frontiers in Education

TYPE Original Research
PUBLISHED 19 June 2026
DOI 10.3389/feduc.2026.1823761

OPEN ACCESS

EDITED BY
Soo Lee,
American Institutes for Research,
United States

REVIEWED BY
Vanessa Scherman,
International Baccalaureate (IBO),
Netherlands
Esra Sozer Boz,
Çanakkale Onsekiz Mart University,
Türkiye

*CORRESPONDENCE
Rolf Strietholt
rolf.strietholt@iea-hamburg.de

RECEIVED 05 March 2026
REVISED 21 May 2026
ACCEPTED 05 June 2026
PUBLISHED 19 June 2026

CITATION
Baghaei P, Strietholt R and Christiansen A (2026) A large-scale empirical investigation of measurement invariance decisions under multiple-group item response theory and multiple-group confirmatory factor analysis.
Front. Educ. 11:1823761.
doi: 10.3389/feduc.2026.1823761

A large-scale empirical investigation of measurement invariance decisions under multiple-group item response theory and multiple-group confirmatory factor analysis

Purya Baghaei[1], Rolf Strietholt[2]* and Andrés Christiansen[1]

[1]International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany, [2]Institute for Social Pedagogy, Adult Education and Pedagogy in Early Childhood, Technische Universität Dortmund, Dortmund, Germany

Measurement invariance (MI) testing is essential for ensuring valid cross-group comparisons in international large-scale assessments (ILSA). There are two major frameworks for establishing MI: multiple-group confirmatory factor analysis (MGCFA) and multiple-group item response theory (MGIRT). This study compares the results of MGCFA and MGIRT in examining measurement invariance using survey data from the International Computer and Information Literacy Study (ICILS) 2023. This study goes beyond prior simulation-based comparisons by providing a large-scale empirical examination of measurement invariance decisions across 33 ICILS questionnaire scales and 32 educational systems under operational assessment conditions. Using common thresholds for model evaluation, the results from the MGIRT suggest invariance across most scales compared to MGCFA, which rejects configural invariance for several scales. This finding suggests that the choice of method, MGCFA or MGIRT, can lead to substantially different conclusions. These findings underscore the urgent need for further methodological research to better understand the conditions under which each approach performs reliably and to guide researchers in making informed choices when assessing measurement invariance.

KEYWORDS
International Computer and Information Literacy Study, measurement invariance, multiple group categorical confirmatory factor analysis, multiple group item response theory, RMSD

Introduction

When survey scales are used to make comparisons across countries or subpopulations such as gender, language, or ethnicity measurement invariance (MI) must hold. MI implies that the test or scale measures the same latent trait in the same way across various groups, time points or test modes (e.g., paper-and-pencil vs. computer). Measurement invariance is essential for making valid comparisons between groups as it ensures that differences observed in the scores are due to true differences in the underlying trait, rather than differences in how the test or scale functions for the groups (Meredith and Teresi, 2006).

In the CFA literature, at least three types or levels of MI are discussed. Configural invariance holds when the number of factors and the pattern of loadings is the same across the groups. Configural invariance guarantees that the same construct is measured for all the groups. The next level of MI after configural invariance is established, is metric invariance. Metric invariance refers to the equality of factor loadings across the groups. It posits that a unit change in the latent factor is translated to same amount of change in the item scores for all groups. Metric invariance supports the comparability of variances across groups. The third level of MI, namely, scalar invariance requires equality of item intercepts (in addition to the equality of factor structures, pattern of loadings, and factor loadings). However, note that for ordinal indicators, scalar invariance refers to the equality of item thresholds (rather than intercepts) across groups, in addition to equal factor loadings and factor structure. This ensures that individuals with the same level of the latent trait have the same probability of endorsing response categories across groups (Kim and Yoon, 2011). Scalar invariance supports the comparability of the latent variable means across the groups (Wu et al., 2007).

In the IRT framework, measurement invariance is examined in the context of differential item functioning (DIF). In fact, DIF refers to measurement non-invariance. An item exhibits DIF, if examinees from different groups with the same level of the latent trait have different probabilities of endorsing an item. Scalar non-invariance translates to uniform DIF where an item is systematically more difficult for members of one group (even when they have the same level of the latent trait). Metric non-invariance translates to non-uniform DIF where the IRT discrimination parameter for an item differs across the two groups. This indicates that the item is mor related to the latent variable in one group and the changes in item difficulty, with reference to the latent trait, are not consistent across groups. Uniform DIF is characterized with parallel item characteristic curves (ICCs) for an item across groups while non-uniform DIF is characterized by intersecting ICCs.

In international comparative large-scale studies, such as TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), and PISA (Programme for International Student Assessment), where many countries with different languages and cultures take part and the goal is to compare the participating countries in the measured skills and abilities, establishing MI is essential. The most commonly used statistical method for examining measurement invariance is multiple-group confirmatory factor analysis (MGCFA; Buchholz and Hartig, 2020; Cieciuch et al., 2014; Greiff and Scherer, 2018; Vandenberg and Lance, 2000). However, multiple-group item response theory (MGIRT) models have also been used recently to examine measurement invariance in the context of international large-scale assessments (ILSA).

Although MGCFA and MGIRT are both used to evaluate measurement invariance, they conceptualize non-invariance differently. MGCFA tests a sequence of model-level hypotheses: whether the same factor structure holds across groups, whether factor loadings are equal, and whether item thresholds are equal. Thus, configural, metric, and scalar invariance are evaluated through global model fit and nested equality constraints. MGIRT, by contrast, evaluates whether item response functions can be represented by common item parameters while allowing group-specific latent trait distributions. Within this framework, DIF is indicated when the response function for an item in each group deviates meaningfully from the common or international item response function.

Consequently, MGIRT-based DIF evidence should not be interpreted as a direct one-to-one test of configural, metric, or scalar invariance in the CFA sense. Rather, low RMSD values indicate that the common MGIRT model reproduces group-specific item response patterns sufficiently well, whereas high RMSD values suggest item-level non-invariance that may require group-specific item parameters. This distinction is important because MGCFA and MGIRT may lead to different practical conclusions about cross-group comparability, especially in many-group ILSA settings with ordinal items and heterogeneous populations. The present study therefore does not assume that MGCFA and MGIRT constitute identical tests of invariance but rather examines how their application leads to different practical conclusions regarding cross-group comparability in a large-scale international assessment context.

The International Computer and Information Literacy Study

The International Computer and Information Literacy Study (ICILS) is an international large-scale assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA) to examine how well students are prepared to participate in an increasingly digital world (Fraillon and Rožman, 2023). ICILS assesses eighth-grade students' computer and information literacy (CIL), defined as the ability to use digital technologies to investigate, create, and communicate information effectively, as well as computational thinking (CT), which refers to the ability to recognize and formulate problems in ways that can be addressed using computational tools and concepts. In addition to the cognitive assessments, ICILS administers extensive contextual questionnaires to students, teachers, school principals, and ICT coordinators to measure constructs such as ICT self-efficacy, attitudes toward technology, digital learning practices, classroom climate, professional development, and the use of ICT in teaching and learning (Fraillon et al., 2025).

Because ICILS is designed to support comparisons across countries, educational systems, and demographic subgroups, it is essential that these questionnaire scales measure the same underlying constructs equivalently across groups. Otherwise, observed differences may reflect cultural or linguistic differences in how respondents interpret the items rather than true differences in the constructs themselves. Consequently, ICILS provides an important context for examining measurement invariance and for comparing how different psychometric frameworks, such as MGCFA and MGIRT, evaluate cross-group comparability in many-group international assessment settings.

Measurement invariance under MGCFA and MGIRT

In single-group confirmatory factor analysis, the observed response of a person to an item is modelled as a linear

combination of a latent variable, an item intercept/thresholds, an item slope or factor loading, and some residual or error score. Here one set of parameters explain the relationship between the observed responses and the model parameters while in MGCFA each group has a set of unique parameters (Byrne, 2012). To establish MI, the equivalence of the parameters across groups can be tested. More specifically, configural MI is established by fitting a MGCFA model where the same number of factors and pattern loadings is imposed across the groups while the slope and intercept parameters are freely estimated. The fit of this model is examined with the familiar indices such RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), and TLI (Tucker–Lewis Index). If the model fits, the first level of MI, i.e., configural invariance is established.

In the next stage, equality of factor loadings is imposed across all the groups. Then, the fit of this constrained model is assessed against the fit of the configural model, where all parameters were estimated freely. If the fit of the constrained model does not worsen substantially, metric invariance is established. The decline in fit is evaluated by examining the differences between the two models' CFI and RMSEA denoted to as ΔCFI and ΔRMSEA. Several cut-off criteria are suggested for ΔCFI and ΔRMSEA to establish metric MI (see Svetina et al., 2020 for an overview). And finally, to evaluate scalar invariance, both slope and intercept parameters are constrained across the groups and the differences in fit indices between this model and the metric model are examined. A different set of cut-off values are recommended to establish scalar MI (Svetina et al., 2020).

Another framework that is increasingly used in large-scale assessments to examine measurement invariance and DIF is MGIRT (OECD, 2017; von Davier et al., 2023; Yamamoto et al., 2013). While in a single-group IRT model the probability of a correct response to an item by a person with a given ability is modelled, in MGIRT, the probability of a correct response to an item by a person selected at random from a group is modeled as a function of the ability of the group and one or more item parameters (Mislevy, 1983).

The predominant estimation technique in IRT, i.e., the Marginal Maximum Likelihood (MML) estimation assumes a normal distribution for the ability parameters, although this assumption is rarely substantiated with empirical evidence. However, the plausibility of assuming a single normal distribution across populations is questionable, particularly given that most ILSAs report substantial variability in country-level performance averages. If this heterogeneity is ignored and a single distribution is incorrectly assumed, the resulting parameter estimates become statistically inconsistent. Therefore, while MGIRT may not be pertinent under Joint or Conditional Maximum Likelihood estimation, it becomes essential under MML with normality assumptions (Baghaei and Robitzsch, 2025). Furthermore, MGIRT provides a flexible approach to solve numerous measurement problems including DIF analysis, equating, and multiple-matrix sampling in ILSAs, amongst others (Bock and Zimowski, 1997).

The primary aim of this study is not merely to demonstrate that different approaches to MI can yield different results, a finding already documented in prior methodological research, but to examine the practical consequences of applying MGCFA and MGIRT in a real ILSA context. Specifically, using ICILS 2023 questionnaire data from more than 30 educational systems and 33 scales, this study provides a large-scale empirical assessment of how MI conclusions depend on the chosen psychometric framework under operational conditions (ordinal items, many groups, heterogeneous populations).

In contrast to previous comparison studies that rely primarily on simulations or a small number of groups (e.g., Buchholz and Hartig, 2020; Kim et al., 2017), the present study contributes by (a) documenting framework-dependent MI decisions across a wide range of substantive scales, (b) evaluating their implications for cross-national comparability in ILSAs, and (c) identifying systematic patterns in the relative stringency of MGCFA and MGIRT when applied to ordinal survey data.

Methodology

Data: International Computer and Information Literacy Study

For this purpose, we use data from the International Computer and Information Literacy Study (ICILS). ICILS is a large-scale international assessment coordinated by the International Association for the Evaluation of Educational Achievement (IEA). The study investigates the development of computer and information literacy, as well as computational thinking skills, needed for effective participation in the digital age among 8th graders. ICILS adopts a comprehensive design, collecting data not only through achievement tests but also through extensive contextual questionnaires administered to students, teachers, principals, and ICT coordinators.

The present study focuses specifically on the student and teacher survey data. In each participating country a sample of about 150 schools was drawn, and one full 8th grade class of about 25 students and 15 teachers are sampled within each school, yielding comparatively large and statistically powerful datasets. The total sample, contains 132,889 students and 60,835 teachers from 31 countries and one benchmarking entity (i.e., North Rhine Westfalia in Germany)

This facilitates robust cross-national measurement invariance analyses. By contrast, only one principal and one ICT coordinator are surveyed per school, limiting the potential for measurement invariance analysis across countries. Consequently, our analyses concentrate on student and teacher responses.

Measures: student and teacher questionnaires

This study draws on multi-item questionnaire data collected from students and teachers in ICILS 2023. The focus lies on data from scales that aim to measure latent variables including attitudes, practices, and experiences related to teaching and learning with ICT. Specifically, we analyze 13 student and 20 teacher scales. Each scale is composed of 3 to 12 items each and employs Likert-type or frequency response formats with 3 to 5 response categories such as "Strongly disagree" to "Strongly agree" or from "Never" to "Always."

The13 student scales address both general school-related attitudes and ICT-specific experiences. One set of scales assesses students' general attitudes towards school and their self-efficacy, such as how confident they feel in managing academic demands. Another set of scales targets students' attitudes towards ICT in general, including their beliefs about its usefulness for everyday life and future careers. Further scales capture how students perceive ICT use in the classroom—whether they find it engaging, inclusive, and safe—and the extent to which they use digital tools to learn autonomously, collaborate with peers, or engage in creative work. For example, students were asked how often they used computers to discuss schoolwork with classmates or to design presentations.

The teacher questionnaire complements the student perspective by focusing on 20 scales on professional development, attitudes towards ICT, and reported classroom practices. Teachers responded to items about their participation in training activities, their perceived needs for further professional learning, and how often they use digital tools to support different instructional purposes. Scales also capture teachers' views on the value and challenges of integrating ICT into teaching. For instance, they rated their agreement with statements about whether ICT improves learning outcomes or increases workload (Frailon et al., 20).

Analysis

All MGCFA models were estimated using *lavaan* package (Rosseel, 2012) in R (R Core Team, 2025). Given the ordinal nature of the Likert-type items (2–5 categories), items were specified as categorical indicators and models were estimated using the Diagonally Weighted Least Squares (DWLS) estimator. Configural, metric, and scalar invariance models were specified following the standard sequence for categorical MGCFA, where scalar invariance was tested by constraining item thresholds (rather than intercepts) to equality across countries. For model identification, the latent factor mean was fixed to zero and variance to one in the reference group, and factor loadings were freely estimated in the configural model. In the metric model, factor loadings were constrained equal across groups, and in the scalar model, both loadings and thresholds were constrained. The MGIRT models (PCM and GPCM) were estimated using the TAM package (Robitzsch et al., 2025) in R with marginal maximum likelihood estimation and group-specific latent distributions.

To examine MI across the scales, categorical MGCFA was run for each scale separately. While in a single-group confirmatory factor analysis, one set of parameters explain the relationship between the observed response and the model parameters, in MGCFA each group has a set of unique parameters (Byrne, 2012). For evaluating configural invariance, MGCFA without cross group equality constraints was run. The fit of this model establishes configural invariance. For metric invariance, factor loadings were fixed to be equal across countries and the model fit was comparted to that of the configural model. For scalar invariance, both factor loadings and thresholds were constrained to be equal across countries and the fit was compared to that of the metric model.

The questionnaire items analyzed in this study were all ordinal Likert-type items with two to five response categories. In such cases, treating items as continuous can lead to biased parameter estimates and misleading fit indices, particularly when category distributions are skewed or the number of categories is limited. Therefore, categorical CFA models based on underlying latent response variables are recommended (Lubke and Muthén, 2004; Muthén, 1984: Muthén and Kaplan, 1985). In categorical CFA, observed ordinal responses are assumed to arise from an underlying continuous latent response variable that is partitioned by a set of thresholds. Consequently, model parameters differ from those in continuous CFA, i.e., instead of item intercepts, threshold parameters are estimated, and scalar invariance is evaluated through the equality of thresholds across groups rather than intercepts. This distinction is particularly important in measurement invariance testing, as the traditional hierarchy of configural, metric, and scalar invariance must be interpreted in terms of factor loadings and thresholds when ordinal indicators are used (Wu and Estabrook, 2016).

For configural invariance to hold, CFI > .95 and RMSEA <.10 were used as criteria (Gorges et al., 2017; Schermelleh-Engel et al., 2003). Differences in RMSEA and CFI were examined to establish MI or the lack thereof. The criteria suggested by Rutkowski and Svetina (2014) and Svetina et al. (2020) were followed:

- For metric Invariance: ΔRMSEA $\leq$ .05 in conjunction with a significant $\Delta\chi 2$ and a ΔCFI $\geq$ -.004
- For scalar Invariance: ΔRMSEA $\leq$ .01 in conjunction with a significant $\Delta\chi 2$ and a ΔCFI $\geq$ -.004

The cut-offs above are based on a simulation condition where the data distribution is ordinal, there is only one factor, there are 10 or 20 groups, and 600 to 6,000 participants are in each group. This is the closest condition to the operational conditions in the ICILS and other international large-scale projects.

In the next step, multiple-group partial credit model (PCM, Masters, 1982) and multiple-group generalized partial credit model (GPCM, Muraki, 1992) which are extensions of the Rasch model for polytomous items were estimated to examine DIF or measurement non-invariance for each scale. While in the PCM, item discrimination is fixed and all the items are assumed to have the same discrimination power, in the GPCM the assumption of uniform discriminating power for items is relaxed. The GPCM was estimated to enhance the comparability of IRT and CFA, as GPCM models item discrimination which is equivalent to CFA factor loading (Kamata and Bauer, 2008; Takane and de Leeuw, 1987).

To investigate DIF within the MGIRT framework, a statistic called root mean square deviation (RMSD) (Oliveri and von Davier, 2011) is examined. To compute RMSD, a MGIRT model is estimated where there is a mixture of normal population distributions (one for each group, so each has a unique mean and variance) and item parameters are constrained across groups. With this, a single two-parameter logistic item response curve (per item) that corresponds to the international item parameter (i.e., an average across all groups) is estimated. To estimate the RMSD, observed proportions of correct responses at various points along the proficiency scale from each

subpopulation (i.e., country) is compared with the two-parameter logistic item response curve that corresponds to the international item parameters (OECD, 2017).

RMSD statistic quantifies the distance between the group specific IRF (item response function) and the common international IRF (Buchholz and Hartig, 2020). An RMSD value of zero indicates a perfect fit for the item and the presence of MI. In other words, it indicates how well the group/country performance can be explained by the international joint parameters. Oliveri and von Davier (2011, 2014) recommended an RMSD value of greater than.10 as a cutoff criterion to identify DIF items. Nevertheless, in PISA 2015, cutoff values of.12 and.30 were used for the cognitive and non-cognitive scales, respectively (Khorramdel et al., 2020; OECD, 2017). RMSD values greater than the cutoff value indicate non-invariance. When this occurs for an item, the equality constraints for the noninvariant items can be relaxed. In this case, the item has group-specific item parameters while the other items have joint parameters. This is equivalent to partial invariance in MGCFA which allows comparison across groups with theta parameters that are on the same scale (Byrne, 2012; Oliveri and von Davier, 2014; Sandoval-Hernandez et al., 2025).

In MGIRT, item parameters are constrained to be equal across groups (i.e., countries) but the latent ability distribution is allowed to vary for each group. RMSD item fit statistics (Oliveri and von Davier, 2011) (the distance between the country specific ICC and the joint international ICC) was estimated as a measure of DIF or measurement non-invariance. Furthermore, residual-based infit and outfit mean square values (Wright and Masters, 1982) were also obtained. An RMSD value of > .25 was set as a criterion to flag misfitting items (Chen et al., 2025; OECD, 2017). The boundary of.70 to 1.30 was considered as the acceptable range for the infit and outfit values (Bond et al., 2020).

Results

Student scales

Table 1 shows the CFA configural model statistics and IRT item fit values for the 13 student scales. As Table 1 shows, all the items fit the PCM and GPCM based on the infit and outfit values. The RMSD values for all the items and across all the countries, are below the cutoff criterion of.25 which means that the scales are invariant. There is only one exception which is limited to a single item in one country where the RSMD exceeds 0.25. This is consistent for both PCM and GPCM. It should be noted that while RMSDs do not show a noticeable difference between the fits of GPCM and PCM, infit and outfit values are closer to their perfect value (i.e., one) in the GPCM model. That is, items have a better fit in the GPCM compared to the PCM.

In contrast, for the MGCFA, the RMSEA values for 8 scales are greater than.10 which means that they do not even satisfy the conditions for configural invariance according to the CFA criteria, but the rest of the scales do.

Table 2 shows the ΔCFI and ΔRMSEA for the metric (constrained factor loadings) and scalar (constrained factor loadings and thresholds) models for the 13 student scales. CFA metric, and scalar invariance of the scales were evaluated using Rutkowski and Svetina's (2014) criteria (see above). Metric and scalar invariance were not evaluated for the scales that were not configural invariant and these scales were eliminated from further examination (within the CFA framework).

The ΔCFI and ΔRMSEA values for the metric models showed that metric invariance can be established for only one scale. The ΔCFI and ΔRMSEA values for the scalar models showed that none of the 13 student scales achieve scalar invariance. Nevertheless, IRT-based RMSD item fit statistics showed that all

TABLE 1 CFA configural model statistics and IRT item Fit values for the student scales.

Scale	CFA		PCM			GPCM		
	CFI	RMSEA	RMSD Range	Infit Range	Outfit Range	RMSD Range	Infit Range	Outfit Range
S_ACMULT	0.972	0.115	.016–.119	.902–1.179	.857–1.233	.019–.111	1.015–1.039	.957–1.041
S_CTCLS	0.937	0.142	.024–.135	.926–1.061	.899–1.035	.022–.128	1.005–1.035	.973–.996
S_GENCLASS	0.996	0.059	.03–.201	.882–1.096	.878–1.094	.021–.192	1.003–1.026	.981–1.003
S_GENEFF	0.961	0.086	.017–.17	.843–1.241	.771–1.247	.013–.141	1–1.033	.976–1.045
S_ICTFUT	0.983	0.129	.036–.129	.881–1.206	.862–1.174	.023–.132	1.02–1.032	.986–1.013
S_ICTNEG	0.978	0.112	.019–.14	.956–1.072	.933–1.075	.018–.148	1.015–1.024	.981–1.023
S_ICTPOSG	0.984	0.155	.025–.112	.932–.968	.915–.953	.019–.106	1.007–1.02	.983–.993
S_ICTPOSS	0.990	0.096	.019–.116	.926–1.067	.898–1.057	.017–.117	1.011–1.032	.972–1.032
S_LRNINTO	0.953	0.104	.017–.169	.98–1.081	.966–1.078	.014–.163	1.017–1.036	1.001–1.012
S_LRNINTS	0.949	0.123	.018–.16	.965–1.065	.95–1.058	.02–.148	1.013–1.022	1.001–1.01
S_LRNSAFE	0.986	0.217	.019–.173	.901–1.085	.869–1.08	.017–.176	1.013–1.023	.951–1.032
S_SPECEFF	1.000	0.000	.026–.123	.893–1.064	.881–1.059	.014–.137	1.002–1.018	.975–1.009
S_SPECLASS	0.988	0.057	.024–.25	.858–1.158	.82–1.199	.019–.252	1.006–1.027	1.002–1.071

RMSD range is across all items and countries. S_SPECEFF contains only three items. For three-item scales, the configural one-factor CFA model may be just-identified; therefore, CFI = 1.000 and RMSEA = 0.000 do not indicate perfect substantive fit.

TABLE 2 CFA and IRT invariance statistics for the student scales.

Scale	Delta CFI Metric	Delta RMSEA Metric	Delta CFI Scalar	Delta RMSEA Scalar	Countries >.25 PCM	Countries >.25 GPCM	Configural MI	Metric MI	Scalar MI	PCM MI	GPCM MI
S_ACMULT	−.001	−.02	−.037	.002	0	0	No	–	–	Yes	Yes
S_CTCLS	.011	−.028	−.022	−.002	0	0	No	–	–	Yes	Yes
S_GENCLASS	−.016	.03	−.095	.049	0	0	Yes	No	–	Yes	Yes
S_GENEFF	.007	−.016	−.026	.009	0	0	Yes	Yes	No	Yes	Yes
S_ICTFUT	0	−.031	−.019	.003	0	0	No	–	–	Yes	Yes
S_ICTNEG	−.002	−.038	−.054	.012	0	0	No	–	–	Yes	Yes
S_ICTPOSG	−.005	−.043	−.011	−.022	0	0	No	–	–	Yes	Yes
S_ICTPOSS	−.008	−.014	−.031	.005	0	0	Yes	No	–	Yes	Yes
S_LRNINTO	.006	−.026	−.073	.02	0	0	No	–	–	Yes	Yes
S_LRNINTS	.007	−.031	−.077	.023	0	0	No	–	–	Yes	Yes
S_LRNSAFE	.003	−.094	−.016	.002	0	0	No	–	–	Yes	Yes
S_SPECEFF	−.003	.052	−.036	.055	0	0	Yes	No	–	Yes	Yes
S_SPECLASS	−.019	.022	−.025	.006	1	1	Yes	No	–	No	No

10.3389/feduc.2026.1823761

the scales, except one, are invariant under both PCM and GPCM. Table 2 shows that there was only one scale in one country with an RMSD>.25. Because RMSD is affected by deviations in item response functions, including differences related to item discrimination and threshold/difficulty parameters, low RMSD values provide evidence that a common MGIRT item parameterization adequately describes group-specific response patterns. However, they should not be interpreted as a direct test of metric or scalar invariance in the strict MGCFA sense. Model fit indices for the metric and scalar models along with the scale alpha reliabilities are presented in the Supplementary Materials.

Teacher scales

Table 3 shows the CFA configural model statistics and IRT item fit values for the 20 teacher scales. Table 3 shows that all the items fit the PCM and GPCM based on their infit and outfit values. The partial credit model RMSD values for all the items and across all the countries are below the cutoff criterion of.25 which means that all the scales are invariant according to the multiple-group PCM. Under the multiple-group GPCM, only a few items in one scale and in one country had RMSDs greater than.25. Therefore, only one scale is noninvariant based on multiple-group GPCM. Similar to the students' scales, the RMSDs are not noticeably different across GPCM and PCM but infit and outfit values are closer to their perfect value in the GPCM model. Here again we observed that the items have a better fit to the GPCM compared to the PCM. With respect to the MGCFA, the RMSEA values for 11 scales are greater than.10 which indicates that they do not satisfy the conditions for configural invariance according to the CFA criteria.

Table 4 shows the ΔCFI and ΔRMSEA for the metric and scalar models for the 20 teacher scales. The ΔCFI and ΔRMSEA values for the metric models showed that metric invariance can only be established for five scales. They also showed that scalar invariance can be established for only one scale. Nevertheless, IRT-based RMSD item fit statistics showed that all the scales are invariant under the PCM and all but one are invariant under the GPCM.

Discussion

This study provides a large-scale empirical evaluation of how measurement invariance decisions differ across psychometric frameworks in an operational international large-scale assessment (ICILS 2023), rather than a purely methodological or simulation-based comparison. In this study, we compared two widely used methods for measurement invariance analysis: MGCFA and MGIRT. We adhered to established procedures

TABLE 3 CFA configural model statistics and IRT item Fit values for the teacher scales.

Scale	CFA		PCM			GPCM		
	CFI	RMSEA	RMSD Range	Infit Range	Outfit Range	RMSD Range	Infit Range	Outfit Range
T_CODEMP	0.981	0.118	.031–.164	.822–1.088	.789–1.066	.026–.172	1.019–1.064	.976–1.039
T_COLICT	0.975	0.157	.035–.139	.771–1.037	.749–1.001	.019–.134	1.002–1.046	.924–1.004
T_EMPITE	0.996	0.062	.016–.122	.674–.918	.456–.85	.007–.136	.983–1.017	.712–.998
T_EPICOG	0.985	0.127	.033–.135	.894–1.104	.867–1.108	.031–.132	.986–1.068	.976–1.061
T_EPICON	0.998	0.048	.026–.113	.835–1.189	.83–1.194	.015–.109	1.001–1.056	.982–1.061
T_EPIEMC	0.995	0.089	.031–.132	.853–1.252	.831–1.263	.025–.139	.982–1.073	.974–1.079
T_ICTCLASA	0.989	0.084	.024–.138	.795–1.131	.676–1.133	.02–.139	1.017–1.067	.918–1.117
T_ICTCLASB	0.982	0.110	.017–.138	.848–1.125	.819–1.114	.019–.138	1.014–1.09	.973–1.075
T_ICTEFF	0.949	0.108	.027–.228	.832–1.172	.801–1.194	.019–.229	.988–1.02	.93–1.024
T_ICTEMP	0.962	0.113	.025–.174	.832–1.186	.808–1.187	.023–.172	1.019–1.055	.974–1.049
T_PROFLRN	0.970	0.107	.025–.164	.797–1.188	.678–1.317	.015–.176	.978–1.039	.916–1.025
T_PROFNEED	0.986	0.076	.017–.191	.757–1.027	.567–1	.006–.176	.985–1.027	.813–.959
T_RESRC	0.985	0.116	.035–.197	.756–1.274	.746–1.287	.024–.209	.989–1.026	.951–1.04
T_TEAPEX	1.000	0.000	.024–.168	.832–.98	.827–.929	.012–.185	1.001–1.027	.934–1.022
T_TEAPIN	0.996	0.079	.032–.135	.775–1.038	.756–1.018	.019–.138	.972–1.052	.948–1.02
T_TEAPTC	0.961	0.145	.022–.236	.896–1.164	.878–1.165	.022–.239	1.017–1.042	.989–1.045
T_USETOOL	0.988	0.053	.017–.146	.829–1.108	.778–1.107	.015–.146	.967–1.032	.99–1.113
T_USEUTIL	0.966	0.110	.019–.195	.851–1.279	.84–1.292	.025–.204	1.007–1.045	.967–1.049
T_VWNEG	0.974	0.096	.027–.247	.836–1.231	.827–1.232	.026–.266	.976–1.053	.942–1.053
T_VWPOS	0.959	0.125	.029–.141	.856–1.159	.839–1.141	.02–.131	.995–1.025	.972–1.003

RMSD range is across all items and countries. T_TEAPEX contains only three items. For three-item scales, the configural one-factor CFA model may be just-identified; therefore, CFI = 1.000 and RMSEA = 0.000 do not indicate perfect substantive fit.

TABLE 4 CFA and IRT invariance statistics for the teacher scales .

Scale	Delta CFI Metric	Delta RMSEA Metric	Delta CFI Scalar	Delta RMSEA Scalar	Countries >.25 PCM	Countries >.25 GPCM	Configural MI	Metric MI	Scalar MI	PCM MI	GPCM MI
T_CODEMP	.005	−.035	−.027	.023	0	0	No	–	–	Yes	Yes
T_COLICT	.005	−.035	−.005	−.012	0	0	No	–	–	Yes	Yes
T_EMPITE	0	−.011	−.002	.014	0	0	Yes	Yes	No	Yes	Yes
T_EPICOG	.001	−.035	−.022	.003	0	0	No	–	–	Yes	Yes
T_EPICON	−.006	.018	−.036	.025	0	0	Yes	No	–	Yes	Yes
T_EPIEMC	−.002	−.01	−.019	.018	0	0	Yes	Yes	No	Yes	Yes
T_ICTCLASA	−.006	.001	0	−.027	0	0	Yes	No	–	Yes	Yes
T_ICTCLASB	−.006	−.013	−.029	−.006	0	0	No	–	–	Yes	Yes
T_ICTEFF	.015	−.025	−.043	.024	0	0	No	–	–	Yes	Yes
T_ICTEMP	.012	−.027	−.021	.014	0	0	No	–	–	Yes	Yes
T_PROFLRN	.005	−.022	−.027	.025	0	0	No	–	–	Yes	Yes
T_PROFNEED	0	−.011	−.003	.009	0	0	Yes	Yes	Yes	Yes	Yes
T_RESRC	−.001	−.017	−.014	.004	0	0	No	–	–	Yes	Yes
T_TEAPEX	−.004	.069	−.025	.036	0	0	Yes	No	–	Yes	Yes
T_TEAPIN	−.002	−.006	−.009	.01	0	0	Yes	Yes	No	Yes	Yes
T_TEAPTC	−.005	−.028	−.122	.044	0	0	No	–	–	Yes	Yes
T_USETOOL	−.005	0	−.031	.016	0	0	Yes	No	–	Yes	Yes
T_USEUTIL	−.004	−.016	−.056	.018	0	0	No	–	–	Yes	Yes
T_VWNEG	−.001	−.011	−.026	.011	0	1	Yes	Yes	No	Yes	No
T_VWPOS	.003	−.024	−.021	−.001	0	0	No	–	–	Yes	Yes

10.3389/feduc.2026.1823761

for testing measurement invariance within both CFA and IRT frameworks and applied commonly accepted cut-off criteria for model evaluation. Rather than formally assessing the absolute performance of each approach or relying on simulated data, our goal was to compare their outcomes by applying both methods to a large international dataset, using each as a reference for the other. This allowed us to examine the consistency of the results produced by MGCFA and MGIRT in a real-world application. Several key points emerged from this investigation.

Both the PCM and the GPCM were employed within the MGIRT framework. Although RMSD values did not significantly differ between PCM and GPCM, infit and outfit statistics demonstrated better item fit in the GPCM. This aligns with expectations, as GPCM captures the variations in discriminations by allowing items to have different slopes. The results from both methods underscore the practical challenges associated with establishing MI in international contexts like ICILS, where demographic, cultural, and educational differences across countries introduce complexities in psychometric modelling.

The results showed that while MGIRT demonstrated strong invariance across most scales (with low RMSD values), MGCFA was more stringent, suggesting even configural non-invariance for more than half of the scales based on the RMSEA and CFI criteria. Specifically, 8 out of 13 student scales and 11 out of 20 teacher scales exhibited RMSEA values above the.10 cutoff in MGCFA, indicating poor fit of the configural model.

A notable divergence between the two methods occurred when examining metric and scalar invariance. Using the selected global fit criteria, MGCFA did not support metric and scalar invariance for most scales, with only one student scale and five teacher scales showing metric invariance, and no student scales and only one teacher scale achieving scalar invariance. Conversely, MGIRT-based RMSD statistics indicated that almost all the scales (both student and teacher) were invariant across countries, except in very rare cases.

The rejection of configural invariance in several scales in this study can partly be due to model complexity. Estimating exact many-group CFA models with ordinal indicators across 32 educational systems simultaneously is very complex (Asparouhov and Muthén, 2014). Previous methodological research has noted that traditional exact invariance approaches become increasingly restrictive in large international comparisons because global fit indices accumulate numerous small sources of misfit across groups and items (Greiff and Scherer, 2018; Kim et al., 2017; Rutkowski and Svetina, 2014). Consequently, reduced configural model fit does not necessarily imply large substantive differences in item functioning, but may also reflect the sensitivity of global model evaluation in highly heterogeneous many-group settings.

Although both MGCFA and MGIRT are used to evaluate measurement invariance, they operate at different levels of analysis and test conceptually distinct hypotheses. MGCFA evaluates hierarchical, model-level hypotheses regarding the equivalence of factor structure, factor loadings, and item thresholds across groups using nested model comparisons. In contrast, MGIRT-based DIF analyses, such as RMSD statistic, assess the degree to which item response functions deviate from a common international item parameterization, providing item-level diagnostics of model misfit rather than formal tests of configural, metric, or scalar invariance. Consequently, low RMSD values cannot be interpreted as direct evidence of configural, metric, or scalar invariance in the strict CFA sense, as they do not test factor structure equivalence or impose hierarchical equality constraints. Instead, RMSD indicates the extent to which a common measurement model adequately reproduces group-specific item response patterns. Therefore, the present study does not treat MGCFA and MGIRT as testing identical hypotheses but rather compares the practical conclusions about cross-group comparability that emerge when two widely used but conceptually different invariance evaluation frameworks are applied to the same large-scale assessment data.

Our MGCFA findings partly contradict and partly align with previous MI studies in ILSAs. They contradict prior research in that earlier studies consistently established at least configural invariance, whereas our results show that configural invariance cannot be established for a substantial proportion of scales when applying MGCFA (more than half of the scales in our study). At the same time, our findings align with the literature in demonstrating that higher levels of invariance are considerably more difficult to achieve, especially as the number of groups increases. The reason for the inconsistency might be that most prior studies were mostly based on a few groups. For example, Gorges et al. (2017) found configural and largely metric and (full or partial) scalar invariance in small group comparisons (2–3 groups). Courtney et al. (2023) reported full invariance across gender and configural invariance across 9 countries (36 pairwise comparisons), but very limited metric (25%) and scalar (2.8%) invariance cross-nationally, while Jusufi et al. (2026) found only partial scalar invariance when comparing two regions. In contrast, Ding et al. (2023), using 80 groups (40 countries×2 cycles), found metric but not scalar invariance, highlighting the difficulty of achieving full invariance in many-group contexts.

These studies suggest a consistent pattern: configural invariance is typically established while metric invariance is less stable and scalar invariance is rarely achieved, especially in cross-national or many-group settings. In this respect, our findings diverge from previous research by showing that even configural invariance frequently fails under MGCFA in a many-group ILSA setting, while at the same time reinforcing the well-established conclusion that metric and scalar invariance constitute the primary challenges for cross-group comparability.

While it is widely accepted that establishing measurement invariance is a crucial step in comparative research, our findings reveal that the choice of the method, MGCFA or MGIRT, can lead to substantially different conclusions. Specifically, we observed that MGCFA tends to reject measurement invariance more frequently, whereas MGIRT often suggests comparability across groups. Although our study was not intended to determine which approach yields the "correct" result, the inconsistencies between these methods are too significant to ignore.

The contribution of this study is not to demonstrate that different MI methods may disagree but to show the magnitude and practical implications of such disagreement when applied to a large set of ordinal questionnaire scales across more than 30 educational systems in an operational international assessment.

While the study does not isolate the mechanisms driving these discrepancies, it highlights that conclusions about cross-national comparability can vary substantially depending on the chosen psychometric framework and decision criteria. Therefore, our findings point more specifically to the need for clearer theoretical guidance on how to interpret and reconcile invariance and DIF evidence in many-group, large-scale assessment contexts.

The present findings also have implications for validity arguments in international comparative assessments. Measurement invariance is closely linked to the validity of cross-group score interpretations because meaningful comparisons across countries depend on the assumption that the underlying constructs are measured equivalently. If different psychometric frameworks lead to different conclusions regarding invariance, then the validity evidence supporting cross-national comparisons may also vary depending on the chosen analytical approach. In this sense, the present study highlights that validity in large-scale international assessments is not only a substantive issue but also a methodological one, shaped by how measurement equivalence is conceptualized and evaluated. Consequently, decisions regarding MGCFA or MGIRT are relevant not only for statistical modeling but also for the interpretability and fairness of international comparisons.

Our findings show that choosing between MGCFA and MGIRT is not about selecting a single correct method but about balancing their different strengths and limitations. MGCFA offers a strict assessment of overall construct comparability across groups, though it can be overly sensitive in large, diverse international datasets with ordinal data, detecting many minor deviations. In contrast, MGIRT focuses on item-level functioning and is better suited for evaluating whether differential item functioning meaningfully affects score comparability in practice. Therefore, the approaches are best used together, with MGCFA assessing global equivalence and MGIRT identifying specific item-level issues with substantive impact.

Despite its contributions, this study has several limitations that warrant consideration. First, our study revealed that the results of MGIRT and MGCFA are inconsistent when applying commonly used evaluation criteria and cut-off values for measurement invariance testing. This inconsistency highlights a critical issue in the interpretation of measurement invariance across frameworks. However, it is important to emphasize that our findings do not imply that one approach is inherently correct or incorrect. Both methods are grounded in rigorous theoretical foundations and have been widely applied in the literature. Nonetheless, the observed differences between the two are striking and underscore the need for further investigation to better understand the sources and implications of these discrepancies. Second, the specific cutoffs used for determining MI may influence the results, and alternative cutoffs could yield different conclusions. Third, while the study focused on ICILS data, the generalizability of the findings to other large-scale assessments, such as PISA or TIMSS, remains to be explored. Future research could replicate this comparison across different datasets to further validate the findings. Additionally, exploring the impact of different sample sizes, item types, and levels of group heterogeneity on the performance of these methods would provide a more comprehensive understanding of their strengths and weaknesses.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.iea.nl/data-tools/repository/icils.

Author contributions

PB: Methodology, Writing – review & editing, Writing – original draft. RS: Writing – original draft, Conceptualization, Writing – review & editing. AC: Writing – original draft, Formal analysis, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work 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) declared that generative AI was not used in the creation of this manuscript.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2026.1823761/full#supplementary-material

References

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* 21 (4), 495–508. doi: 10.1080/10705511.2014.919210

Baghaei, P., and Robitzsch, A. (2025). A tutorial on item response modeling with multiple groups using TAM. *Educ. Methods Psychometr.* 3, 14. doi: 10.61186/emp.2025.1

Bock, R. D., and Zimowski, M. F. (1997). "Multiple group IRT," in *Handbook of Modern Item Response Theory*, eds. W. J. van der Linden, and R. K. Hambleton (New York: Springer), 433–448. doi: 10.1007/978-1-4757-2691-6_25

Bond, T. G., Zi, Y., and Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 4th edn. New York: Routledge.

Buchholz, J., and Hartig, J. (2020). Measurement invariance testing in questionnaires: a comparison of three multigroup-CFA and IRT-based approaches. *Psychol. Test. Assess. Model.* 62 (1), 29–53.

Byrne, B. M. (2012). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. New York: Routledge.

Chen, M., Christiansen, A., Rožman, M., and Strietholt, R. (2025). "Scaling procedures for ICILS 2023 questionnaire items," in *IEA International Computer and Information Literacy Study 2023: Technical Report*, eds. J. Fraillon, M. Rožman, S. Meyer, L. Musu, Y.-L. Liaw, A. Christiansen, and S. Tieck (Hamburg: IEA), 209–267.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front. Psychol.* 5, 982. doi: 10.3389/fpsyg.2014.00982

Courtney, M. G. R., Hernández-Torrano, D., Karakus, M., and Singh, N. (2023). Measuring student well-being in adolescence: proposal of a five-factor integrative model based on PISA 2018 survey data. *Large-Scale Assess. Educ.* 11, 20. doi: 10.1186/s40536-023-00170-y

Ding, Y., Yang Hansen, K., and Klapp, A. (2023). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *Eur. J. Psychol. Educ.* 38, 709–732. doi: 10.1007/s10212-022-00623-y

Fraillon, J., and Rožman, M. (2023). *IEA International Computer and Information Literacy Study 2023: Assessment Framework*. Cham: Springer. doi: 10.1007/978-3-031-61194-0

Fraillon, J., Rožman, M., Meyer, S., Musu, L., Liaw, Y.-L., Christiansen, A., and Tieck, S. (2025). *IEA International Computer and Information Literacy Study 2023: Technical Report*. Hamburg: IEA.

Gorges, J., Koch, T., Maehler, D. B., and Offerhaus, J. (2017). Same but different? Measurement invariance of the PIAAC motivation-to-learn scale across key socio-demographic groups. *Large-Scale Assess. Educ.* 5, 13. doi: 10.1186/s40536-017-0047-5

Greiff, S., and Scherer, R. (2018). Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing. *Eur. J. Psychol. Assess.* 34 (3), 141–144. doi: 10.1027/1015-5759/a000487

Jusufi, D., Yavuz Temel, G., and Schwippert, K. (2026). Books as educational resource? Assessing cross-country measurement invariance of the TIMSS home resources for learning scale in post-conflict societies. *Large-Scale Assess. Educ.* 14, 5. doi: 10.1186/s40536-025-00269-4

Kamata, A., and Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Struct. Equ. Modeling.* 15 (1), 136–153. doi: 10.1080/10705510701758406

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.

Kim, E. S., Cao, C., Wang, Y., and Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Struct. Equ. Model. Multidiscip. J.* 24 (4), 524–544. doi: 10.1080/10705511.2017.1304822

Kim, E. S., and Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Struct. Equ. Model. Multidiscip. J.* 18 (2), 212–228. doi: 10.1080/10705511.2011.557337

Lubke, G. H., and Muthén, B. O. (2004). Applying multiple group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Struct. Equ. Modeling.* 11 (4), 514–534. doi: 10.1207/s15328007sem1104_2

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika.* 47 (2), 149–174. doi: 10.1007/BF02296272

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care.* 44 (11, Suppl 3), S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Mislevy, R. J. (1983). Item response models for grouped data. *J. Educ. Stat.* 8, 271–288. doi: 10.3102/10769986008004271

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16 (2), 159–176. doi: 10.1177/014662169201600206

Muthén, B. O. (1984). A general structural equation model for dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika.* 49, 115–132. doi: 10.1007/BF02294210

Muthén, B. O., and Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *Br. J. Math. Stat. Psychol.* 38, 171–189. doi: 10.1111/j.2044-8317.1985.tb00832.x

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

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

Oliveri, M. E., and von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *Int. J. Test.* 14 (1), 1–21. doi: 10.1080/15305058.2013.825265

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (Accessed August 18, 2025).

Robitzsch, A., Kiefer, T., and Wu, M. (2025). TAM: Test Analysis Modules. R package version 4.3-25. Available online at: https://CRAN.R-project.org/package=TAM

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48 (2), 1–36. doi: 10.18637/jss.v048.i02

Rutkowski, L., and Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educ. Psychol. Meas.* 74, 31–57. doi: 10.1177/0013164413498257

Sandoval-Hernandez, A., Carasco, D., and Eryilmaz, N. (2025). Alignment optimization in international large-scale assessments: a scoping review and future directions. *Educ. Methods Psychom.* 3, 16. doi: 10.61186/emp.2025.3

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: test of significance and descriptive goodness-of-fit measures. *Methods Psychol. Res. Online.* 8 (2), 23–74.

Svetina, D., Rutkowski, L., and Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using Mplus and the lavaan/semTools packages. *Struct. Equ. Model. Multidiscip. J.* 27 (1), 111–130. doi: 10.1080/10705511.2019.1602776

Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika.* 52 (3), 393–408. doi: 10.1007/BF02294363

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the MI literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods.* 3, 4–69. doi: 10.1177/109442810031002

von Davier, M., Mullis, I. V. S., Fishbein, B., and Foy, P. (2023). Methods and procedures: PIRLS 2021 technical report. Boston College, TIMSS and PIRLS International Study Center. Available online at: https://pirls2021.org/methods

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: Mesa Press.

Wu, A. D., Li, Z., and Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Pract. Assess. Res. Eval.* 12 (3), 1–26. doi: 10.7275/mhqa-cd89

Wu, H., and Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika.* 81 (4), 1014–1045. doi: 10.1007/s11336-016-9506-0

Yamamoto, K., Khorramdel, L., and von Davier, M. (2013). "Scaling PIAAC cognitive data," in *Technical Report of the Survey of Adult Skills (PIAAC), 2nd edn.* (Paris: OECD), 1–33. Available online at: https://www.oecd.org/skills/piaac/technical_report_2nd_edition_chapters_17-23.pdf