide deeper explanations for DIF and enhance the interpretability of findings.

In the study, the items coded "BSBE03" were not included in the analysis because they failed to achieve unidimensionality across countries, while the items coded "BSBE01" were not included in the analysis because they were evaluated with binary scoring. In future studies, validity evidence related to invariance and DIF of these items, which could not be examined with different models to be established, can be examined.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This study utilized publicly available data. Therefore, ethical approval was not required. The authors confirm that all ethical guidelines were followed.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3–23). Lawrence Erlbaum Associates, Inc

Asil, M. & Gelbal, S. (2012). Cross-cultural Equivalence of the PISA Student Questionnaire. *Education and Science*, 37(166), 236-249.

Atalay, K. (2010). *PISA 2006 öğrenci anketinde yer alan tutum maddelerinin değişen madde fonksiyonu açısından incelenmesi*. (Yüksek Lisans Tezi) Hacettepe Üniversitesi, Ankara.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe Üniversitesi Eğitim Bilimleri Dergisi, 41*, 36–47.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological methods*, *25*(6), 673.

Bialo, J. A., & Li, H. (2022). Fairness and Comparability in Achievement Motivation Items: A Differential Item Functioning Analysis. *Journal of Psychoeducational Assessment*, 40(6), 722-743. https://doi.org/10.1177/07342829221090113

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, N. J.: Guilford Press.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif : An R package for detecting differential item functioning using iterative ordinal logistic regression/ item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30.

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning (pp. 35–66).* Lawrence Erlbaum Associates, Inc.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35.

Fishbein, B., Martin, M. O., Mullis, I. V., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*(1), 1–23.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, *7*(3), 286.

Gök, B., Kabasakal, K. A., & Kelecioglu, H. (2014). Analysis of attitude items in PISA 2009 Student Questionnaire in terms of differential item functioning based on culture. *Journal of Measurement and Evaluation in Education and Psychology*, 5(1), 72-87. https://doi.org/10.21031/epod.64124

Gören, S., Sayın, A., & Gelbal, S. (2024). An Analysis of Item Bias in the PISA 2018 Reading Understanding and Memorising Strategies Questionnaire, *Kastamonu Education Journal*, 32(2), 345-356. doi: 10.24106/kefdergi.X

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, *19*(10), 3595-3617.

Huang, X. (2010). *Differential Item Functioning: The Consequence of Language, Curriculum, or Culture?* (Doktora Tezi). University of California, Berkeley.

Kaiser, H.F. (1974). An Index of Factorial Simplicity. *Psychometrika, 39*, 31-

Keengwe, J., & Hussein, F. (2014). Using computer-assisted instruction to enhance achievement of English language learners. *Education and information technologies*, 19, 295-306.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford.

Liu, R. (2019). *DIF Among English Language Learners on a Large-Scale*

Meredith, W. (1993). *Measurement invariance, factor analysis, and factorial invariance*. Psychometrika, 58(4), 525– 543. https://doi.org/10.1007/BF02294825

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, *26*, 573-596.

Mohorić, T., & Takšić, V. (2016). *DIF analiza upitnika emocionalne kompetentnosti Mantel-Haenszel metodom: kros-kulturalna usporedba (DIF analysis of ESCQ using Mantel-Haenszel procedure: Cross-cultural comparison).* In XX. Danipsihologije u Zadru.

Molenaar, D., & Feskens, R. (2024). Relating violations of measurement invariance to group differences in response times. *Psychological Methods.* Advanced online publication. https://doi.org/10.1037/met0000655

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*.

OECD (2023). *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/dfe0bf9c-en.

Oliveri, M.E., Olson, B.F., Ercikan, K. & Zumbo, B. (2012). Methodologies for Investigating Item- and Test-Level Measurement Equivalence in International Large-Scale Assessments, *International Journal of Testing*, *12*:3, 203–223, doi: 10.1080/15305058.2011.617475

Opesemowo, O. A. G. (2025). Exploring undue advantage of differential item functioning in high-stakes assessments: Implications on sustainable development goal 4. *Social Sciences & Humanities Open*, *11*, 101257.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*(2), 215-230. https://doi.org/10.1111/j.1745- 3984.1996.tb00490.x

Sharma, S., (1996). *Applied Multivariate Techniques*. John Wiley and Sons, Inc., New York

Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77–105. https://doi.org/10.1111/j.1745-3984.1985.tb01050.x

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of cross-cultural psychology*, *35*(1), 50–61.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Şekercioğlu, G. (2018). Measurement invariance: Concept and implementation. *International Online Journal of Education and Teaching (IOJET)*, 5(3). 609-634.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). New York: Allyn and Bacon.

Uzun, N. B., & Gelbal, S. (2017). PISA fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi. *Kastamonu Eğitim Dergisi*, *25*(6), 2427-2446.

Valverde-Berrocoso, J., Acevedo-Borrega, J., & Cerezo-Pizarro, M. (2022, June). Educational technology and student performance: A systematic review. In *Frontiers in Education* (Vol. 7, p. 916502). Frontiers Media SA.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. https://doi.org/10.1177/109442810031002

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics*, *21*(2), 110–130.

Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: PISA 2012 örneği. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 243-253.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251.