Studies in Educational Evaluation 87 (2025) 101519

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

A critical evaluation of alignment optimization for improving cross-national comparability in international large-scale assessments

Andres Sandoval-Hernández [a], Diego Carrasco [b], Nurullah Eryilmaz [c,d,*]

[a] University of Bath, UK
[b] Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile
[c] IEA Hamburg, Germany
[d] Leuphana University, Luneburg, Germany

ARTICLE INFO

Keywords:
TALIS 2018
Alignment optimization
Measurement invariance
Cross-national comparability
Principal scales
Teacher scales
Multiple-group confirmatory factor analysis (MGCFA)
International large-scale assessments (ILSAs)
Educational measurement

ABSTRACT

This study critically examines the use of alignment optimization to improve cross-national comparability of teacher and principal scales from the Teaching and Learning International Survey (TALIS) 2018. By investigating key psychometric properties, including dimensionality, reliability, and measurement invariance, the study highlights critical challenges in international large-scale assessments. While unidimensionality and high internal consistency were established for all scales, traditional multiple-group confirmatory factor analysis (MGCFA) suggested that scalar invariance could not be fully established for most scales, raising concerns about the robustness of cross-national comparisons under strict invariance assumptions. In contrast, alignment optimization emerged as a flexible and robust method, significantly enhancing the comparability of principal scales, all of which met alignment criteria. However, persistent challenges were identified for many teacher scales, which fell below alignment thresholds, emphasizing unresolved methodological complexities. This study demonstrates the transformative potential of alignment optimization for advancing psychometric rigor in global educational research and underscores the need for innovative approaches to address lingering comparability issues in international assessments.

1. Introduction

Ensuring cross-national comparability in International Large-Scale Assessments (ILSAs) is a cornerstone for understanding global educational trends and informing policy. With the growing participation of diverse countries, the demand for robust methodologies to ensure valid cross-country comparisons has intensified (Treviño et al., 2021). ILSAs, such as those led by the OECD and other international bodies (2019), rely heavily on contextual questionnaires to capture nuanced insights into the educational environments that shape outcomes at the student, teacher, and school levels (Senden et al., 2023). These questionnaires not only contextualize achievement scores but also serve as tools for evaluating the equity and effectiveness of educational systems globally.

The pursuit of comparability in ILSAs, however, is fraught with methodological challenges. Multi-Group Confirmatory Factor Analysis (MGCFA), the conventional approach to establishing measurement invariance, often falls short when applied to complex, culturally diverse datasets. Its inability to consistently achieve scalar invariance across groups limits the validity of cross-national comparisons and undermines the interpretability of findings. In response to these limitations, alignment optimization has emerged as a promising alternative, offering greater flexibility through the concept of approximate invariance. Despite its potential, the application of alignment optimization in ILSAs remains uneven, with most studies focusing predominantly on student-level data. These studies, while illuminating, have yielded mixed results, highlighting the need for more comprehensive investigations that address both their strengths and limitations.

This study seeks to bridge this gap by critically examining the application of alignment optimization using data from the Teaching and Learning International Survey (TALIS) 2018. By focusing on teacher and principal scales—key dimensions of educational systems—we aim to provide empirical evidence on the efficacy of alignment methods in enhancing cross-national comparability. We note that invariance alignment (IA) is also referred to as alignment optimization, and we have used these terms interchangeably, as seen in previous studies (Robitzsch, 2023; Cieciuch et al., 2018; Pokropek et al., 2019). Beyond

* Corresponing author at: IEA Hamburg, Germany.
E-mail address: nurullah.eryilmaz@leuphana.de (N. Eryilmaz).

https://doi.org/10.1016/j.stueduc.2025.101519
Received 15 December 2024; Received in revised form 14 September 2025; Accepted 17 September 2025
Available online 22 September 2025

assessing their utility, we explore the boundaries of these methods, identifying unresolved challenges and areas for methodological refinement. By situating our analysis within the broader discourse on ILSA methodologies, this study contributes to advancing the rigor and interpretability of international educational research.

1.1. Multi-group confirmatory factor analysis (MGCFA)

Multiple Group Confirmatory Factor Analysis (MGCFA) is a fundamental method used in structural equation modeling (SEM) to assess measurement invariance (MI) across different groups (Joreskog, 1971; Van de Vijver et al., 2019). The method allows researchers to simultaneously fit a predefined factor structure to various groups, facilitating comparisons of measurement parameters such as factor loadings and intercepts.

Multiple Group Confirmatory Factor Analysis (MGCFA) is widely recognized as one of the most effective techniques for testing measurement invariance (MI) in structural equation modeling (SEM). While there are various approaches to examine MI (Braun & Johnson, 2010; see also Kim et al., 2017; Davidov et al., 2018a, b), MGCFA remains the most discussed method in the literature (Jöreskog, 1971; Sörbom, 1974).

The MGCFA model is structured as follows: let $i = 1, \ldots, n$ denote individual observations, $p = 1, \ldots, P$ the number of observed indicators, $g = 1, \ldots, G$ the groups to be compared, and $\mathbf{y}_{ig}$ a $P\,x\,1$ vector of observed indicator scores for individual i in group g. The model further includes:

- Λg: a P × 1 group-specific vector of factor loadings,
- νg: a vector of intercepts,
- ηig: the latent factor scores for individual i in group g,
- ϵig: a P x 1 vector of residual terms assumed to follow N (0, Θg), where Θg = Cov(ϵg, ϵ′g).

The measurement model can be expressed as:
yig = νg + Λgηig + ϵig.
The mean and covariance structure (MACS) for the model is given by:
μg = νg + Λgαg,
Cov(yg, yg′) = Σg = ΛgΨgΛ′g + Θg.
In these equations:

- $\boldsymbol{\mu}g$ represents the observed means of $\mathbf{y}_g$,
- $\boldsymbol{\alpha}_g$ and $\boldsymbol{\Psi}_g$ denote the latent means and variances of η_g,
- $\eta_g \sim N\,(\boldsymbol{\alpha}_g, \boldsymbol{\Psi}_g)$.

MGCFA tests MI hierarchically, beginning with configural invariance, which evaluates whether the same factor structure is applicable across groups. The next level, metric invariance, examines whether factor loadings are equivalent across groups, ensuring comparability of factor scores. Finally, scalar invariance, which involves the equality of item intercepts across groups, is considered a necessary condition for valid comparisons of latent means (Chen et al., 2008; Chen, 2007; Vandenberg & Lance, 2000). However, recent studies suggest that scalar invariance is not always necessary to ensure comparability. While traditionally considered essential, newer research argues that meaningful comparisons can still be made under conditions of partial or approximate invariance (Robitzsch & Lüdtke, 2023; Funder & Gardiner, 2024; Welzel & Inglehart, 2016; Liu et al., 2017). This perspective is particularly relevant for international large-scale assessments, where strict scalar invariance is often difficult to achieve due to cultural and linguistic differences.

While MGCFA is widely used, achieving scalar invariance can be challenging in studies with large numbers of groups. Small deviations in parameter estimates may result in significant model misfit due to the stringent assumptions underlying MGCFA. These challenges are especially pronounced in international large-scale assessments (ILSAs), where cultural and linguistic differences between groups often lead to non-invariance. As a result, researchers have sought alternative methods, such as alignment optimization, to address these limitations (Thissen, 2024).

For instance, the TALIS 2018 study applied this method to assess the cross-cultural comparability of teacher and principal scales across participating countries. The analysis revealed that only one teacher scale and one principal scale achieved scalar-level invariance, underscoring the challenges of achieving full comparability in large-scale assessments (refer to Table 9 and Figs. X and Y in the Appendix for detailed results).

1.2. The alignment method

The alignment method, developed by Asparouhov and Muthén (2014), addresses many of the challenges associated with traditional MGCFA. Unlike MGCFA, which requires strict invariance for meaningful comparisons, alignment optimization operates under the principle of approximate MI. This approach enables the identification of meaningful latent mean differences across groups while tolerating minor non-invariances in measurement parameters.

Let's provide an overview of the alignment method used to estimate measurement invariance. For illustration purposes, we employ a multiple-group factor analysis model with a single latent factor η, which is measured by $p = 1, \ldots, P$ observed indicators across $g = 1, \ldots, G$ groups. Let y_{ipg} denote the p-th observed indicator for individual i in group g. The factor model is expressed as:

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\eta_{ig} + \epsilon_{ipg}$$

In this equation:

1.3. ν_{pg} and λ_{pg} represent the intercept and loading parameters, respectively

- $\epsilon_{ipg} \sim N\,(0,\ \theta_{pg})$ is the residual term.
- $\eta_{ig} \sim N\,(\alpha_g,\ \psi_g)$ is the latent factor for individual i in group g.

The alignment method estimates all group-specific parameters, including νpg, λpg, αg, ψg, and θpg. Unlike traditional methods that assume exact measurement invariance (MI), the alignment method relies on approximate MI. This allows for the estimation of αg and ψg by minimizing the overall differences between the measurement parameters across groups. The approach assigns a preference to models that maximize the number of invariant parameters while minimizing non-invariant parameters, thereby achieving an optimal balance. This methodology ensures that the group-specific factor means and variances are estimated accurately without strictly adhering to the assumption of exact MI.

The process starts with a configural model that provides the best fit for the data without enforcing invariance constraints. The alignment algorithm then optimizes model parameters to minimize non- invariance while maintaining the fit of the configural solution. This procedure ensures that the majority of parameters are invariant, while allowing for localized non-invariances that do not significantly affect the interpretation of results (Asparouhov & Muthén, 2014).

One of the major advantages of alignment is its ability to differentiate between invariant and non- invariant parameters in a single estimation step, eliminating the need for the iterative adjustments required in MGCFA. Additionally, alignment is particularly effective in studies with large group numbers, where full scalar invariance is unlikely to hold. However, the method assumes that the majority of parameters are invariant; violations of this assumption can impact the accuracy of estimated group differences and lead to biased interpretations (Fischer et al., 2019; Klieme, 2020).

The alignment optimization approach is an exploratory multiple-group factor analysis technique designed to identify the most suitable configuration of partial measurement invariance, allowing for approximate rather than strict equivalence across groups. This method has been proposed as a potential approach for estimating group means and

variances when full measurement invariance is not achieved, thereby allowing for approximate cross-group comparisons (Asparouhov & Muthén, 2014). However, the extent to which alignment optimization provides meaningful comparisons depends on the specific analytical context, and it should be considered as one of several possible solutions rather than a definitive method.

The procedure consists of two key stages. First, a configural invariance model is fitted, allowing factor loadings and intercepts to vary freely across groups while constraining factor means to zero and fixing factor variances at one. In the second stage, factor means and variances are estimated for each group without assuming full invariance. During this step, the alignment method applies a simplicity function, comparable to rotation criteria in exploratory factor analysis (EFA), to identify the optimal invariance structure (Asparouhov & Muthén, 2014, p. 496). More broadly, invariance alignment (IA) can be understood as a particular robust linking procedure for polytomous items, as commonly used in item response theory (IRT) (Robitzsch, 2020, Stats).

Following the alignment estimation, a comprehensive review is conducted to determine which measurement parameters exhibit approximate invariance and which deviate significantly. This is achieved by analyzing both visual representations and the degree of misfit in factor loadings and intercepts. To ensure the validity and reliability of the results, a benchmark of no more than 25 % non-invariance is typically recommended, as outlined by Asparouhov and Muthén (2014) and Muthén and Asparouhov (2014).

Alignment optimization has demonstrated significant utility in ILSAs, such as PISA and TALIS, where group-level heterogeneity complicates traditional MI testing. By identifying subsets of parameters and groups that meet invariance criteria, alignment optimization offers a more flexible approach compared to traditional methods for assessing measurement invariance across groups. However, its effectiveness in enabling meaningful cross-group comparisons depends on the extent of non-invariance and the specific analytical context (Asparouhov & Muthén, 2014; Robitzsch & Lüdtke, 2023).

1.4. From MG-CFA to alignment: approaches to measurement invariance in ILSAs

A critical distinction in testing measurement invariance lies between traditional multi-group confirmatory factor analysis (MG-CFA) and alignment optimization. MG-CFA is primarily a theory-driven approach that evaluates the fit of a proposed structure of implied covariances across groups by applying sequential equality constraints on parameters (metric and scalar invariance) (Kline, 2016). This stepwise model-trimming strategy often works well in two-group comparisons but becomes increasingly restrictive and prone to **false detection of** noninvariance as the number of groups rises, as is typically the case in international large-scale assessments (Byrne & van de Vijver, 2017; Kim et al., 2017; Rutkowski & Svetina, 2014). In contrast, alignment optimization seeks to minimize the overall degree of noninvariance by freely estimating factor means and variances through a simplicity function, thereby allowing a certain level of parameter variation while still producing comparable results across groups (Asparouhov & Muthén, 2014; Marsh et al., 2018; Munck et al., 2018; Seddig & Lomazzi, 2019). This method is better suited for analyses involving many groups, as it avoids relying on extensive pairwise comparisons and provides a practical way to evaluate approximate rather than exact invariance (Fischer & Karl, 2019; Lamm et al., 2019; Muthén & Asparouhov, 2013, 2014, 2018; van de Vijver et al., 2019). Importantly, alignment allows researchers to interpret factor means without imposing equality constraints on loadings and intercepts, which is especially relevant in **cross-cultural settings** where exact invariance is rarely attainable (Byrne & van de Vijver, 2017; Somaraju et al., 2022). Taken together, the two approaches reflect different philosophies of invariance testing: MG-CFA aims for exact equivalence, whereas alignment provides a framework for **approximate invariance** that acknowledges and accommodates heterogeneity across diverse populations.

1.5. Previous studies related to alignment optimization

The increasing application of alignment optimization methods in cross-national studies highlights their growing importance in addressing challenges related to measurement invariance. These methods have been applied across various educational and non-cognitive constructs, from teaching quality and job satisfaction to ICT readiness and political trust. Studies such as those analyzing teacher self-efficacy, distributed leadership, and student motivation have demonstrated the potential of alignment optimization to approximate invariance where traditional Multi-Group Confirmatory Factor Analysis (MGCFA) fails. However, despite these promising applications, the results are inconsistent across studies. For instance, while some constructs, like school ICT readiness or distributed leadership, exhibit low non-invariance rates that facilitate cross-group comparisons, others, such as teacher self-efficacy or mathematics self-concept, show higher levels of non-invariance. While this may introduce complexities in comparability, meaningful comparisons can still be made under certain conditions (Robitzsch & Lüdtke, 2023). These discrepancies suggest that the success of alignment methods heavily depends on the nature of the construct, item design, and group characteristics (Sandoval-Hernández et al., 2025).

Moreover, the lack of consensus on thresholds for acceptable non-invariance and the variability in methodological rigor across studies further complicate the interpretation of results. While some researchers adopt stringent criteria (e.g., <20 % non-invariance), others allow for higher thresholds, leading to divergent conclusions about construct comparability. Additionally, the diverse educational and cultural contexts in which these methods are applied introduce challenges related to linguistic and contextual biases, as seen in studies on teaching quality and political trust. This underscores the need for more standardized practices and methodological advancements, including simulation studies, to refine the alignment optimization approach. As the reliance on alignment methods continues to grow in large- scale assessments, establishing robust guidelines or integrating them with complementary methods, such as Bayesian approximate invariance, will be critical to enhancing their utility and thoughtful use (Sandoval-Hernández et al., 2025).

In the literature on student data, various methodological approaches have been proposed to address the challenges posed by exact measurement invariance. Among these, invariance alignment (IA) and Bayesian measurement invariance represent distinct approaches, with IA optimizing parameter structures across groups and Bayesian methods relying on probabilistic modeling with prior distributions (Robitzsch, 2022a). However, existing research has not demonstrated substantial improvements in the cross-national comparability of student questionnaire scales when employing these methods (Senden et al., 2023; see also Ding et al., 2022; Fischer et al., 2019; Odell et al., 2021; Wurster, 2022). This underscores the limitations of current statistical approaches in resolving invariance issues within student questionnaire data across diverse cultural contexts. To enhance cross-national comparability, it may be necessary to address potential sources of bias at all stages of cross-cultural research, from study design to data analysis (Senden et al., 2023).

1.6. The objective of this study

The primary aim of this study is to critically assess the effectiveness of alignment optimization as a methodological approach for facilitating cross-cultural comparisons across participating countries in international large-scale assessments (ILSAs). We seek to determine whether alignment optimization serves as a robust alternative to traditional methods for ensuring measurement invariance and, importantly, to identify the specific conditions under which it proves most effective.

To address this aim, we analyze data from the OECD's TALIS 2018

survey, encompassing responses from teachers and principals. As one of the most comprehensive datasets in educational research, TALIS 2018 provides a standardized framework for examining the comparability of constructs across diverse national and cultural contexts. By leveraging this dataset, our study aims to offer rigorous, evidence-based insights into the applicability, strengths, and limitations of alignment optimization in achieving meaningful and reliable cross-national comparisons. This work contributes to advancing methodological innovation in ILSAs and underscores the critical importance of ensuring comparability in educational research and policymaking.

2. Methodology

2.1. Data

This study utilizes data from the 2018 Teaching and Learning International Survey (TALIS), conducted by the Organisation for Economic Co-operation and Development (OECD). TALIS is a large-scale international survey that provides insights into the working conditions of teachers and school leaders and the learning environment in schools. The 2018 cycle of TALIS collected data from 48 education systems, offering comprehensive information on teaching practices, school leadership, and professional development.

For this study, we focused on the data collected from lower secondary school teachers (ISCED 2) and their school leaders. TALIS 2018 employs a two-stage sampling design, where schools are sampled first, followed by a random sample of teachers within each participating school. The dataset includes responses from approximately 260,000 teachers and 13,000 school leaders, representing diverse educational systems globally. The data was weighted using country-specific sampling weights to account for the sampling design and ensure accurate cross-national comparisons.

2.2. Measures

The TALIS 2018 questionnaires provide rich data on a variety of constructs relevant to teaching and school leadership. For this study, we focused on a subset of variables to examine alignment optimization and measurement invariance across countries. These variables include:

Teacher Constructs: Constructs such as job satisfaction, workload stress, teacher-student relations, and self-efficacy were analyzed to capture teachers' perceptions of their work environment and professional practices (see Table 7 for more detailed information).

Principal Constructs: Variables related to distributed leadership, stakeholder involvement, and school leadership were selected to reflect the leadership styles and practices at the school level.

Each construct was measured using a set of items grouped into scales. For example, teacher job satisfaction was assessed through subscales focusing on satisfaction with the profession and work environment. Similarly, distributed leadership was evaluated using items that reflect the participation of stakeholders in decision-making processes.

All scales were derived from TALIS 2018 and validated through a rigorous development process, ensuring their reliability and validity across participating countries (OECD, 2020). These pre-established measures, rather than being constructed specifically for this study, were adopted to maintain consistency with international assessment standards. The scale development process included extensive piloting, psychometric analysis, and expert review, ensuring their appropriateness for large-scale educational research. This approach aligns with the methodological framework outlined in the TALIS 2018 Technical Report and provides a robust foundation for the alignment optimization analysis conducted in this study. These measures formed the foundation for the alignment optimization analysis conducted in this study.

2.3. Analytical strategy

Our analytical strategy comprised several steps to assess the comparability and validity of the scales derived from the TALIS 2018 dataset, specifically focusing on teacher and principal responses at the ISCED 2 level. The primary goal was to ensure that the scales used in the analysis were suitable for cross-country comparisons and reliable for measuring the intended constructs. All analyses were conducted using the rd3c3 R library (Carrasco et al., 2024), a versatile tool developed by the authors for advanced psychometric analysis with a focus on cross-national comparability. The rd3c3 package serves as a wrapper for the alignment procedure implemented in Mplus, facilitating its application within the R environment through *MplusAutomation* (Hallquist & Wiley, 2018).

2.4. Evaluation of dimensionality

The first step involved evaluating the dimensionality of each scale to confirm that they were unidimensional and thus measured a single underlying construct. We utilized the parallel analysis method as recommended by Lubbe (2019) for this purpose. Parallel analysis involves comparing the observed eigenvalues from item correlation matrices, with those obtained from randomly generated data correlation matrices of the same size. If the observed eigenvalues exceed the random ones, it indicates the presence of a significant factor. Lubbe's (2019) parallel analysis is designed for ordered categorical indicators. It uses polychoric correlation matrices instead or pearsonian correlations matrices, and accounts for item indicator response distribution. Traditional parallel analysis using pearsonian correlations tend overestimate number of factors, while parallel analysis on polychoric correlations may also overestimate the number of factors, if item category responses are not taken into account. By applying this method, we systematically assessed each scale to determine the appropriate number of factors to retain. Assessing unidimensionality is crucial because it validates the assumption that the scale items collectively shared a single source of variance, which is a prerequisite for subsequent analyses like measurement invariance testing and alignment optimization.

2.5. Reliability

The second phase of the analytical strategy focused on evaluating the reliability of the scales. This step involved calculating key reliability metrics, including person separation reliability and Cronbach's alpha. Assessing reliability is a critical precursor to structural analyses, as it ensures that the scales are measuring constructs consistently across individuals and contexts. Without sufficient reliability, subsequent validity and invariance testing may be compromised. Person separation reliability assesses the scale's ability to differentiate between groups based on the factor realizations and their standard errors (Verhavert et al., 2018; Wright & Masters, 1982), while Cronbach's alpha estimates internal consistency and is widely used as a measure of reliability in psychometric research (Cronbach, 1951; Cronbach & Shavelson, 2004). These measures were calculated to ensure the robustness of the scales and their suitability for use in subsequent analyses.

2.6. Measurement invariance testing

To evaluate the measurement invariance of the scales, we employed a model-based approach using the graded response model (GRM), which is also referred to as confirmatory factor analysis (CFA) with ordinal indicators. This approach follows the guidelines established by Wu and Estabrook (2016), Rutkowski and Svetina (2017), and Tse et al. (2024). The GRM is particularly suited for scales with ordinal data, as it accounts for the graded nature of responses and allows for robust testing of measurement properties across groups. While we acknowledge that other researchers often use the normal distribution factor model for

Likert-scale items, our choice was driven by the categorical nature of the TALIS 2018 data. Some studies (e.g., Rhemtulla et al., 2012) suggest that categorical variables can be treated as continuous under certain distributional assumptions and with sufficient response categories; however, this decision remains partially assumption-dependent and therefore a methodological choice (Robitzsch, 2022b).

Our analysis adopted a trimming strategy, beginning with the most constrained model (strict invariance) and progressively relaxing constraints to test less restrictive models—such as scalar and threshold invariance—in line with current practices for ordered categorical data. The primary focus was on achieving scalar and strict invariance, which are necessary to support comparisons of both latent and observed means across countries (Tse et al., 2024).

The evaluation criteria for model fit were based on recommendations by Rutkowski and Svetina (2017). Specifically, we applied a strict RMSEA threshold of < 0.055 as the primary cut-off for determining acceptable model fit in scalar invariance testing. However, in the initial pooled CFAs, we allowed for a more relaxed RMSEA threshold of < 0.10 to accommodate the complexity of the data structure and to assess general model adequacy (He et al., 2014; Hu & Bentler, 1999). This two-phase strategy ensured that exploratory assessments (pooled CFA) maintained feasibility while the final multigroup comparisons adhered to conservative fit standards for ordinal data (Chang et al., 2017).

Importantly, our approach diverges from the IEA team's practice, which applies CFA using maximum likelihood (ML) estimation, treating Likert-scale items as continuous. In contrast, given the ordinal-categorical structure of the TALIS data, we applied DWLS (diagonally weighted least squares) estimation, which is more appropriate for modeling categorical outcomes. This distinction is critical, as RMSEA values derived from ML-based CFA are not directly comparable to those from DWLS-based models, and traditional thresholds like RMSEA < 0.08 or 0.10 may be too lenient in the context of ordinal indicators (Xia & Yang, 2019). Therefore, importing fit thresholds from continuous-data CFA and applying them to ordinal-data CFA is methodologically inappropriate (Kite et al., 2018).

Our work is grounded in the approach described by Svetina et al. (2020), who offer updated guidelines for invariance testing using categorical data and provide model specifications tailored for multigroup comparisons. These specifications include:

- Pooled: a graded response model applied to the combined sample with design weights;
- Configural: a multigroup model with freely estimated parameters across groups;
- Threshold: a model where thresholds are held equal across groups;
- Scalar: a model with equal loadings and thresholds but free scale factors;
- Strict: a fully invariant model except for group-level means.

Their simulations recommend a RMSEA < 0.055 threshold as a rule of thumb for scalar-level invariance with ordinal indicators across many groups. We followed this threshold when evaluating our multigroup CFA models using DWLS estimation.

There are thus three methodological approaches in play throughout this paper:

1. Pooled CFA with DWLS, using RMSEA < 0.10 for exploratory model fit;
2. Multigroup CFA with ordinal indicators (DWLS), using RMSEA < 0.055 as a criterion for scalar invariance;
3. Alignment optimization applied over DWLS-based CFA, providing parameter-level evidence of measurement invariance.

While RMSEA served as our primary fit index due to its sensitivity to structural misfit in measurement invariance models, we also considered CFI as a complementary measure, especially in the pooled CFA phase. In Tables 3 and 4, CFI values consistently exceeded standard thresholds, which supports the overall robustness of the models.

Table 3
Fit Indices for Teacher Scales.

Scale	CFI	RMSEA	SRMR
Workload stress	0.96	0.13	0.04
Workplace well-being and stress	1	0.04	0.01
Satisfaction with target class autonomy	0.99	0.11	0.02
Personal utility motivation to teach	1	0.06	0.01
Clarity of instruction (subscale)	0.98	0.11	0.03
Cognitive activation (subscale)	0.96	0.15	0.03
Classroom management (subscale)	0.98	0.19	0.02
Exchange and co-ordination among teachers (subscale)	1	0.03	0.01
Professional collaboration in lessons among teachers (subscale)	0.98	0.05	0.01
Effective professional development	0.99	0.04	0.03
Professional development barriers	0.96	0.07	0.03
Self-efficacy in classroom management (subscale)	1	0.07	0.01
Self-efficacy in instruction (subscale)	1	0.05	0.01
Self-efficacy in student engagement (subscale)	0.99	0.12	0.02
Job satisfaction with work environment (subscale)	0.99	0.1	0.02
Job satisfaction with profession (subscale)	0.97	0.2	0.03
Teachers' perceived disciplinary climate	1	0.02	0
Teacher-student relations	1	0.06	0.01
Participation among stakeholders, teachers	0.95	0.21	0.05
Team innovativeness	1	0.06	0
Self-related efficacy in multicultural classrooms	0.94	0.22	0.05
Diversity practices	0.96	0.13	0.06
Teaching practices	0.91	0.12	0.04

Table 4
Fit Indices for Principal Scales.

Scale	CFI	RMSEA	SRMR
Workload stress	1	0	0
Job satisfaction with work environment (subscale)	0.94	0.24	0.06
Job satisfaction with profession (subscale)	0.97	0.16	0.03
School leadership	1	0	0
Participation among stakeholders, principals	0.94	0.16	0.06
Academic pressure	0.97	0.16	0.05
Stakeholder involvement, partnership	1	0	0
Lack of special needs personnel	1	0	0
School delinquency and violence	1	0.01	0
Organisational innovativeness	0.98	0.2	0.03
Diversity beliefs	0.99	0.17	0.02
Distributed leadership	1	0	0
Diversity practices, school	0.98	0.06	0.04
Diversity policies, school	0.93	0.14	0.22
Equity beliefs	0.99	0.14	0.03

We also note that alignment optimization offers additional value by directly identifying which factor loadings and intercepts are non-invariant, rather than relying on global model fit indices alone. This approach complements the traditional MGCFA framework and provides a more flexible, parameter-level understanding of comparability across countries.

Finally, we recognize that the discussion on fit index benchmarks for categorical data is still evolving. As recommended in recent literature (e.g., Rutkowski & Svetina, 2017; Svetina & Rutkowski, 2020), continued refinement of cut-off values and modeling strategies is necessary, particularly in large-scale, multi-group contexts like TALIS where strict invariance is often difficult to achieve.

2.7. Alignment optimization

The alignment optimization process employed in this study used the same graded response model (GRM) as in the previous routine, ensuring consistency in the analytical approach. This model facilitates the comparison of latent means across groups by accounting for the ordinal nature of the data.

To anchor the alignment process, we fixed the latent mean to the

most central comparison group. This central group was identified based on its mean scores over the observed item responses, providing a stable reference point for aligning other groups.

For evaluating the success of the alignment process, we adopted a threshold of 75 % common parameters across the compared groups. This threshold ensures that the alignment meets a satisfactory level of parameter invariance, enabling meaningful cross-group comparisons.

Overall, the model estimation process followed the framework established by Svetina et al. (2020) model-based invariance testing. However, a key distinction in this study was the inclusion of the strict invariance model, aligning more closely with recent advancements in the field as discussed by Tse et al. (2024) and Padgett (2023). This approach allowed for a more comprehensive evaluation of measurement invariance across the specified groups, ensuring robust comparisons of latent means and variances.

An important consideration in the alignment procedure is the choice of the epsilon (ε) value, which affects the statistical properties of the estimator. Research suggests that the default *Mplus* setting of 0.01 may be less optimal compared to a more stringent value of 0.001, which can lead to improved estimation accuracy (Robitzsch, 2024).

In the results section, we present a detailed report for one selected teacher scale out of the 23 evaluated, and one principal scale out of the 15 analyzed. This comprehensive report showcases all the generated results using the programmed alignment routines, providing a clear and thorough overview of the alignment procedure and its outcomes.

3. Results

3.1. The results of dimensionality

The dimensionality analysis, conducted using parallel tests, confirmed that all teacher and principal scales are unidimensional. For teachers, the analysis included 23 scales, such as "Workload stress," "Cognitive activation," and "Job satisfaction with work environment," all of which demonstrated a single underlying dimension. Similarly, for principals, the 15 scales, including "School leadership," "Academic pressure," and "Organisational innovativeness," were found to be unidimensional. These findings support the assumption that each scale measures a coherent and singular construct, ensuring their validity for subsequent analyses.

3.2. The results of reliability

In the Appendix, Table 1 presents reliability metrics for teacher scales from TALIS 2018, showing that most scales achieved values above the acceptable threshold of 0.70 for both Cronbach's alpha and separation reliability. While scales like "Team innovativeness" and "Participation among stakeholders, teachers" demonstrated strong reliability, others such as "Exchange and coordination among teachers" and "Professional collaboration in lessons" showed slightly lower values, indicating areas for improvement. Overall, the results affirm the robustness of the teacher-related constructs.

Table 2 displays reliability results for principal scales, also indicating generally strong internal consistency across most measures. High reliability was observed for scales like "Diversity beliefs" and "Organizational innovativeness," while lower but acceptable values were found for "Diversity policies" and "Diversity practices." All scales were retained to allow a comprehensive assessment in the subsequent measurement invariance analysis, even if a few fell slightly below the conventional cut-off.

3.3. The results of measurement invariance testing

Table 3 presents the pooled data fit indices (CFI, RMSEA, and SRMR) for the teacher scales in TALIS 2018, with the following threshold values applied: 0.90 for the Comparative Fit Index (CFI), 0.08 for the Standardized Root Mean Square Residual (SRMR), and 0.10 for the Root Mean Square Error of Approximation (RMSEA).

The CFI values indicate that most teacher scales meet or exceed the 0.90 threshold, suggesting an acceptable fit for the majority of scales. Similarly, the SRMR values remain below 0.08 across all scales, reinforcing adequate model fit. However, the RMSEA values exhibit more variability, with several scales approaching or exceeding the 0.10 threshold. Notably, some scales, such as Classroom Management (0.19), Job Satisfaction with Profession (0.20), and Participation Among Stakeholders (0.21), exceed this threshold, signaling potential concerns regarding model fit. Overall, these results suggest that while most teacher scales achieve acceptable model fit, some may require further investigation or refinement to improve their alignment. The findings highlight the importance of evaluating multiple fit indices to ensure measurement comparability in large-scale assessments.

Table 4 presents the pooled fit indices (CFI, RMSEA, and SRMR) for the principal scales in TALIS 2018, with the following thresholds applied: 0.90 for the Comparative Fit Index (CFI), 0.08 for the Standardized Root Mean Square Residual (SRMR), and 0.10 for the Root Mean Square Error of Approximation (RMSEA).

The CFI values demonstrate that all principal scales meet or exceed the 0.90 threshold, indicating good model fit across all scales. Similarly, SRMR values remain below the 0.08 threshold, reinforcing an acceptable fit for all principal scales. However, RMSEA values show greater variability, with some scales exceeding the 0.10 threshold, particularly Job Satisfaction with Work Environment (0.24), Organisational Innovativeness (0.20), and Diversity Policies (0.14). These results suggest potential models fit concerns for a few scales, warranting further investigation into possible sources of misfit.

In cases where RMSEA values are reported as 0, this is due to the scale comprising only three items. RMSEA is known to be an unreliable fit index for models with very few degrees of freedom, particularly for models with only three indicators, where it often returns values near or equal to zero regardless of model fit (Kenny et al., 2015). Therefore, these values should be interpreted with caution and in conjunction with other fit indices.

Overall, the findings provide strong evidence of acceptable model fit for most principal scales, with only a few scales requiring additional evaluation, particularly in relation to RMSEA. This highlights the importance of assessing multiple fit indices to ensure robust model alignment and comparability in international assessments.

Table 5 summarizes the scalar-level invariance results for teacher scales from TALIS 2018, evaluated using a stringent RMSEA threshold of 0.055, which is often applied to ensure meaningful comparability of latent means across countries. Only a limited number of teacher scales—*Self-efficacy in classroom management*, *Self-efficacy in instruction*, *Teacher–student relations*, and *Team innovativeness*—met this threshold, as indicated by the "yes" outcome in the rmsea_test column. These results suggest that for these particular constructs, mean comparisons across countries may be considered valid and interpretable within an international context.

In contrast, most teacher scales either exceeded the RMSEA threshold or failed to converge during scalar invariance testing. For instance, constructs such as *Job satisfaction with profession*, *Classroom management*, and *Participation among stakeholders* demonstrated RMSEA values well above the cut-off, indicating potential misfit in the scalar model. Scales with missing RMSEA entries (e.g., *Workplace well-being and stress*, *Effective professional development*, and *Diversity practices*) likely reflect non-convergence issues, which may stem from insufficient variability, local model misfit, or data sparsity across countries—such as low frequencies in specific item response categories within certain national samples. Sparsity limits the reliability of polychoric correlations and can prevent the estimation algorithm from identifying stable threshold parameters, especially in models with many groups and ordinal indicators. These findings highlight the persistent measurement challenges in achieving scalar invariance for teacher-reported

Table 5
Testing Means Comparability Results for Teacher Scales.

scale	*model*	*RMSEA*	*CFI*	*SRMR*	*rmsea_test*
Workload stress	scalar	0.08	0.96	0.05	no
Workplace well-being and stress	scalar				
Satisfaction with target class autonomy	scalar	0.07	0.99	0.03	no
Personal utility motivation to teach	scalar	0.06	1	0.03	no
Clarity of instruction (subscale)	scalar	0.06	0.98	0.04	no
Cognitive activation (subscale)	scalar	0.08	0.95	0.04	no
Classroom management (subscale)	scalar	0.1	0.98	0.04	no
Exchange and co-ordination among teachers (subscale)	scalar	0.07	0.95	0.04	no
Professional collaboration in lessons among teachers (subscale)	scalar	0.07	0.83	0.05	no
Effective professional development	scalar				
Professional development barriers	scalar	0.07	0.92	0.05	no
Self-efficacy in classroom management (subscale)	scalar	0.05	1	0.02	yes
Self-efficacy in instruction (subscale)	scalar	0.04	0.99	0.02	yes
Self-efficacy in student engagement (subscale)	scalar				
Job satisfaction with work environment (subscale)	scalar	0.07	0.99	0.03	no
Job satisfaction with profession (subscale)	scalar	0.11	0.96	0.05	no
Teachers' perceived disciplinary climate	scalar	0.06	1	0.03	no
Teacher-student relations	scalar	0.05	1	0.02	yes
Participation among stakeholders, teachers	scalar	0.12	0.97	0.06	no
Team innovativeness	scalar	0.05	1	0.02	yes
Self-related efficacy in multicultural classrooms	scalar	0.12	0.96	0.06	no
Diversity practices	scalar				
Teaching practices	scalar	0.07	0.9	0.05	no

*If some scale lines are empty, which means may not converge the model

constructs, underscoring the importance of cautious interpretation when comparing means across countries. Researchers are advised to consider alternative approaches such as partial invariance models or alignment optimization to explore comparability more flexibly.

Table 6
Testing Means Comparability Results for Principal Scales.

scale	*model*	*RMSEA*	*CFI*	*SRMR*	*rmsea_test*
Workload stress	scalar	0.07	0.96	0.04	no
Job satisfaction with work environment (subscale)	scalar				
Job satisfaction with profession (subscale)	scalar				
School leadership	scalar				
Participation among stakeholders, principals	scalar				
Academic pressure	scalar				
Stakeholder involvement, partnership	scalar				
Lack of special needs personnel	scalar	0.07	0.99	0.04	no
School delinquency and violence	scalar				
Organisational innovativeness	scalar				
Diversity beliefs	scalar				
Distributed leadership	scalar				
Diversity practices, school	scalar				
Diversity policies, school	scalar				

*If some scale lines are empty, which means may not converge the model

Table 6 presents the scalar-level invariance results for principal scales. Here, the pattern is even more pronounced: none of the principal scales achieved scalar invariance based on the strict RMSEA < 0.055 threshold. For several scales—such as *Lack of special needs personnel*—results are available, but the RMSEA exceeds the threshold, indicating lack of scalar-level comparability. For most other principal scales, scalar models failed to converge, and thus RMSEA values could not be estimated. This is marked by empty cells in the table.

The failure to converge may be partially attributed to the smaller sample sizes of principals in comparison to teachers, which increases model instability and reduces statistical power for detecting invariance. Additionally, sparse response patterns in certain countries—due to limited use of specific response categories or highly skewed item distributions—can exacerbate estimation difficulties, particularly in ordinal CFA models. Another plausible explanation is that constructs related to school leadership (e.g., *Distributed leadership*, *Participation among stakeholders*) are inherently more context-dependent and may vary substantially across national systems due to differing school governance structures, accountability frameworks, and leadership norms. Such cultural and systemic diversity can introduce severe non-invariance, further complicating model convergence and comparability.

These results warrant caution in interpreting mean comparisons of principal scales across countries using traditional MGCFA-based methods. They also suggest that relying solely on scalar invariance criteria may obscure meaningful insights. Alternative techniques, such as alignment optimization, may offer more practical solutions by identifying and adjusting for non-invariant parameters, as discussed in subsequent sections of this paper. Nonetheless, these findings underscore the ongoing methodological complexity in evaluating and achieving cross-national comparability in international large-scale assessments (ILSAs), especially for school leader-reported data.

3.4. The results of alignment optimization

Table 7 presents the alignment test results for teacher scales in TALIS 2018. The invariance parameter (*inv_par)* is used to assess the degree of comparability across countries, with a threshold of 0.75 indicating acceptable alignment. Scales meeting this criterion (align_test = "yes") suggest that their latent means can be meaningfully compared across groups. However, the majority of teacher scales fall below this threshold, indicating limited cross-country comparability. Notably, only

Table 7
Alignment Results with Teacher Scales.

scale	inv_par	align_test
Workload stress	0.48	no
Workplace well-being and stress	0.53	no
Satisfaction with target class autonomy	0.56	no
Personal utility motivation to teach	0.48	no
Clarity of instruction (subscale)	0.65	no
Cognitive activation (subscale)	0.52	no
Classroom management (subscale)	0.46	no
Exchange and co-ordination among teachers (subscale)	0.43	no
Professional collaboration in lessons among teachers (subscale)	0.42	no
Effective professional development	0.81	yes
Professional development barriers	0.47	no
Self-efficacy in classroom management (subscale)	0.66	no
Self-efficacy in instruction (subscale)	0.67	no
Self-efficacy in student engagement (subscale)	0.61	no
Job satisfaction with work environment (subscale)	0.59	no
Job satisfaction with profession (subscale)	0.50	no
Teachers' perceived disciplinary climate	0.61	no
Teacher-student relations	0.59	no
Participation among stakeholders, teachers	0.53	no
Team innovativeness	0.52	no
Self-related efficacy in multicultural classrooms	0.55	no
Diversity practices	0.74	no
Teaching practices	0.55	no

one scale (Effective professional development) surpasses the alignment threshold, reinforcing the challenges of achieving comparability in teacher-reported constructs within international assessments. These findings highlight the need for methodological refinement or alternative approaches to improve cross-national measurement invariance in large-scale assessments.

Table 8 presents the alignment test results for principal scales in TALIS 2018. The invariance parameter (inv_par) assesses the degree of cross-country comparability, with a threshold of 0.75 indicating acceptable alignment. Notably, all principal scales surpass this threshold (align_test = "yes"), suggesting that their latent means can be meaningfully compared across countries. These results indicate strong measurement consistency and robustness for principal-reported constructs across different national contexts, reinforcing their suitability for cross-national analysis in international large-scale assessments.

4. Discussion

This study examined the comparability of teacher and principal scales derived from the TALIS 2018 dataset across participating countries, using both traditional multiple-group confirmatory factor analysis (MGCFA) and alignment optimization techniques. The findings highlight the potential of alignment optimization as a methodological tool for enhancing cross-national comparability in ILSAs. Alignment optimization offers a practical alternative to traditional MGCFA by allowing for approximate measurement invariance, which is particularly useful in ILSAs where strict invariance is often unattainable due to cultural and linguistic diversity. Unlike MGCFA, which requires strict scalar invariance for valid cross-national comparisons, alignment optimization identifies subsets of parameters and groups that meet invariance criteria, enabling meaningful comparisons even in the presence of partial non-invariance. This approach can be integrated into the existing scaling and reporting processes of ILSAs by providing additional insights into the robustness of scale scores and identifying areas where scales may require refinement. In the following sections, we discuss how these insights can inform the interpretation and use of ILSA data, as well as future methodological advancements (Robitzsch & Lüdtke, 2023; Fischer et al., 2025; Funder & Gardiner, 2024; Rosseel & Loh, 2022).

The results showed significant differences between the two methods in terms of achieving scalar-level invariance, as well as notable disparities between teacher and principal scales. Under the MGCFA framework, almost all scales—both teacher and principal—failed to reach scalar invariance. This outcome underscores the well-documented challenges associated with ensuring measurement equivalence in large-scale, cross-national assessments. The absence of full scalar invariance suggests that mean comparisons across countries should be interpreted with caution, given potential differences in how constructs are measured across contexts.

Table 8
Alignment Results with Principals' Scales.

scale	inv_par	align_test
Workload stress	0.92	yes
Job satisfaction with work environment (subscale)	0.92	yes
Job satisfaction with profession (subscale)	0.91	yes
School leadership	0.97	yes
Participation among stakeholders, principals	0.92	yes
Academic pressure	0.98	yes
Stakeholder involvement, partnership	0.90	yes
Lack of special needs personnel	0.90	yes
School delinquency and violence	0.96	yes
Organisational innovativeness	0.95	yes
Diversity beliefs	0.95	yes
Distributed leadership	0.96	yes
Diversity practices, school	0.97	yes
Diversity policies, school	0.91	yes
Equity beliefs	0.97	yes

Interestingly, on the TALIS 2018 Technical Report only "team innovativeness" from teacher scales reached the scalar level invariance which ensure mean score comparability across countries, however, in our study in addition to team innovativeness, Self-efficacy in classroom management, Self-efficacy in instruction and Teacher-student relations scales also reached the scalar level invariance. For principals, on the technical report diversity beliefs reached scalar level invariance which ensured the cross-cultural comparison across countries for this scale, however, in our analysis, none of principal scales reached the scalar level invariance.

In contrast, alignment optimization yielded markedly different results, particularly for the principal scales. All principal scales successfully achieved comparability under this method, enabling meaningful cross-national comparisons of their latent means. This suggests that alignment optimization is a powerful tool for addressing the limitations of traditional MGCFA, particularly for constructs measured at the school leadership level. This result may raise questions, especially considering that the principal sample sizes are considerably smaller than those for teachers. It is plausible that the smaller principal samples limited the statistical power to detect non-invariance, which might have led the alignment method to yield more favorable comparability results. While alignment optimization is designed to tolerate minor non-invariances, its sensitivity may differ depending on group size and response patterns. Readers should be cautious when interpreting these results and consider the possibility that alignment comparability in smaller samples may reflect reduced power to detect violations rather than genuinely higher measurement invariance. This issue merits further methodological investigation in future simulation studies.

However, for teacher scales, alignment optimization was less effective, with many scales still failing to meet the criteria for comparability. This highlights persistent measurement challenges for constructs assessed at the teacher level, even with advanced techniques. The observed high levels of non-invariance may stem from minor parameter differences across groups that, due to the large sample sizes in this study, become statistically significant (Asparouhov & Muthén, 2022). This raises the question of whether the statistical significance threshold should be recalibrated when applying the alignment method to analyze measurement invariance in International Large-Scale Assessment (ILSA) data (Senden et al., 2023). Further methodological research is needed to determine the necessity of such an adjustment and to develop effective strategies for its implementation.

While our findings illustrate the promise of alignment optimization in addressing cross-national comparability, we caution against interpreting it as a universally superior alternative to multi-group CFA. As Luong and Flake (2023) emphasize, alignment should not be treated as an accessory analysis to be applied automatically, nor as a universal solution. Its scope is limited to latent mean and variance comparisons, and questions of generalizability remain open. We therefore position alignment as a **valuable methodological advancement** for large-scale, multi-group contexts—particularly ILSAs—while recognizing that it complements, rather than replaces, traditional CFA approaches. Our findings align with recent applications of alignment in ILSAs. For example, Fang et al. (2025), using TALIS 2018, demonstrated that alignment optimization offered clear advantages over MG-CFA in identifying comparable patterns across many groups and in detecting item-level noninvariance. At the same time, they emphasized that unlike MG-CFA, which is supported by well-established guidelines, alignment still lacks universally agreed-upon standards for implementation and reporting. This reinforces the need for careful application and transparent documentation when using alignment in cross-national research.

These findings contribute to the ongoing debate on how much measurement invariance violations impact substantive research conclusions. While alignment optimization provides a more flexible approach to assessing cross-national comparability, prior research suggests that results from different invariance assumptions often yield highly similar substantive conclusions (Desa et al., 2019). This raises an

important question about the extent to which strict measurement invariance criteria should guide the interpretation of ILSA scale scores. In practical terms, even when full scalar invariance is not achieved, ILSAs report scale scores on an international metric, implicitly consenting comparison of means across countries. This practice, while can be considered useful for broad cross-national analyses, also introduces risks when comparability assumptions are not fully met. Rather than advocating solely for methodological refinements, this study underscores the need for a more transparent discussion on how ILSA scores should be used in secondary analyses. One possible recommendation is to promote model-based approaches in reporting, where adjustments for partial measurement invariance are incorporated into substantive analyses rather than relying exclusively on pre-scaled scores.

These findings have important implications for the interpretation and use of TALIS 2018 data. For teacher scales that failed to achieve alignment optimization comparability, cross-national mean comparisons should be avoided or treated with extreme caution. Researchers may consider alternative approaches, such as focusing on within-country analyses or examining trends rather than direct comparisons or building on prior research (e.g., Fischer et al., 2019; Klieme, 2020; Scherer & Gustafsson, 2015; Scherer & Nilsen, 2016; Eryilmaz et al., 2020; Eryilmaz & Sandoval Hernandez, 2024; Kaya et al., 2024), a practical alternative is to restrict cross-national mean score comparisons to a select group of culturally and linguistically similar countries. For principal scales, the successful application of alignment optimization provides confidence in their comparability and supports their use in cross-national analyses of leadership practices and perceptions.

Future research should delve deeper into the reasons for the differential performance of teacher and principal scales, considering factors such as construct clarity, item design, and the influence of cultural and contextual variables unique to teachers and principals. Additionally, methodological advancements in alignment optimization are essential to enhance its application in complex hierarchical models and diverse educational contexts, further addressing the unique challenges of large-scale assessments. To fully validate this promising method, comprehensive simulation studies are required to determine the optimal sample sizes, item numbers, and response category designs necessary for reliable and meaningful cross-national comparisons.

Researchers can apply the practices discussed in this study using the rd3c3 R package, which facilitates alignment optimization in international large-scale assessments. This package allows users to efficiently implement alignment optimization, providing a more flexible alternative to traditional MGCFA. By leveraging rd3c3, researchers can better evaluate cross-national comparability, visualize alignment results, and ensure more transparent and robust interpretations of ILSA data.

In conclusion, while alignment optimization helps address some of the limitations of traditional invariance testing, its effectiveness is closely tied to how fit indices are used to assess model adequacy. In this study, we relied primarily on RMSEA and CFI, two commonly used indices in measurement invariance testing. However, it is well-documented that RMSEA behaves differently in ML-based CFA compared to DWLS-based CFA (ordinal models), and conventional cut-offs may not be directly transferable (Xia & Yang, 2019). Applying ML-based RMSEA thresholds to ULS and DWLS estimation methods can lead to the accumulation of models with severe misfit that are nonetheless considered acceptable, particularly in research settings with large sample sizes (Xia & Yang, 2019).

While RMSEA is highly sensitive to minor misfits, especially in large samples, it can sometimes overstate measurement problems when small deviations occur across groups. Conversely, CFI tends to be more stable, but it may not adequately capture localized areas of misfit. In our pooled CFA analyses, some elevated RMSEA values may also reflect local dependence among similarly worded items, which can inflate residual covariances not accounted for by the latent construct (Marsh et al., 2013). Additionally, RMSEA values of zero—observed in some scales—are explained by the limited number of items (e.g., three-item scales), where the model's low degrees of freedom can lead to artificially suppressed fit statistics (Kenny et al., 2015). Given these limitations, our findings should be interpreted with caution, particularly regarding teacher scales where alignment thresholds were not consistently met.

By acknowledging these limitations and leveraging advanced techniques, researchers can better support evidence-based educational policymaking on a global scale. Future research should continue refining methodological benchmarks, ensuring that measurement invariance decisions are informed by appropriate fit criteria for specific estimation methods.

Future research should prioritize refining alignment methodologies to address the complexities inherent in culturally and contextually diverse constructs. Simulation studies are also necessary to determine optimal conditions for implementing alignment methods, including considerations such as sample size, item design, and construct clarity. Ensuring valid cross-national comparability would lead to higher-quality data that can be used to benchmark educational performance, identify best practices, and understand differences in the performance of education systems within and between countries. Such data would also facilitate unveiling systemic relationships between inputs and outputs in education, support progress toward global educational agendas (e.g., Sustainable Development Goals), and provide critical cultural and contextual insights. In sum, advancing methodological approaches like alignment optimization is necessary for enabling reliable global educational comparisons, fostering innovation, and informing evidence-based policies that address the diverse realities of education systems worldwide.

This study shows that alignment optimization is a useful tool for identifying non-invariant factor loadings and intercepts across countries in international large-scale assessments. It proved particularly effective for principal scales, where it enabled meaningful cross-national comparisons despite the failure of traditional MGCFA to establish scalar invariance. However, for teacher scales, substantial non-invariance persisted even under alignment, limiting the comparability of these constructs. Thus, while alignment optimization enhances our ability to detect and manage non-invariance, its success in improving comparability depends on the nature of the constructs and respondent group.

CRediT authorship contribution statement

Carasco Diego: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Andres Sandoval-Hernandez:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **ERYILMAZ NURULLAH:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Acknowledgement

This research was supported by funding from the International Association for the Evaluation of Educational Achievement (IEA) Research and Development (R&D) funding program. We are deeply grateful to the IEA for providing the financial resources and institutional support that made this work possible. Their commitment to advancing educational research has been instrumental in enabling this study. We also extend my thanks to the colleagues and experts associated with the IEA for their invaluable insights and encouragement throughout the research process.

Appendix

Table 1
The reliability results for teacher scales.

scale	separation	alpha
Workload stress	0.81	0.79
Workplace well-being and stress	0.82	0.75
Satisfaction with target class autonomy	0.80	0.83
Personal utility motivation to teach	0.85	0.85
Clarity of instruction (subscale)	0.72	0.72
Cognitive activation (subscale)	0.74	0.73
Classroom management (subscale)	0.85	0.87
Exchange and co-ordination among teachers (subscale)	0.77	0.74
Professional collaboration in lessons among teachers (subscale)	0.64	0.63
Effective professional development	0.45	0.55
Professional development barriers	0.69	0.64
Self-efficacy in classroom management (subscale)	0.81	0.84
Self-efficacy in instruction (subscale)	0.78	0.80
Self-efficacy in student engagement (subscale)	0.82	0.83
Job satisfaction with work environment (subscale)	0.79	0.78
Job satisfaction with profession (subscale)	0.80	0.79
Teachers' perceived disciplinary climate	0.86	0.85
Teacher-student relations	0.76	0.80
Participation among stakeholders, teachers	0.85	0.87
Team innovativeness	0.86	0.90
Self-related efficacy in multicultural classrooms	0.85	0.84
Diversity practices	0.62	0.70
Teaching practices	0.60	0.56

Table 2
The reliability results for principal scales.

scale	separation	alpha
Workload stress	0.64	0.63
Job satisfaction with work environment (subscale)	0.75	0.76
Job satisfaction with profession (subscale)	0.77	0.74
School leadership	0.80	0.82
Participation among stakeholders, principals	0.77	0.76
Academic pressure	0.77	0.77
Stakeholder involvement, partnership	0.73	0.67
Lack of special needs personnel	0.75	0.75
School delinquency and violence	0.80	0.82
Organisational innovativeness	0.81	0.85
Diversity beliefs	0.77	0.90
Distributed leadership	0.73	0.76
Diversity practices, school	0.54	0.61
Diversity policies, school	0.47	0.64
Equity beliefs	0.66	0.84

Table 9
TALIS 2018 teachers and principals' scales and invariance levels.

Scale Label	Variable Name	Scale Type	Invariance Level
Academic pressure	T3PACAD	Principal Scales	Metric
Stakeholder involvement, partnership	T3PCOM	Principal Scales	Metric
School delinquency and violence	T3PDELI	Principal Scales	Configural
Diversity beliefs	T3PDIVB	Principal Scales	Scalar
Job satisfaction, overall, teacher	T3PJOBSA	Principal Scales	Configural
Job satisfaction with work environment, principal	T3PJSENV	Principal Scales	Configural
Job satisfaction with profession, principal	T3PJSPRO	Principal Scales	Configural
Lack of special needs personnel	T3PLACSN	Principal Scales	Metric
Participation among stakeholders, principals	T3PLEADP	Principal Scales	Metric
School leadership	T3PLEADS	Principal Scales	Metric
Organisational innovativeness	T3PORGIN	Principal Scales	Configural
Workload stress	T3PWLOAD	Principal Scales	Configural
Clarity of instruction	T3CLAIN	Teacher Scales	Metric
Classroom management	T3CLASM	Teacher Scales	Configural
Cognitive activation	T3COGAC	Teacher Scales	Metric
Professional collaboration in lessons among teachers	T3COLES	Teacher Scales	Metric

(*continued on next page*)

Table 9 (*continued*)

Scale Label	Variable Name	Scale Type	Invariance Level
Teacher cooperation, overall	T3COOP	Teacher Scales	Configural
Teachers perceived disciplinary climate	T3DISC	Teacher Scales	Metric
Diversity practices, teacher	T3DIVP	Teacher Scales	Configural
Effective professional development	T3EFFPD	Teacher Scales	Configural
Exchange and cooperation among teachers	T3EXCH	Teacher Scales	Configural
Job satisfaction, overall, teacher	T3JOBSA	Teacher Scales	Metric
Job satisfaction with work environment, teacher	T3JSENV	Teacher Scales	Metric
Job satisfaction with profession, teacher	T3JSPRO	Teacher Scales	Metric
Professional development barriers	T3PDBAR	Teacher Scales	Configural
Need prof. devel. for teaching for diversity	T3PDIV	Teacher Scales	Metric
Need prof. devel. in subject matter and pedagogy	T3PDPED	Teacher Scales	Metric
Personal utility value	T3PERUT	Teacher Scales	Metric
Satisfaction with target class autonomy	T3SATAT	Teacher Scales	Metric
Self-efficacy in classroom management	T3SECLS	Teacher Scales	Metric
Self-efficacy in student engagement	T3SEENG	Teacher Scales	Metric
Self-related efficacy in multicultural classrooms	T3SEFE	Teacher Scales	Metric
Self-efficacy in instruction	T3SEINS	Teacher Scales	Metric
Teacher self-efficacy, overall	T3SELF	Teacher Scales	Metric
Social utility value	T3SOCUT	Teacher Scales	Metric
Participation among stakeholders, teachers	T3STAKE	Teacher Scales	Metric
Student behaviour stress	T3STBEH	Teacher Scales	Configural
Teacher-student relations	T3STUD	Teacher Scales	Metric
Team innovativeness	T3TEAM	Teacher Scales	Scalar
Teaching practices, overall	T3TPRA	Teacher Scales	Configural
Perceptions of value and policy influence	T3VALP	Teacher Scales	Metric
Workplace well-being and stress	T3WELS	Teacher Scales	Metric
Clarity of instruction	T3WLOAD	Teacher Scales	Metric

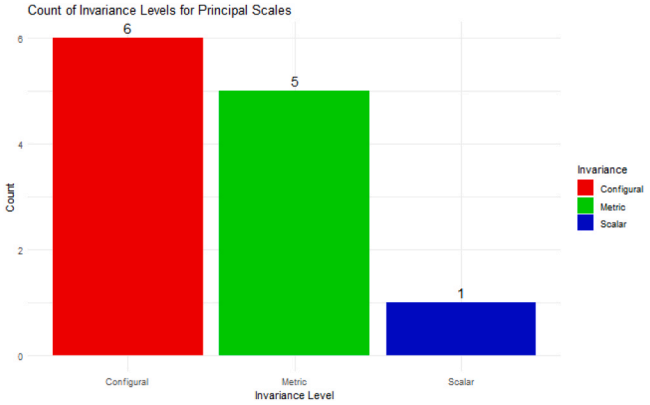

Fig. X. Measurement Invariance Level of Principal Scales in TALIS 2018 Report.

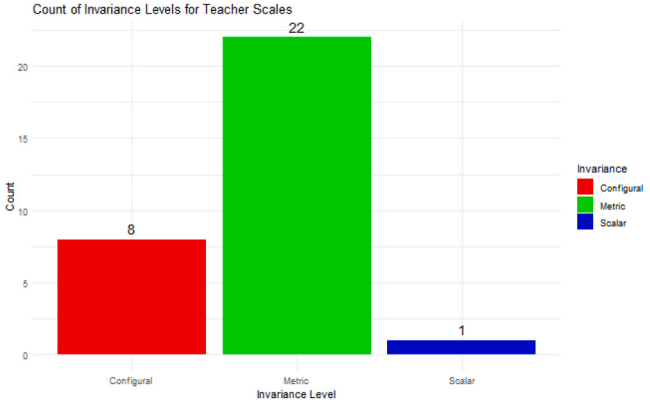

Fig. Y. Measurement Invariance Level of Teacher Scales in TALIS 2018 Report.

References

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

Asparouhov, T., & Muthén, B. (2022). Multiple group alignment for exploratory and structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–23. https://doi.org/10.1080/10705511.2022.2127100

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. *375–393*). Wiley-Blackwell.

Byrne, B. M., & Van de Vijver, F. J. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: a paradigmatic cross-cultural application. *Psicothema, 29*(4), 539–551.

Carrasco, D., Eryilmaz, N., & Sandoval-Hernandez, A. (2024). *rd3c3: R library for Advanced psychometric Analysis*. ⟨https://github.com/dacarras/rd3c3⟩.

Chang, Y. W., Hsu, N. J., & Tsai, R. C. (2017). Unifying differential item functioning in factor analysis for categorical data under a discretization of a normal variant. *Psychometrika, 82*(2), 382–406. https://doi.org/10.1007/s11336-017-9562-0

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods Research, 36*(4), 462–494.

Cieciuch, J., Davidov, E., & Schmidt, P. (2018). Alignment optimization: estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. *571–592*). Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Davidov, E., Muthén, B. O., & Schmidt, P. (2018a). Measurement invariance in cross-national studies: challenging traditional approaches and evaluating new ones. *Sociological Methods Research, 47*(4), 631–636. https://doi.org/10.1177/0049124118789702

Davidov, E., Schmidt, P., Billiet, J., & Meuleman, B. (2018b). *Cross-cultural analysis: methods and applications (2nd ed.* Routledge.

Desa, D., Van de Vijver, F. J. R., Carstens, R., & Schulz, W. (2019). Measurement invariance in international large-scale assessments: integrating theory and method. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology* (pp. 881–910). New York, NY: Wiley.

Eryilmaz, N., & Sandoval Hernandez, A. (2024). Improving cross-cultural comparability: does school leadership mean the same in different countries? *Educational Studies, 50*(5), 917–938.

Ding, Y., Yang Hansen, K., & Klapp, A. (2023). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education, 38*(2), 709–732.

Eryilmaz, N., Rivera-Gutiérrez, M., & Sandoval-Hernández, A. (2020). Should different countries participating in PISA interpret socioeconomic background in the same way? A measurement invariance approach. *Revista Iberoamericana Délelőtt EducaciÓN, 84*(1), 109–133.

Fang, G., Teo, T., & Chan, P. W. K. (2025). Testing for approximate measurement invariance of instructional quality in the Teaching and Learning International Survey (TALIS) 2018. *Humanities and Social Sciences Communications, 12*(1), 1–10.

Fischer, J., Praetorius, A. K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability, 31*, 201–220.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in r. *Frontiers in Psychology, 10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507

Fischer, R., Karl, J. A., Luczak-Roesch, M., & Hartle, L. (2025). Why we need to rethink measurement invariance: the role of measurement invariance for cross-cultural research. *Cross-Cultural Research, 59*(2), 147–179.

Funder, D. C., & Gardiner, G. (2024). MIsgivings about measurement invariance. *European Journal of Personality, 38*(6), 889–895.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: an r package for facilitating Large-Scale latent variable analyses in mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

He, J., Van de Vijver, F., Espinosa, A. D., & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: a multilevel study. *International Journal of Cross-Cultural Management, 14*(3), 306–322. https://doi.org/10.1177/1470595814541424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426.

Kaya, S., Eryilmaz, N., & Yuksel, D. (2024). A cross-cultural comparison of self-efficacy as a resilience measure: evidence from PISA 2018. *Youth Society, 56*(3), 597–621.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods Research, 44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kite, B. A., Jorgensen, T. D., & Chen, P. Y. (2018). Random permutation testing applied to measurement invariance testing with Ordered-Categorical indicators. *Structural Equation Modeling, 25*(4), 573–587. https://doi.org/10.1080/10705511.2017.1421467

Klieme, E. (2020). Policies and practices of assessment: a showcase for the use (and Misuse) of international Large-Scale assessments in educational effectiveness research. In In. J. Hall, A. Lindorff, & P. Sammons (Eds.), *International Perspectives in Educational Effectiveness Research* (pp. 147–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-44810-3_7.

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

Lamm, R., Do, T., & Rodriguez, M. C. (2019). April). Measurement invariance of an internationaldevelopmental assets measure: alignment of 29 countries [Paper presentation]. annual meetingof The National council on measurement in education. *Toronto, Canada*.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3). https://doi.org/10.1037/met0000075

Lubbe, D. (2019). Parallel analysis with categorical variables: impact of category probability proportions on dimensionality assessment accuracy. *Psychological Methods, 24*(3), 339.

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: a tutorial for transparent analysis planning and reporting. *Psychological Methods, 28*(4), 905.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2013). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 25*(1), 107–126. https://doi.org/10.1037/a0023313

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., & Muthén, B. (2018). What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524–545. https://doi.org/10.1037/met0000113

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research, 47*(4), 687-728.

Muthén, B., Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups.* ⟨http://www.statmodel.com/download/⟩.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology, 5*, 978.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods Research, 47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Odell, B., Gierl, M., & Cutumisu, M. (2021). Testing measurement invariance of PISA 2015 mathematics, science, and ICT scales using the alignment method. *Studies in Educational Evaluation, 68*, 100965.

OECD. (2020). *TALIS 2018 technical report*. OECD Publishing. https://doi.org/10.1787/799337c2-en

Padgett, R. N. (2023). A Tutorial on Cross Wave Measurement Invariance Testing with Item Factor Analysis. *Practical Assessment. Research & Evaluation, 28*, 13.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling, 26*(5), 724–744.

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2020). Lp loss functions in invariance alignment and haberman linking with few or many groups. *Stats, 3*(3), 246–283.

Robitzsch, A. (2022b). On the bias in confirmatory factor analysis when treating discrete variables as ordinal instead of continuous. *Axioms, 11*(4), 162.

Robitzsch, A. (2023). Implementation aspects in invariance alignment. *Stats, 6*(4), 1160–1178.

Robitzsch, A. (2024). Examining differences of invariance alignment in the mplus software and the r package sirt. *Mathematics, 12*(5), 770.

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling, 0*(0), 1–12. https://doi.org/10.1080/10705511.2023.2191292

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*. https://doi.org/10.1037/met0000503

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and psychological Measurement, 74*(1), 31–57.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international large-scale assessments: CFA versus IRT. *Applied Measurement in Education, 30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243537

Sandoval-Hernandez, A., Carasco, D., & Eryilmaz, N. (2025). Alignment optimization in international Large-Scale assessments: a scoping review and future directions. *Educational Methods Psychometrics, 3*(16), 16.

Senden, B., Teig, N., & Nilsen, T. (2023). Studying the comparability of student perceptions of teaching quality across 38 countries. *International Journal of Educational Research Open, 5*, Article 100309.

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: an application of multilevel bifactor structural equation modeling. *Frontiers in Psychology, 6*, 1550. https://doi.org/10.3389/fpsyg.2015.01550

Scherer, R., & Nilsen, T. (2016). The relations among school climate, instructional quality, and achievement motivation in mathematics. In In. T. Nilsen, & J.-E. Gustafsson (Eds.), *Teacher Quality, Instructional Quality and Student Outcomes* (pp. 51–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-41252-8_3.

Seddig, D., & Lomazzi, V. (2019). Using cultural and structural indicators to explain measurement noninvariance in gender role attitudes with multilevel structural equation modeling. *Social Science Research, 84*, Article 102328.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: what's old, what's new, what's next? *Organizational Research Methods, 25*(4), 741–785.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group invariance with categorical outcomes using updated guidelines: an illustration using mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Thissen, D. (2024). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research, 0*(0), 1–25. https://doi.org/10.1080/00273171.2024.2396148

Treviño, E., Sandoval-Hernández, A., Miranda, D., Rutkowski, D., & Matta, T. (2021). Invariance of socioeconomic status scales in international studies. In J. Manzi, M. R. García, & S. Taut (Eds.), *Validity of Educational Assessments in Chile and Latin America.* Cham: Springer. https://doi.org/10.1007/978-3-030-78390-7_10.

Tse, W. W. Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. *Behavior Research Methods, 56*(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

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), 4–70. https://doi.org/10.1177/109442810031002

van de Vijver, F. J. R., Fischer, R., & Fontaine, J. R. J. (2019). *Methods for the analysis of cross-cultural data.* Cambridge University Press.

Van de Vijver, F. J. R., Van de, Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Donné, N., Le, Lek, K., Meuleman, B., Paccagnella, M., & Schoot, R. van de (2019). Invariance analyses in large-scale studies. *OECD Education Working Papers, 201*, 1–110. https://doi.org/10.1787/19939019

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Applied Psychological Measurement, 42*(6), 428–445. https://doi.org/10.1177/0146621617748321

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies, 49*(8), 1068–1094.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA press.

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

Wurster, S. (2022). Measurement invariance of non-cognitive measures in TIMSS across countries and across time. An application and comparison of Multigroup Confirmatory Factor Analysis, Bayesian approximate measurement invariance and alignment optimization approach. *Studies in Educational Evaluation, 73*, 101143.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the effect of model misspecification. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2