Enhancing Cross-Cultural Comparisons in ILSAs: A Comprehensive Investigation and Guideline Development

Complementary studies

Andrés Sandoval-Hernández[1]

Diego Carrasco[2]

Nurullah Eryilmaz[3]

§

University of Bath | Pontificia Universidad Católica de Chile | IEA

General Abstract

Abstract

This research integrates findings from two complementary studies to advance understanding of alignment optimization as a method for addressing measurement invariance challenges in international large-scale assessments (ILSAs). **Study 1** provides a comprehensive scoping review of 18 empirical studies across assessments, including PISA, TALIS, TIMSS, and ICCS, analyzing the application of alignment optimization for enhancing cross-national comparability. The review reveals promising yet inconsistent outcomes, particularly in culturally and linguistically diverse contexts, and highlights methodological limitations and future directions for strengthening alignment approaches. **Study 2** presents systemic evidence from TALIS 2018, examining the dimensionality, reliability, and cross-national comparability of teacher and principal scales. While classical multiple-group confirmatory factor analysis (MG-CFA) demonstrated limited scalar-level invariance for many scales, alignment optimization successfully enabled comparability for all principal scales but failed to meet thresholds for many teacher scales. Together, these studies underscore the potential of alignment optimization to improve comparability, especially for principal constructs, while identifying persistent challenges for teacher scales. This work contributes to the theoretical and practical discourse on ILSA methodologies, offering actionable recommendations for achieving more valid and equitable educational assessments globally.

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

Study 1: Alignment Optimization in International Large-Scale Assessments: A Scoping Review and Future Directions

Published in:
Sandoval-Hernández, A., Carrasco, D., & Eryilmaz, N. (2025). Alignment Optimization in International Large-Scale Assessments: A Scoping Review and Future Directions. *Educational Methods & Psychometrics*, *3*(16). https://doi.org/10.61186/emp.2025.3

Abstract

This scoping review critically examines the application of alignment optimization in international large-scale assessments (ILSAs) to address measurement invariance challenges. By systematically analyzing 18 empirical studies across assessments, including PISA, TALIS, TIMSS, and ICCS, and incorporating data from student, teacher, and principal questionnaires, this review provides a comprehensive synthesis assessment of the method's effectiveness in enhancing cross-national comparability. While alignment optimization shows promise, findings reveal its inconsistent performance in achieving measurement invariance across diverse cultural and linguistic contexts. The review highlights specific limitations of the method, especially where cultural and linguistic diversity influence item interpretation, and proposes a set of targeted recommendations for advancing methodological robustness in future research. This study contributes to the theoretical and practical discourse on improving the validity of educational assessments globally, offering a foundation for more equitable and reliable cross-national comparisons in education research.

Keywords: Alignment Optimization, International Large-Scale Assessments (ILSAs), Measurement Invariance, Cross-National Comparability, Educational Assessment

1. Introduction

International large-scale assessments (ILSAs) such as the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), the Teaching and Learning International Survey (TALIS), and the International Civic and Citizenship Education Study (ICCS) have become invaluable resources in the global landscape of educational research and policy evaluation. By examining educational outcomes across multiple countries, these assessments provide a unique opportunity to evaluate and compare student knowledge, skills, and attitudes within a broad cross-national framework. ILSAs have increasingly informed educational reform efforts by identifying factors that may contribute to achievement disparities, educational equity, and the development of lifelong learning competencies (Mullis et al., 2020; OECD, 2019). However, a primary methodological challenge underlying the analysis of ILSA data is the assurance of *measurement invariance*—a statistical property indicating that a construct holds the same meaning across diverse populations. Without achieving measurement invariance, cross-national comparisons of student, teacher, or principal responses to items can yield biased or misleading interpretations, ultimately affecting the validity of policy recommendations derived from these comparisons (Putnick & Bornstein, 2016; Rutkowski & Svetina, 2017). As Van de Vijver, F., et al. (2019) highlight, measuring complex constructs in global educational assessments presents significant methodological challenges.

Traditional methods like Multiple-Group Confirmatory Factor Analysis (MG-CFA) have been the primary tools used to assess measurement invariance in ILSA studies. MG-CFA tests invariance across hierarchical

levels—configural, metric, and scalar—by constraining parameters to test if the model fits comparably across groups. Configural invariance establishes that items load on the same factors across groups, metric invariance tests equality of factor loadings, and scalar invariance further constrains item intercepts to be equal across groups. Despite their rigorous structure, these traditional approaches often face limitations in culturally heterogeneous settings. Specifically, achieving scalar invariance in cross-national studies can be challenging due to differences in response styles, language nuances, and educational contexts that may influence how items are understood and endorsed (Eryilmaz et al., 2020; Eryilmaz & Sandoval Hernandez, 2024; Kaya *et al.,* 2024; Davidov et al., 2014; Chen, 2008; Treviño *et al.,* 2021; Sandoval-Hernández *et al.,* 2019). In such cases, the need for approximate invariance—where the model allows some level of variability in item parameters—is increasingly recognized to facilitate more realistic interpretations of cross-cultural data.

To address these challenges, Asparouhov and Muthén (2014) proposed the *alignment optimization* method, which enables approximate measurement invariance by relaxing some constraints on item parameters. Unlike traditional approaches that necessitate strict invariance across all items, alignment optimization seeks to balance invariance and model fit by allowing for a proportion of item non-invariance while maintaining comparability. This method, also referred to as the "alignment" approach, thus aims to preserve the validity of latent mean comparisons across groups even when some level of non-invariance exists. In recent years, alignment optimization has gained traction in ILSA research as a means to conduct more flexible and pragmatic invariance testing, particularly in studies that involve diverse samples such as PISA, TIMSS, TALIS, and ICCS data (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2018).

The alignment optimization approach is grounded in an exploratory multiple-group factor analysis aimed at identifying an optimal measurement invariance design. This method seeks to ensure satisfactory estimation of group means and variances, even in the presence of non-invariance, thereby enabling meaningful group mean comparisons (Asparouhov & Muthén, 2014). The process involves two main stages. Initially, a configural invariance model is applied across groups, where factor loadings and intercepts are allowed to vary freely, while factor means are constrained to zero, and factor variances are fixed at one. In the second stage, factor means and variances are released, allowing the estimation of these parameters in each group without assuming full measurement invariance. This step simultaneously identifies the optimal measurement invariance pattern by incorporating a simplicity function, akin to rotation criteria in exploratory factor analysis (EFA) (Asparouhov & Muthén, 2014, p. 496).

After the alignment estimation is completed, a detailed review is conducted to identify which measurement parameters meet approximate invariance and which do not. This is achieved through both visualizations and calculations of misfit in factor loadings and intercepts. To assess the reliability and quality of the alignment results, a threshold of no more than 25% non-invariance is recommended, as outlined by Asparouhov and Muthén (2014) and Muthén & Asparouhov (2014).

The increasing application of alignment optimization in ILSA studies has produced promising yet mixed results. For example, studies on constructs such as teaching quality, intrinsic motivation, distributed leadership, and ICT readiness have utilized alignment optimization to achieve partial or approximate invariance (Eryilmaz & Sandoval-Hernandez, 2023; Zakariya 2021), facilitating more nuanced insights across countries (Rutkowski et al., 2019; Jerrim & Micklewright, 2014). However, a critical review of these applications reveals substantial variability in the effectiveness of alignment optimization, especially

when applied to culturally sensitive constructs like political trust and intrinsic motivation. This variability raises questions regarding the reliability of alignment optimization as a universal solution for invariance testing in cross-national contexts, particularly when comparing latent means of constructs that may be interpreted differently across cultural and linguistic boundaries.

This scoping review aims to map the application of alignment optimization in ILSAs, providing a comprehensive overview of its usage across different assessments, constructs, and respondent types (e.g., students, teachers, principals). By synthesizing findings from multiple studies, we seek to evaluate the practical effectiveness of alignment optimization, outline its limitations, and identify areas requiring further methodological development. Through this analysis, the review contributes to ongoing discussions regarding the robustness of alignment optimization as a tool for achieving cross-national comparability in educational research, emphasizing the need for additional empirical testing and methodological refinement to address the complex landscape of cultural and contextual variability in ILSA data.

Objective of the Review

The principal aim of this scoping review is to critically examine the application and efficacy of alignment optimization in addressing measurement invariance challenges within ILSAs. As alignment optimization is increasingly adopted to enhance the comparability of latent constructs across culturally and linguistically diverse contexts, understanding the breadth and effectiveness of its application is essential for refining cross-national comparisons in educational research. This review, therefore, seeks to provide a comprehensive synthesis of alignment optimization's use across ILSAs such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Teaching and Learning International Survey (TALIS), and International Civic and Citizenship Education Study (ICCS). Through this synthesis, the review aims to contribute to the broader discourse on the methodological rigor and limitations inherent in measuring complex constructs in global educational assessments.

More specifically, the objectives of this review are fourfold:

1. *Summarize the Constructs and Contexts in which Alignment Optimization Has Been Utilized:* This review will provide an overview of the various constructs—ranging from teaching quality and political trust to ICT readiness and intrinsic motivation—examined in ILSAs where alignment optimization has been employed. Additionally, we will assess how alignment optimization has been applied across different respondent groups, including students, teachers, and principals, and within different survey types such as student and teacher questionnaires. By cataloging the diversity of applications, we aim to map the extent to which alignment optimization has facilitated the comparability of educational constructs across the heterogeneous cultural, linguistic, and socio-economic settings of participating countries.
2. *Assess the Effectiveness of Alignment Optimization in Achieving Measurement Invariance Across Diverse Cultural and Linguistic Groups:* As measurement invariance is foundational for valid cross-national comparisons, it is essential to evaluate alignment optimization's success in meeting this requirement. This review will analyze the degree to which alignment optimization has succeeded in achieving configural, metric, and scalar invariance for various constructs, as well as the challenges encountered in attaining these levels across multiple ILSA studies. We will

scrutinize the outcomes of alignment optimization in terms of its ability to approximate invariance, thereby allowing partial non-invariance while still enabling latent mean comparisons. This objective will provide a critical assessment of alignment optimization's effectiveness, highlighting cases where it has provided robust comparability as well as those where significant non-invariance remains.

3. *Identify Limitations and Challenges Associated with Alignment Optimization:* Despite its utility, alignment optimization has limitations, particularly when applied to constructs influenced by distinct cultural and linguistic factors. This review will explore common challenges faced in applying alignment optimization, including high levels of non-invariance in culturally sensitive constructs and inconsistencies in reporting non-invariance percentages across studies. The review will also examine specific methodological and practical challenges, such as the potential for interpretive bias when approximate invariance thresholds are applied. In doing so, we aim to underscore the contexts where alignment optimization may fall short and to outline the limitations researchers should consider when interpreting cross-national comparisons derived from alignment-optimized models.
4. *Provide Recommendations for Future Research and Methodological Improvements:* Building on the synthesized findings, this review will offer targeted recommendations for future research on alignment optimization in ILSAs. These recommendations will include methodological suggestions for enhancing the robustness of alignment optimization, such as integrating alternative approaches like Bayesian approximate invariance testing and utilizing machine learning to detect patterns of non-invariance. Additionally, we will advocate for improved reporting standards, including consistent documentation of non-invariance rates and clear justifications for acceptable thresholds in cross-national analyses. Furthermore, we will propose avenues for exploring alignment optimization's applicability to less commonly studied constructs and suggest design strategies to account for cultural and linguistic influences on measurement invariance. By providing these recommendations, we aim to support the advancement of alignment optimization as a methodologically sound approach for improving the comparability of educational measures across global contexts.

In summary, this scoping review intends to present a detailed assessment of alignment optimization's application in ILSAs, with a focus on examining its scope, effectiveness, limitations, and potential for methodological enhancement. Through this analysis, we aim to contribute to the refinement of tools for international educational research, promoting more reliable and valid cross-national comparisons that can inform educational policy and practice on a global scale.

3. Methodology

This scoping review follows the methodological framework established by Arksey and O'Malley (2005) and is also informed by the procedures for data analysis and interpretation outlined by Van de Vijver and colleagues (2019), which offer a structured approach to exploring and synthesizing research on alignment optimization in International Large-Scale Assessments (ILSAs). The review process encompasses the following stages:

1. Identifying the Research Question

The central research question guiding this review is: "How has alignment optimization been applied in ILSAs to address measurement invariance challenges?" This inquiry aims to examine the contexts, constructs, and effectiveness of alignment optimization in enhancing the comparability of educational measures across diverse cultural and linguistic groups within ILSAs. By focusing on alignment optimization's application, the review seeks to elucidate its role in overcoming the limitations of traditional measurement invariance methods, such as Multiple-Group Confirmatory Factor Analysis (MG-CFA), which often encounter difficulties in achieving full invariance in multicultural contexts.

2. Identifying Relevant Studies

A comprehensive literature search was conducted across academic databases, including ERIC, PsycINFO, Web of Science, and Google Scholar. Search terms were selected to capture studies focused on alignment optimization, measurement invariance, and specific ILSAs such as TIMSS, PISA, TALIS, and ICCS. Keywords and Boolean operators were used in various combinations, including terms such as "alignment optimization," "measurement invariance," "cross-national comparability," and "large-scale assessment." The initial search was supplemented by reviewing references cited in relevant articles to identify additional studies not surfaced in the database search. The search process was iterative, with periodic adjustments to keywords and search terms as new insights emerged from preliminary results.

3. Study Selection

The selection process involved screening studies based on predetermined inclusion and exclusion criteria. Studies were eligible for inclusion if they (1) applied alignment optimization in the context of an ILSA, (2) specifically addressed issues related to measurement invariance across diverse national, cultural, or linguistic groups, (3) were published in peer-reviewed journals, and (4) were written in English. Exclusion criteria included studies focusing on alignment optimization in non-educational fields or using traditional measurement invariance techniques without integrating alignment optimization. Abstracts and titles were initially screened for relevance, followed by a full-text review to ensure each study met the inclusion criteria. This two-step screening process was designed to enhance reliability and reduce the risk of bias in study selection.

4. Charting the Data

A data extraction form was developed to systematically chart relevant information from each study. Extracted data included study characteristics (e.g., year, authorship, and publication source), constructs examined (e.g., teaching quality, intrinsic motivation, political trust, ICT readiness), ILSA survey type (e.g., student, teacher, or principal questionnaires), and the specific alignment optimization methods applied. Information on key findings related to measurement invariance levels (e.g., configural, metric, and scalar invariance) and non-invariance percentages were also recorded. The data extraction process aimed to capture both quantitative and qualitative aspects of each study, allowing for a thorough analysis of the methods and outcomes of alignment optimization in various international educational settings. By charting these variables, the review seeks to produce a structured overview that highlights patterns and differences in alignment optimization's application across constructs and assessments (see table 1).

5. Collating, Summarizing, and Reporting the Results

In the final stage, findings were synthesized thematically to address the research objectives and answer the guiding research question. Thematic analysis involved categorizing studies based on constructs, ILSA

survey types, and measurement invariance outcomes. Key themes that emerged included the consistency of alignment optimization across different constructs, challenges posed by cultural and linguistic diversity, and the extent to which alignment optimization succeeded in achieving approximate measurement invariance across groups. A critical evaluation of non-invariance percentages was also conducted, with comparisons drawn between studies that achieved acceptable levels of non-invariance and those with higher levels of misalignment. This synthesis provided insights into the strengths and limitations of alignment optimization in cross-national contexts, informing recommendations for its future use and potential methodological enhancements. In reporting the results, visual aids such as summary tables and figures were employed to present the findings clearly and systematically, facilitating a deeper understanding of alignment optimization's application across the diverse educational settings encompassed by ILSAs.

Through this structured approach, the review offers a thorough examination of alignment optimization's role in international educational assessments, identifying methodological gaps and providing recommendations for future research to enhance the robustness and validity of cross-national comparisons in education.

Table 1. The studies used in this review

Number	*Year*	*Study_Name*	*ILSA (study)*	*Target*	*MG-CFA*	*Alignment Optimization*
1	2023	School and Teacher Information, Communication and Technology (ICT) readiness across 57 countries: The alignment optimization method	PISA	teacher and school	N/A	Below 25% (School ICT Readiness), Above 25% (Teacher ICT Readiness)
2	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	TIMSS	principal, teacher and student	Configural	Above 25% (for some models)
3	2023	Is Distributed Leadership Universal? A Cross-Cultural, Comparative Approach across 40 Countries: An Alignment Optimization Approach	TALIS	principal	Metric	Below 25%
4	2021	Assessing the comparability of teacher-related constructs in TIMSS 2015 across 46 education systems: an alignment optimization approach	TIMSS	teacher	Metric (for some constructs)	Below 25% (for most constructs)
5	2024	Cross-National Measurement of Mathematics Intrinsic Motivation: An Investigate of Measurement Invariance with MG-CFA and Alignment Method Across Fourteen Countries	TIMSS	students	Configural (4^{th}grade), Metric (8^{th}grade)	Below 25%
6	2021	The evolution of the association between ICT use and reading achievement in 28 countries	PISA	students	N/A	N/A (Monte-Carlo Simulation)
7	2024	Determinants of trust in order and representative institutions among adolescents across Europe: testing rational and cultural approaches	ICCS	students	Configural	Below 25%
8	2024	Political trust among European youth: Evaluating multi-dimensionality and cross-national measurement comparability	ICCS	students	Metric	Below 25%
9	2021	Tailoring a measurement model of socioeconomic status: Applying the alignment optimization method to 15 years of PISA	PISA	students	Metric	Below 25%

10	2024	Measuring and explaining political tolerance among adolescents: insights from the International Civic and Citizenship Education Study2016	ICCS	students	Configural	Below 25%
11	2019	The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality	PISA	students	N/A	Above 25%
12	2020	Measurement Invariance Testing with Alignment Method: Many Groups Comparison	PISA	students	Configural	Above 25%
13	2023	Studying the comparability of student perceptions of teaching quality across 38 countries	TIMSS	students	Metric	Above 25%
14	2022	Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method	PISA	students	Metric	Above 25%
15	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	ICCS, CIVED	students	N/A	Below 25%
16	2020	Teacher job satisfaction across 38 countries and economies: An alignment optimization approach to a cross-cultural mean comparison	TALIS	teachers	Metric	Above 25%
17	2021	Application of an innovative alignment optimisation method to a cross-cultural mean comparison of teacher self-efficacy: A cross-country study	TALIS	teachers	Metric	Above 25%
18	2024	Measurement invariance across immigrant and nonimmigrant populations on PISA non-cognitive scales	PISA	students	N/A	Above 25%

4. Results

Upon reviewing the 18 studies, it was observed that five achieved configural invariance, indicating a consistent factor structure across groups. Eight studies reached metric invariance, demonstrating equal factor loadings across groups. Notably, none of the studies attained scalar invariance, which would require equal item intercepts across groups. Five studies did not report on these levels of measurement invariance.

Regarding alignment optimization results, nine studies reported non-invariance levels below the 25% threshold, suggesting acceptable cross-group comparability. In contrast, eight studies exceeded this threshold, indicating significant measurement non-invariance and potential challenges in cross-cultural comparisons. One study did not provide alignment optimization results. They evaluated alignment optimization using Monte Carlo Simulation Studies (Monte Carlo).

These findings highlight the complexities involved in achieving measurement invariance across diverse cultural contexts, underscoring the need for careful consideration when interpreting cross-national assessments.

4.1. Thematic Findings from the Review of Alignment Optimization Studies

The analysis of 18 studies applying alignment optimization in ILSAs revealed several key themes regarding measurement invariance and the method's effectiveness. These themes reflect the complexities and challenges of achieving cross-national comparability in diverse cultural and linguistic contexts.

4.1.1. Levels of Measurement Invariance Across Constructs

One of the foundational objectives of alignment optimization is to address measurement invariance at different levels—configural, metric, and scalar. These levels serve as benchmarks for the consistency of constructs across groups, such as countries or regions (see Figure 1).

- *Configural Invariance:* Five studies achieved configural invariance, confirming that the constructs share a similar factor structure across groups. Constructs such as teaching quality and intrinsic motivation demonstrated a consistent underlying structure across countries. This outcome provides a foundational level of comparability but is insufficient for comparing factor loadings or latent means.
- *Metric Invariance:* Eight studies reached metric invariance, enabling the comparison of factor loadings across groups. Constructs such as distributed leadership, teacher job satisfaction, and some dimensions of teaching quality successfully achieved this level, indicating that the relationships between items and their latent constructs were similar across contexts.
- *Scalar Invariance:* None of the reviewed studies attained scalar invariance. Scalar invariance requires equality in item intercepts, allowing for valid comparisons of latent means. The absence of scalar invariance across all studies highlights the significant challenges in ensuring full comparability of constructs in multicultural assessments.
- *Unreported Invariance Levels:* Five studies did not explicitly report on measurement invariance levels, limiting the ability to evaluate their success in addressing comparability challenges. This

gap in reporting highlights the need for greater transparency in the application of alignment optimization.

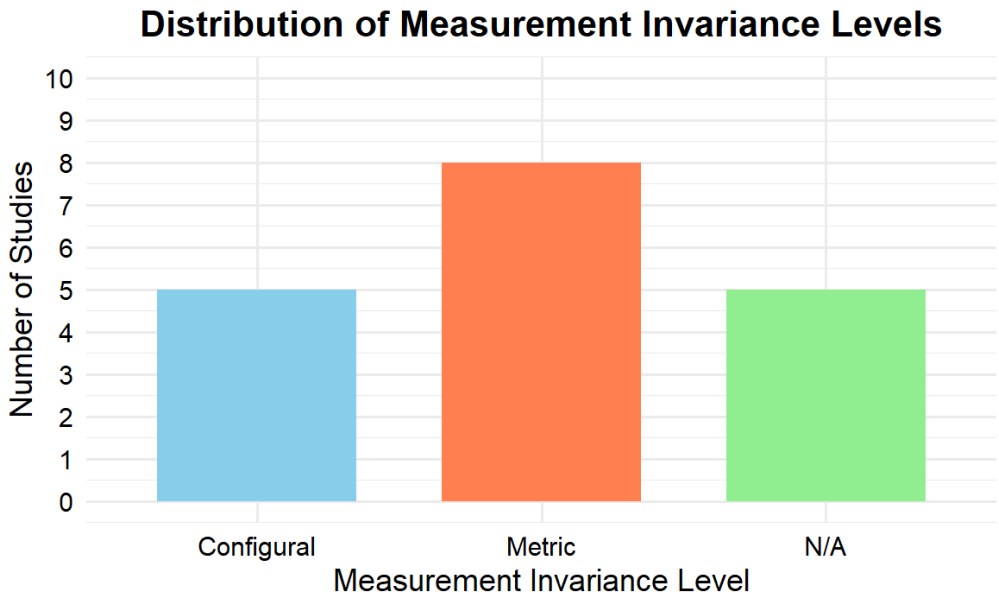

Figure 1. The distribution of Invariance Level

4.1.2. Alignment Optimization: Successes and Challenges

Alignment optimization has been proposed as an alternative to traditional methods, particularly in cases where scalar invariance cannot be achieved. The review of these studies provided insights into its effectiveness and limitations (see Figure 2).

- *Studies Below 25% Non-Invariance Threshold:* Nine studies reported alignment optimization results with non-invariance levels below the 25% threshold, suggesting acceptable cross-group comparability. For example, distributed leadership (8.75% non-invariance) and political trust (16% non-invariance) demonstrated relatively low levels of item misalignment, indicating that these constructs could be meaningfully compared across countries.
- *Studies Above 25% Non-Invariance Threshold:* Eight studies exceeded the 25% threshold, revealing significant challenges in achieving cross-national comparability. Constructs such as teacher self-efficacy (44% non-invariance) and teacher job satisfaction (46% non-invariance) showed high variability in how items were interpreted across cultural contexts. These findings highlight the limitations of alignment optimization in mitigating cultural and linguistic differences for certain constructs.
- *Monte Carlo Simulations:* One study utilized Monte Carlo simulations to explore non-invariance patterns, offering a complementary methodological approach. However, this study did not report specific non-invariance percentages, leaving some uncertainty about the effectiveness of alignment optimization in these contexts.

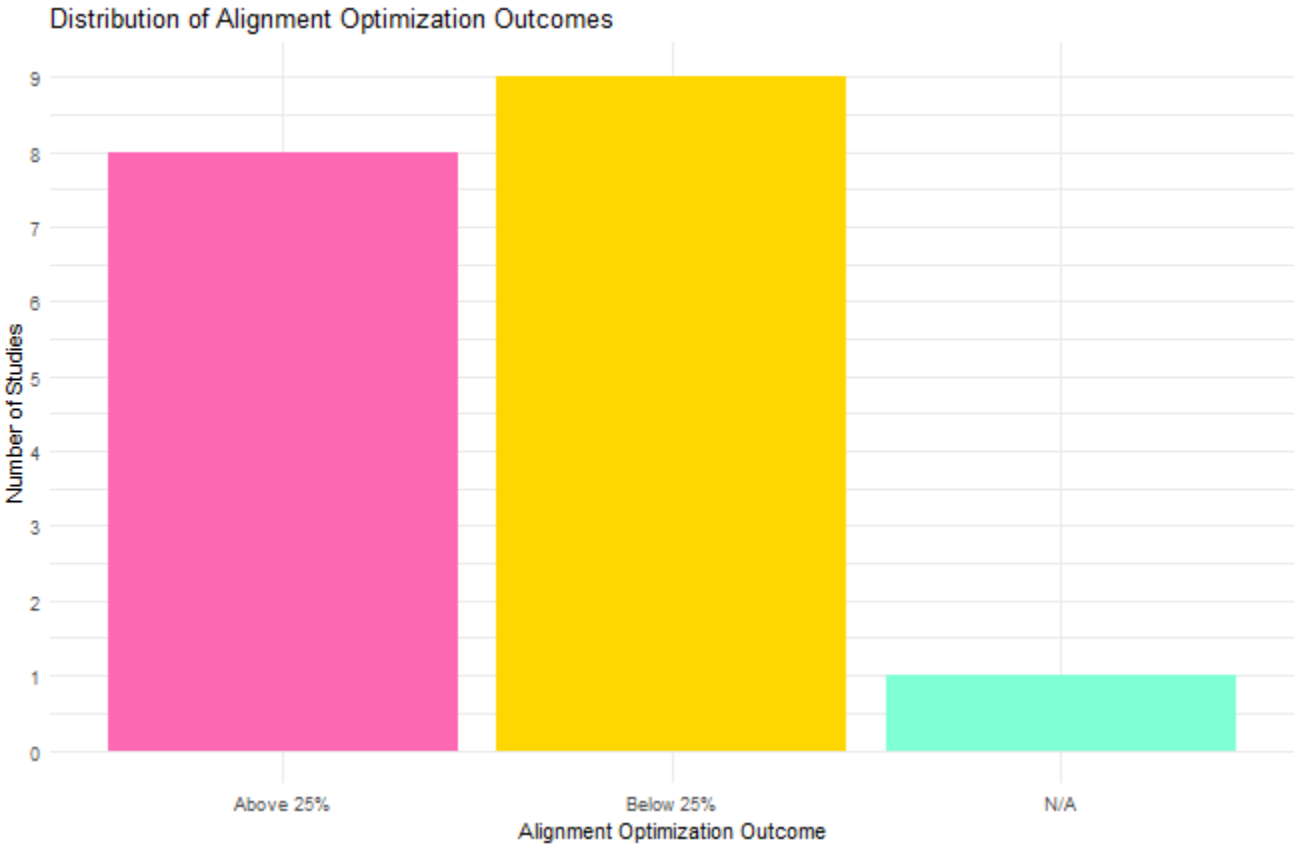

Figure 2. The Distribution of Alignment Optimization

4.1.3. Construct-Specific Challenges

The studies covered a wide range of constructs, revealing varying degrees of success in achieving measurement invariance. There are some examples of the following constructs were included in the studies:

- *Teaching Quality:* Constructs related to teaching quality, such as behavioral management and instructional clarity, exhibited high non-invariance rates (49.6% and 32.9%, respectively). This suggests significant challenges in measuring these constructs consistently across countries, likely due to cultural differences in how teaching practices are perceived.
- *ICT Readiness:* School ICT readiness achieved below-threshold non-invariance (8%), indicating a relatively high level of comparability. In contrast, teacher ICT readiness exhibited 30.1% non-invariance, reflecting more variability in item interpretation for this construct.
- *Student Belonging and Bullying:* The "Sense of Belonging at School" scale was close to the 25% threshold (25.1%), while the "Bullying" scale surpassed it. These findings highlight cultural differences in how students perceive, and report experiences related to school belonging and bullying, complicating cross-national comparisons.

4.1.4. Target Groups and Assessment Contexts

The studies spanned a diverse range of target populations and assessment contexts, showcasing the versatility of alignment optimization. The following populations and studies emerged:

- *Target Populations:* The studies included data from students, teachers, principals, and schools. Constructs ranged from socio-emotional measures like political trust and bullying to professional measures such as teacher self-efficacy and job satisfaction. This diversity highlights the broad applicability of alignment optimization in ILSAs.
- *ILSAs Represented: The* studies primarily focused on well-established ILSAs, including PISA, TIMSS, TALIS, and ICCS. Each assessment brought unique challenges and opportunities for applying alignment optimization, depending on the constructs measured and the populations surveyed.

4.1.5. Cultural and Linguistic Influences

Several studies underscored the significant role of cultural and linguistic diversity in shaping alignment results.

- *Linguistic Similarity:* In some studies, linguistic similarity between groups improved comparability. For example, teaching quality assessments showed greater alignment when linguistic differences were minimal.
- *Cultural Differences:* Socio-emotional measures, such as political trust and bullying, exhibited more pronounced cultural differences. These differences often led to higher levels of non-invariance, reflecting varying interpretations of survey items across countries.

4.1.6. Methodological Variability

The studies employed a range of methodological approaches, highlighting both the potential and limitations of alignment optimization.

- *Traditional vs. Alignment Optimization:* Traditional MG-CFA methods often achieved configural or metric invariance but struggled with scalar invariance. Alignment optimization allowed for approximate invariance in some cases, but high non-invariance levels persisted for several constructs, particularly those influenced by cultural and linguistic factors.
- *Item-Level Analysis***:** Some studies adopted innovative item-level analysis approaches, such as applying a 25% non-invariance threshold for individual items. This method provided a nuanced understanding of where misalignment occurred, aiding in the interpretation of results.

The review of these 18 studies reveals the complexities of achieving measurement invariance in ILSAs and the mixed effectiveness of alignment optimization. While the method shows promise in addressing cross-cultural comparability challenges, persistent issues with scalar invariance and high non-invariance levels for certain constructs highlight its limitations. The findings underscore the need for continued methodological innovation, greater transparency in reporting, and culturally sensitive approaches to improve the robustness of cross-national educational assessments.

5. Research Gap Analysis on the Use of Alignment Optimization for Cross-National and Longitudinal Comparability

The application of alignment optimization in addressing measurement invariance within International Large-Scale Assessments (ILSAs) presents both promising advancements and notable challenges. Traditional methods, such as Multi-Group Confirmatory Factor Analysis (MG-CFA), often fail to achieve scalar invariance, particularly when comparing latent constructs across diverse educational and cultural settings. For instance, studies on teaching quality, mathematics intrinsic motivation, and mathematics self-concept/self-efficacy have struggled to reach full measurement invariance using MG-CFA. In response, alignment optimization has emerged as an alternative approach, allowing for approximate invariance and enabling cross-national comparisons where traditional methods fall short.

Strengths of Alignment Optimization

Alignment optimization has demonstrated effectiveness in achieving approximate invariance for certain constructs, facilitating meaningful cross-national comparisons. For example, constructs such as distributed leadership and political trust have shown low levels of non-invariance, indicating the method's utility in capturing subtle cross-cultural differences while preserving comparability. By permitting a degree of flexibility in parameter invariance, alignment optimization enables researchers to work with complex, multicultural datasets without excluding groups or compromising sample sizes.

Persistent Challenges and Limitations

Despite its advantages, alignment optimization is not universally effective. Studies have reported high levels of non-invariance in constructs like mathematics self-efficacy and self-concept, with non-invariance rates of 48.43% and 56.75%, respectively. Similarly, the Instrumental Motivation scale exhibited a 49.5% non-invariance rate, suggesting limitations in the method's ability to fully account for cross-country differences. These discrepancies often stem from cultural and linguistic differences that influence how respondents interpret survey items and response scales.

Moreover, inconsistencies in reporting alignment optimization results, such as non-invariance percentages and their implications, hinder the method’s transparency and replicability. For instance, some studies utilizing Monte Carlo simulations have omitted specific non-invariance percentages, creating gaps in understanding the full extent of alignment optimization's applicability.

Methodological and Contextual Considerations

The effectiveness of alignment optimization varies by construct and population. Constructs related to socio-emotional factors, such as sense of belonging and bullying, are particularly susceptible to cultural nuances, complicating their alignment across countries. Conversely, constructs with more standardized interpretations, such as ICT readiness and school emphasis on academic success, exhibit higher levels of comparability.

Additionally, the limited representation of non-Western contexts in these studies highlights a critical gap. Most reviewed studies have focused on data from Western or developed countries, leaving open questions about how alignment optimization performs in underrepresented regions such as Africa, South Asia, and Latin America.

Implications for Future Research

To advance the application of alignment optimization, researchers should consider the following:

1. *Diverse Item Groupings and Constructs*: Investigate varied educational constructs, such as academic resilience, digital literacy, or collaborative problem-solving, across international datasets like PISA or TIMSS. Comparing constructs relevant to specific subpopulations, such as teacher leadership styles or student well-being, may reveal varying levels of invariance across cultural and linguistic groups.

2. *Larger and More Heterogeneous Sample Sizes:* Increase participant samples by including data from more countries and larger participant pools. For example, comparing constructs like student engagement using data from over 80 countries in PISA, with participant numbers exceeding 500,000 students, would provide a more comprehensive understanding of measurement invariance across global populations. Additionally, designing studies that focus on subgroups within countries, such as urban vs. rural populations or students from different socio-economic backgrounds, can test how alignment optimization performs when comparing heterogeneous groups.

3. *Multiple Time Points for Longitudinal Comparisons:* Expand analyses by including more time points to track changes in constructs like student motivation or teacher job satisfaction over decades. Utilizing data from PISA cycles spanning 20 years can examine how constructs evolve and test the consistency of alignment optimization over time. Conducting research that compares age cohorts longitudinally, such as analyzing educational outcomes from elementary to secondary school, can assess whether alignment optimization is sensitive to developmental changes.

4. *Diverse Cultural and Linguistic Contexts:* Incorporate more data from underrepresented regions like Africa, South Asia, or Latin America. By using constructs measured in culturally distinct groups, researchers can better assess how alignment optimization handles diverse linguistic and cultural contexts. Conducting studies that test constructs across multiple linguistic groups within a single country, such as Canada (English and French) or Switzerland (German, French, and Italian), can reveal whether alignment optimization is robust within countries with multiple languages and cultural contexts.

5. *Alternative Methodologies and Hybrid Approaches:* Compare alignment optimization with Bayesian methods to determine which approach better handles non-invariance issues. For instance, testing the invariance of teacher professional development constructs using both methods can compare their effectiveness. Utilizing machine learning models to identify patterns of non-invariance and predict which constructs are more likely to exhibit measurement challenges could be particularly useful in large datasets like TIMSS, with thousands of schools and teachers.

6. *Studies Involving Teachers and Principals:* Explore invariance in constructs like teacher instructional practices or principal leadership styles using datasets like TALIS. Studies involving more than 10,000 teachers and principals across over 40 countries would allow for a comprehensive analysis of educational leadership practices.

7. *Use of Simulated Data for Method Comparison:* Design simulation studies to test the effectiveness of alignment optimization under different conditions of non-invariance. This can help refine the method and establish clearer guidelines for its application.

Conclusion

Alignment optimisation has made a valuable contribution to educational measurement by providing a practical approach to address measurement invariance in International Large-Scale Assessments (ILSAs). Its significance lies in offering a more flexible alternative to traditional methods, particularly in contexts where full scalar invariance is unattainable.

The method's primary strength is its capacity to navigate the complexities of cultural and linguistic diversity inherent in ILSAs. By permitting approximate invariance, alignment optimisation enables researchers to make comparisons that would otherwise be impossible using conventional methods such as MG-CFA. However, this flexibility comes with limitations, as its effectiveness varies across constructs and contexts.

Some considerations for future developments could include developing clearer protocols for applying alignment optimisation. Researchers must establish more precise thresholds for acceptable non-invariance and adopt transparent reporting standards to enhance the method's rigour and reproducibility. This can be achieved by extending the scope of application by investigating a wider range of constructs, incorporating more geographically and culturally diverse settings, and examining longitudinal patterns in measurement invariance.

While alignment optimisation offers significant potential, it is not a universal solution. Researchers must apply the method with a clear and explicit understanding of its limitations, particularly for culturally sensitive measures. The variability in performance across constructs—ranging from relatively stable measures like ICT readiness to more complex socio-emotional dimensions—highlights the need for a context-specific approach.

Ultimately, alignment optimisation represents a critical advancement in addressing the methodological challenges of cross-national educational research. By providing a viable alternative to strict invariance testing, it allows for more meaningful comparisons in challenging contexts. However, its use demands careful interpretation, methodological rigour, and a nuanced understanding of cultural diversity.

As the field of international educational assessment evolves, alignment optimisation is set to play an increasingly central role. Advancing this approach will require ongoing research, methodological innovation, and a commitment to engaging with the complexities of both ILSA data and cultural and linguistic diversity. Researchers must remain critical and transparent in their application of this method, ensuring that cross-national comparisons are as accurate and meaningful as possible.

References

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

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., & Muth´en, 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

Borgonovi, F., & Pokropek, M. (2021). The evolution of the association between ICT use and reading achievement in 28 countries. *Computers and Education Open*, *2*, 100047.

Casas, M., & Sireci, S. G. (2024). Measurement invariance across immigrant and nonimmigrant populations on PISA non-cognitive scales. *International Journal of Testing,* 1–20. *Online first*, https://doi.org/10.1080/15305058.2024.2426022

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40,* 55–75.

Ding, Y., Hansen, K. Y., & 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,* 709–732. https://doi.org/10.1007/s10212-022-00623-y

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 de Educación*, *84*(1), 109-133.

Eryilmaz, N., & Sandoval-Hernandez, A. (2023). Is distributed leadership universal? A cross-cultural, comparative approach across 40 Countries: An alignment optimisation approach. *Education Sciences*, *13*(2), 218.

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.

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. https://doi.org/10.1007/s11092-019-09295-7

Glassow, L. N., Rolfe, V., & Hansen, K. Y. (2021). Assessing the comparability of teacher-related constructs in TIMSS 2015 across 46 education systems: An alignment optimization approach. *Educational Assessment, Evaluation and Accountability*, *33*, 105-137. https://doi.org/10.1007/s11092-020-09348-2

Jerrim, J., & Micklewright, J. (2014). Socioeconomic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *International Journal of Educational Research, 66,* 63–78.

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.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

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), 767–799. https://doi.org/10.1177/0049124117729691

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. *Educational Research and Evaluation, 24*(1-2), 14–28.

OECD. (2019). *PISA 2018 Results (Volume I).* OECD Publishing.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39–51.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2019). *Handbook of International Large-Scale Assessment: Background, technical issues, and methods of data analysis.* Springer.

Sandoval-Hernández, A., David, R., Matta, T., & Miranda, D. (2019). Back to the drawing board: Can we compare socioeconomic background scales?. *Revista de Educación*, *2018*(383), 37-61.

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*, 100309.

Sırgancı, G., Uyumaz, G., & Yandı, A. (2020). Measurement invariance testing with alignment method: Many groups comparison. *International Journal of Assessment Tools in Education, 7*(4), 657–673. https://doi.org/10.21449/ijate.714218

Stals, L., & Ziemes, J. F. (2024). Determinants of trust in order and representative institutions among adolescents across Europe: Testing rational and cultural approaches. *Acta Politica, 59,* 1–23. https://doi.org/10.1057/s41269-024-00331-z

Treviño, E., Sandoval-Hernández, A., Miranda, D., Rutkowski, D., & Matta, T. (2021). Invariance of socioeconomic status scales in international studies. *Validity of educational assessments in Chile and Latin America*, 235-257.

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

Yiğiter, M. S. (2024). Cross-National Measurement of Mathematics Intrinsic Motivation: An Investigate of Measurement Invariance with MG-CFA and Aligment Method Across Fourteen Countries. *Journal of Theoretical Educational Science*, *17*(1), 1-27.

Zakariya, Y. F. (2021). Application of an innovative alignment optimisation method to a cross-cultural mean comparison of teacher self-efficacy: A cross-country study. *Heliyon, 7*(10). e08212, https://10.1016/j.heliyon.2021.e08212

Zakariya, Y. F., Bjørkestøl, K., & Nilsen, H. K. (2020). Teacher job satisfaction across 38 countries and economies: An alignment optimization approach to a cross-cultural mean comparison. *International Journal of Educational Research, 101,* 101573.

Ziemes, J. F. (2024). Measuring and explaining political tolerance among adolescents: Insights from the International Civic and Citizenship Education Study 2016. *Large-scale Assessments in Education, 12*(1), 17. https://doi.org/10.1186/s40536-024-00206-x

Ziemes, J. F. (2024). Political trust among European youth: Evaluating multi-dimensionality and cross-national measurement comparability. *Studies in Educational Evaluation, 73,* 101321. https://doi.org/10.1016/j.stueduc.2023.101321

Wu, R., Yang, W., Rifenbark, G., & Wu, Q. (2023). School and Teacher Information, Communication and Technology (ICT) readiness across 57 countries: The alignment optimization method. *Education and information technologies*, *28*(2), 1273-1297.

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.

Appendix 1. *Summary of the studies examined*

1. **Studying the comparability of student perceptions of teaching quality across 38 countries**

This study examined student perceptions of teaching quality in 39 countries using data from 221,016 eighth-grade students from TIMSS 2019. Two key scales were employed: behavioral management (6 items, with response categories ranging from 0 to 3 or 4) and instructional clarity (6 items, with the same response range). To evaluate the underlying factor structure and assess cross-national comparability, a multi-group confirmatory factor analysis (MG-CFA) was conducted. Results indicated that while metric invariance was achieved, scalar invariance was not, thus preventing meaningful cross-country comparisons of mean scores.

For the behavioral management scale, an average of 49.6% of the parameters were identified as non-invariant, while the instructional clarity scale showed an average of 32.9% non-invariant parameters. Both scales had non-invariance levels exceeding the 20% threshold proposed by Asparouhov and Muthén (2022), indicating substantial variability in parameter stability across groups. This degree of non-invariance suggests significant challenges in achieving cross-cultural comparability of these scales.

2. **Assessing the comparability of teacher-related constructs in TIMSS 2015 across 46 education systems: an alignment optimization approach**

In the TIMSS 2019 study, which surveyed 13,508 eighth-grade mathematics teachers, several teacher-related constructs were assessed, including teacher job satisfaction, teacher self-efficacy, school emphasis on academic success, school conditions and resources, and perceptions of a safe and orderly school environment. These constructs were measured using Likert-type scales with response categories ranging from 1 to 4.

Three constructs—school emphasis on academic success, school conditions and resources, and teacher self-efficacy—achieved metric invariance, indicating that the construct measurement was consistent across groups, enabling meaningful comparisons. However, the other constructs did not achieve metric invariance, limiting cross-group interpretability.

Teacher Job Satisfaction (7 items): This construct demonstrated a low percentage of non-invariant groups (8.9%), which is well below the recommended threshold of 25% suggested by Muthén and Asparouhov (2014), indicating minimal variance in how different groups interpret the items.

School Emphasis on Academic Success (5 items): Only 7.8% of groups showed significant non-invariance, suggesting that perceptions of the emphasis on academic success were largely comparable across groups.

School Conditions and Resources (7 items): The percentage of significant non-invariance for this construct was 8.39%, again comfortably below the 25% threshold, indicating a high level of cross-group consistency in item interpretation.

Safe and Orderly School (8 items): This construct had a non-invariance rate of 9.8%, indicating that most groups consistently interpreted the items related to school safety and orderliness.

Teacher Self-Efficacy (9 items): The measure of non-invariance was 8.58%, demonstrating that teachers across groups interpreted self-efficacy items in a largely consistent manner.

3. **Cross-National Measurement of Mathematics Intrinsic Motivation: An Investigation of Measurement Invariance with MG-CFA and the Alignment Method Across Fourteen Countries**

In this study, we assessed the measurement invariance of the TIMSS 2019 mathematics intrinsic motivation scale, which includes nine items rated on a four-point Likert scale, across a sample of 70,192 fourth-grade students and 82,800 eighth-grade students from 14 different countries. The analysis employed Multi-Group Confirmatory Factor Analysis (MG-CFA) and the Alignment Method to evaluate the scale's invariance properties at both grade levels. Results revealed that at the fourth-grade level, the mathematics intrinsic motivation scale demonstrated only configural invariance according to MG-CFA but achieved approximate invariance through the Alignment Method. For the eighth grade, the scale met both configural and metric invariance levels according to MG-CFA, while also achieving approximate invariance with the Alignment Method (%8.7). These findings suggest that the mathematics intrinsic motivation scale supports approximate measurement invariance across countries for both grade levels, allowing for meaningful cross-national score comparisons.

4. **Determinants of trust in order and representative institutions among adolescents across Europe: testing rational and cultural approaches**

In the ICCS 2016 study, conducted across 15 European countries, political trust was measured using a scale comprising six items, each with four response categories. The scale achieved configural invariance

when evaluated through Multi-Group Confirmatory Factor Analysis (MG-CFA), indicating a similar factor structure across countries. Using the Alignment Optimization method, non-invariance was identified in 16% of measurement patterns, suggesting minor variations in how political trust was perceived across these nations.

5. **Is Distributed Leadership Universal? A Cross-Cultural, Comparative Approach across 40 Countries: An Alignment Optimisation Approach**

In the TALIS 2018 study, conducted across 40 countries and involving around 8,000 principals, the distributed leadership scale consisted of 5 questions with 4 response categories. The scale was tested for cross-national comparability and reached metric level invariance, indicating consistent factor loadings across countries. This supports the reliability of examining relationships between distributed leadership and other variables. Additionally, the Alignment Optimization method revealed non-invariance in only 8.75% of the measurement patterns, well within acceptable thresholds, further affirming the validity of cross-country comparisons for this scale.

6. **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**

In this study, data from the TIMSS cycles of 2007, 2011, and 2015 were analyzed across 26 countries, drawing on responses from principals, teachers, and students. The constructs analyzed included "School Emphasis on Academic Success" (from principal and teacher questionnaires), and the "Students Confident in Mathematics" (SCM) and "Students Confident in Science" (SCS) scales from the student questionnaire. These constructs met criteria for comparability over time and across a broad range of countries. For analyses spanning these cycles, a pooled dataset was utilized, with sample sizes varying depending on the specific analysis: between 105,739 and 120,315 for student data, 7665 to 8411 for teacher data, and 4117 to 4389 for principal data.

The "School Emphasis on Academic Success" construct was assessed with seven items on a five-point Likert scale, including topics such as teachers' understanding and implementation of curricular goals, expectations for student achievement, and parental involvement and support. The SCM and SCS scales measured students' confidence in mathematics and science through four consistent items across cycles, capturing aspects of students' perceived proficiency and ease in learning these subjects. Measurement invariance analysis indicated that only configural invariance was consistently achieved across countries, with some constructs reaching metric invariance over time. The alignment optimization approach was applied, establishing approximate invariance across most countries and cycles, though about 25% of models displayed high levels of non-invariance, potentially impacting the reliability of latent mean comparisons. Future analyses may benefit from advanced computing approaches to address persistent non-invariance challenges effectively.

7. **Measurement Invariance Testing with Alignment Method: Many Groups Comparison**

In PISA 2015, the Instrumental Motivation scale was analyzed across 56 countries, involving 406,961 students. The scale comprised four questions, each with four response categories. Results indicated that the scale achieved configural invariance across countries, confirming a consistent factor structure. However, when evaluated through alignment optimization, measurement non-invariance was observed

in 49.5% of the parameters, suggesting limitations in cross-country comparability for latent mean comparisons.

8. **Measuring and explaining political tolerance among adolescents: insights from the International Civic and Citizenship Education Study2016**

The ICCS 2016 study analyzed the political tolerance scale, which consisted of 6 questions rated on a 4-point Likert scale. This assessment was conducted with Grade 8 students across 15 countries. The study utilized the alignment optimization method to test for measurement invariance directly and found that 22.22% of the parameters displayed non-invariance.

9. **Political trust among European youth: Evaluating multi-dimensionality and cross-national measurement comparability**

The ICCS 2016 study, conducted across 15 countries with a sample of 52,788 participants, examined the political trust scale, consisting of 6 questions rated on a 4-point Likert scale. Multi-group confirmatory factor analysis (MG-CFA) indicated that the scale achieved metric level invariance. Using the alignment optimization method, 16.11% of the parameters were identified as non-invariant.

10. **School and Teacher Information, Communication and Technology (ICT) readiness across 57 countries: The alignment optimization method**

This study examined the measurement invariance of school and teacher Information, Communication, and Technology (ICT) readiness across 57 countries participating in the Program for International Student Assessment (PISA) 2018. The School and Teacher ICT Readiness scale is an 11-item self-reported measure with two subfactors: School ICT Readiness and Teacher ICT Readiness.

The data consisted of responses from 18,041 schools, with an average of 316 schools per country. The number of schools varied widely, from 142 in Iceland to 1,089 in Spain. The scale uses a four-point Likert-type format, where school principals rate their agreement with statements ranging from “Strongly disagree” to “Strongly agree.”

The School ICT Readiness subscale comprises five items (SC155Q01HA to SC155Q05HA) focused on infrastructure (e.g., “The number of digital devices connected to the Internet is sufficient”). The Teacher ICT Readiness subscale includes six items (SC155Q06HA to SC155Q11HA) assessing teachers’ technical and pedagogical skills (e.g., “Teachers have the necessary technical and pedagogical skills to integrate digital devices in instruction”).

Rather than conducting a Multi-Group Confirmatory Factor Analysis (MG-CFA), the study directly applied the alignment optimization method. The results showed that the School ICT Readiness subscale exhibited approximately 8% non-invariance, whereas the Teacher ICT Readiness subscale had a higher non-invariance rate of 30.1%.

11. **Tailoring a measurement model of socioeconomic status: Applying the alignment optimization method to 15 years of PISA**

This study analyzed data from the Swedish sample of the PISA cycles from 2000 to 2015, comprising 26,043 students, to investigate the measurement invariance of the Socio-Economic Status (SES) measure. The SES construct was assessed using five items with continuous response options.

Longitudinal analysis was conducted to determine the stability and comparability of the SES measure over time. Using Multi-Group Confirmatory Factor Analysis (MG-CFA), the SES construct achieved metric level invariance. However, when using the Alignment Optimization method, the model exhibited a non-invariance rate of 21.67%, indicating some challenges in ensuring comparability across the different cycles.

12. **Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method**

The aim of this study was to investigate the measurement invariance of mathematics self-concept (MSC) and mathematics self-efficacy (MSE) across 40 countries participating in the Programme for International Student Assessment (PISA) for the years 2003 and 2012. The sample consisted of 271,760 students in the 2003 cycle and 333,804 students in the 2012 cycle. Initially, traditional measurement invariance testing was conducted using Multi-Group Confirmatory Factor Analysis (MGCFA). Subsequently, alignment analysis was performed to allow for minimal non-invariance while estimating all parameters.

Results from MGCFA indicated that mathematics self-concept and self-efficacy achieved metric invariance across the 80 groups (i.e., country-by-cycle combinations). However, the alignment method revealed a high degree of non-invariance in both constructs, making it difficult to compare factor means reliably across all participating countries.

Mathematics Self-Efficacy (MSE): MSE in PISA was measured using eight items that reflect students' perceived abilities to solve specific mathematics problems. Students rated their confidence on a 4-point Likert scale: “very confident,” “confident,” “not very confident,” or “not at all confident.” Example tasks included “using a train timetable,” “calculating TV discounts,” and “calculating square meters of tiles.” The alignment analysis showed that 48.43% of the parameters exhibited non-invariance.

Mathematics Self-Concept (MSC): MSC was assessed using five items that capture students’ perceived competence in mathematics. Responses were collected on a 4-point Likert scale, ranging from “strongly agree” to “strongly disagree.” Example items included statements such as “I get good grades,” “I learn quickly,” and “I am not good at math.” The alignment analysis revealed that 56.75% of the parameters were non-invariant

These findings indicate significant challenges in comparing the factor means of mathematics self-concept and self-efficacy across the countries included in this study, given the high levels of non-invariance.

13. **The evolution of the association between ICT use and reading achievement in 28 countries**

This study analyzed data from PISA cycles 2009, 2012, 2015, and 2018 to investigate three categories of ICT use among students: ICT used for fun, ICT used for learning at home, and ICT used for learning at school. Students were asked to indicate the frequency of their engagement in various ICT-related activities using the following response options: "never or hardly ever," "once or twice a month," "once or twice a week," and "every day or almost every day."

Although the initial analysis involved reporting the percentage of non-invariance using alignment optimization, the researchers did not disclose any specific non-invariance percentages from this method. Subsequently, they employed Monte Carlo simulations to further explore the measurement properties of

the ICT use scales. However, the absence of reported non-invariance percentages leaves a gap in understanding the extent of measurement invariance issues across the cycles and categories of ICT use.

14. **The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality**

Using PISA 2012 data, we investigated the comparability of three teaching quality dimensions: student support (5 items on a 4-point Likert scale), classroom management (5 items on a 4-point Likert scale), and cognitive activation (9 items on a 4-point Likert scale) by employing a novel psychometric method, alignment optimization. Our analysis encompassed 15 countries, grouped into five linguistic clusters, to explore the impact of linguistic similarity on data comparability.

According to Asparouhov and Muthén (2014), a maximum of 25% non-invariant item loadings and intercepts is recommended as the threshold for data comparability. If most item loadings are invariant, it becomes relatively straightforward to adhere to this limit. On the other hand, Flake and McCoach (2018) propose a guideline of 29% non-invariant item intercepts as a benchmark for meaningful mean comparisons. Given that the choice of an upper non-invariance threshold significantly affects interpretation, further research is needed to establish robust psychometric criteria for valid cross-country comparisons.

15. **Measurement Invariance in Comparing Attitudes Toward Immigrants Among Youth Across Europe in 1999 and 2009: The Alignment Method Applied to IEA CIVED and ICCS**

To conduct an integrated alignment analysis across CIVED99 and ICCS09, we created a pooled dataset that included student records from both studies, guided by Barber and Torney-Purta (2012). These authors reviewed and compared the frameworks and items from CIVED99 and ICCS09. Data files from all participating European countries were sourced from the IEA data repository (IEA, 2014).

Our comparative design aimed to minimize bias related to country, cohort, and gender. We incorporated data from 22 European countries in CIVED99 and 24 in ICCS09, yielding a total of 46 country samples. The target population comprised 14-year-old native-born students, with separate groups formed by gender, resulting in 92 groups (46 countries across two cohorts, separated by gender). Each group was assigned a unique identifier combining the country ISO code, the study cohort (99 or 09), and the gender of the students. The sample size totaled 79,278 students, with 18 countries participating in both studies, allowing for trend analysis between 1999 and 2009.

The "Support for Immigrants' Rights (SIR)" scale consisted of five items using a 4-point Likert scale. Alignment analysis indicated 19.3% non-invariance across groups.

16. **Teacher job satisfaction across 38 countries and economies: An alignment optimization approach to a cross-cultural mean comparison**

In this analysis of teacher job satisfaction across 38 countries using TALIS 2018 data, they employed a 4-item 4 options Likert scale. The Multiple-Group Confirmatory Factor Analysis (MG-CFA) achieved metric-level invariance, indicating consistent factor loadings across countries. However, the Alignment Optimization method revealed that 46% of the parameters were non-invariant, suggesting significant differences in item intercepts or thresholds among countries. This discrepancy implies that while the

construct of job satisfaction is similarly understood, the way items are interpreted or responded to varies across cultural contexts, potentially affecting the comparability of mean scores.

17. **Application of an innovative alignment optimization method to a cross-cultural mean comparison of teacher self-efficacy: A cross-country study**

In your analysis of the teacher self-efficacy scale from TALIS 2018 across 39 countries, you utilized a 4-item Likert scale. Applying Multiple-Group Confirmatory Factor Analysis (MG-CFA), you achieved metric-level invariance, indicating consistent factor loadings across these countries. However, the Alignment Optimization method revealed that 44% of the parameters were non-invariant, suggesting significant differences in item intercepts or thresholds among countries. This discrepancy implies that while the construct of teacher self-efficacy is similarly understood across these nations, the interpretation or response to specific items varies, potentially affecting the comparability of mean scores.

18. **Measurement invariance across immigrant and nonimmigrant populations on PISA non-cognitive scales**

In a study analyzing PISA student data from 31 countries, researchers examined two distinct immigrant groups, resulting in 62 groups with at least 300 immigrant students each, totaling 218,315 students. They focused on two scales: 'Sense of Belonging at School' and 'Bullying,' each comprising six items with four response options. Instead of traditional measurement invariance testing, the researchers assessed item-level non-invariance by setting a threshold of 25% for each country. On average, the 'Sense of Belonging at School' scale exhibited 25.1% non-invariance, while the 'Bullying' scale exceeded the 25% threshold, indicating higher levels of non-invariance. These findings suggest that the 'Bullying' scale may be less comparable across countries and immigrant groups, potentially due to cultural differences in interpreting survey items.

Study 2: A Critical Evaluation of Alignment Optimization for Improving Cross-National Comparability in International Large-Scale Assessments

Under Review in (2nd Revision):
Sandoval-Hernandez, A., Carrasco, D, & Eryilmaz, N. (2025). A Critical Evaluation of Alignment Optimization for Improving Cross- National Comparability in International Large-Scale Assessments. *Studies in Educational Evaluation.*

Abstract

This study critically evaluates the application of alignment optimization to enhance the cross-national comparability of teacher and principal scales from the Teaching and Learning International Survey (TALIS) 2018. The investigation addresses key psychometric properties, including dimensionality, reliability, and measurement invariance. Initial analyses confirmed the unidimensionality and high internal consistency of all scales. However, classical multiple-group confirmatory factor analysis (MGCFA) revealed that most scales failed to achieve scalar invariance, thereby limiting their utility for valid cross-national comparisons. By contrast, alignment optimization provided a more flexible yet rigorous framework for comparability assessment, yielding significant improvements for principal scales, all of which met the criteria for cross-national comparability. Despite these advancements, many teacher scales remained below alignment thresholds, signalling unresolved comparability challenges. This study demonstrates the potential of alignment optimisation as a promising approach for improving the comparability of international large-scale assessments, while also identifying areas requiring methodological refinement. The findings underscore the necessity of innovative strategies to enhance the interpretability of global educational measures.

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

Introduction

The importance of ensuring cross-national comparability in International Large-Scale Assessments (ILSAs) cannot be overstated, especially as the number and diversity of participating countries continue to grow (Treviño, et al., 2021). As highlighted in OECD Education Working Papers (2019), the primary objectives of ILSAs shape the extent to which efforts and resources are allocated to improve the cross-national comparability of contextual questionnaires. If the goal is to facilitate valid and meaningful cross-country comparisons, prioritizing comparability becomes essential despite the associated challenges and costs. This is particularly critical as contextual questionnaires are instrumental in providing insights into the diverse educational environments and their influence on student, teacher, and school-level outcomes (Senden et al., 2023).

To address comparability issues, traditional methods such as Multi-Group Confirmatory Factor Analysis (MGCFA) have been extensively used in ILSAs. However, MGCFA often struggles to meet stringent invariance criteria, especially at the scalar level, limiting its effectiveness for cross-group mean comparisons. In response, innovative approaches like alignment optimization have gained traction as alternatives. While alignment methods allow for approximate invariance and greater flexibility, evidence remains inconsistent. Notably, recent studies employing alignment optimization have primarily focused on student data, revealing mixed success in addressing non-invariance challenges. Despite the growing adoption of these methods, there is still no standardized framework or robust consensus on their efficacy across diverse educational settings.

This study contributes to this ongoing discourse by leveraging TALIS 2018 data to examine the application of alignment optimization in the context of teacher and principal questionnaires. By focusing on both teacher and principal scales, we aim to provide clear empirical evidence on the utility of alignment optimization in enhancing cross-national comparability within ILSAs. This analysis not only highlights the potential of alignment methods but also identifies their limitations, offering valuable insights for future methodological advancements in large-scale educational assessments.

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 (Jöreskog, 1971). 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.

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 tests the equality of item intercepts, a prerequisite for valid cross-group comparisons of latent means (Meredith, 1993; Millsap, 2011).

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.

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 1 and Figures X and Y in the Appendix for detailed results).

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.

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 lead to oversimplified results (Fischer et al., 2019; Klieme, 2020).

The alignment optimization approach is an exploratory multiple-group factor analysis technique designed to identify the most suitable measurement invariance configuration. This method facilitates accurate estimation of group means and variances even when complete invariance is not achieved, thereby supporting meaningful cross-group comparisons (Asparouhov & Muthén, 2014). 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).

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 & 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 provides a flexible solution for enabling meaningful cross-group comparisons.

Previous Studies Related to Alignment Optimization (Study 1)

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 allow for meaningful cross-group comparisons, others, such as teacher self-efficacy or mathematics self-concept, show high non-invariance, limiting their comparability. These discrepancies suggest that the success of alignment methods heavily depends on the nature of the construct, item design, and group characteristics.

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.

In the literature with student data, innovative statistical techniques, such as alignment optimization and Bayesian approximate invariance, have been suggested as potential solutions to the challenges posed by exact measurement invariance. 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).

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.

Methodology

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 were weighted using country-specific sampling weights to account for the sampling design and ensure accurate cross-national comparisons.

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 1 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. These measures formed the foundation for the alignment optimization analysis conducted in this study.

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 open-source tool developed by the authors for advanced psychometric analysis with a focus on cross-national comparability.

1. *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 design 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. *Reliability*

The second phase of the analytical strategy focused on evaluating the reliability of the scales. This step involved calculating key reliability metrics, including separation reliability and Cronbach's alpha. Separation reliability assesses the scale's ability to differentiate between groups based on the factor realizations and their standard errors (e.g., Verhavert et al. 2018), while Cronbach's alpha produces a similar estimate which are informative of the total score separation between persons (e.g, Cronbach & Shavelson, 2004). These measures were calculated to ensure the robustness of the scales and their suitability for use in subsequent analyses.

3. *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 the responses and allows for robust testing of measurement properties across groups.

Our analysis adopted a trimming model strategy, starting with the most constrained model (strict invariance) and systematically freeing parameters in subsequent models to test less restrictive invariance levels. The primary focus was on achieving scalar and strict invariance, which are necessary for comparing both latent means and observed means across groups (Tse et al., 2024).

The evaluation criteria for model fit were initially grounded in the recommendations of Rutkowski and Svetina (2017), applying a stringent threshold of RMSEA < 0.055 to determine acceptable model fit. However, as the analysis progressed, the threshold was relaxed to RMSEA < 0.10 to account for the complexity of the data and ensure the feasibility of the models tested. This adaptive approach allowed for a balanced assessment of model fit while maintaining rigor in the evaluation of measurement invariance. This threshold accommodates the large number of groups in the TALIS 2018 dataset and

ensures the robustness of the models tested. Although the chi-square corrected test, as described by Svetina and Rutkowski (2019), was not implemented in this analysis, it remains an option for further refinement of the testing procedure.

This approach provides a rigorous framework for testing measurement invariance, ensuring that comparisons across groups are meaningful and unbiased.

4. *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) for 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.

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.

Results

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.

2. *The results of reliability*

Figure 1 presents the reliability metrics for teacher scales from TALIS 2018, comparing Cronbach's Alpha (orange bars) and Separation Reliability (blue bars). Across the scales, both metrics indicate generally

high reliability, with most scales achieving values above 0.7, a common threshold for acceptable reliability.

Notable findings include slightly lower reliability for scales such as "Exchange and coordination among teachers (subscale)" and "Professional collaboration in lessons among teachers (subscale)," indicating room for improvement in these measures. Conversely, scales like "Team innovativeness" and "Participation among stakeholders, teachers" exhibit strong reliability across both metrics, underscoring their robustness.

The comparison demonstrates consistent results between the two metrics, affirming the reliability of the scales for capturing teacher-related constructs in TALIS 2018.

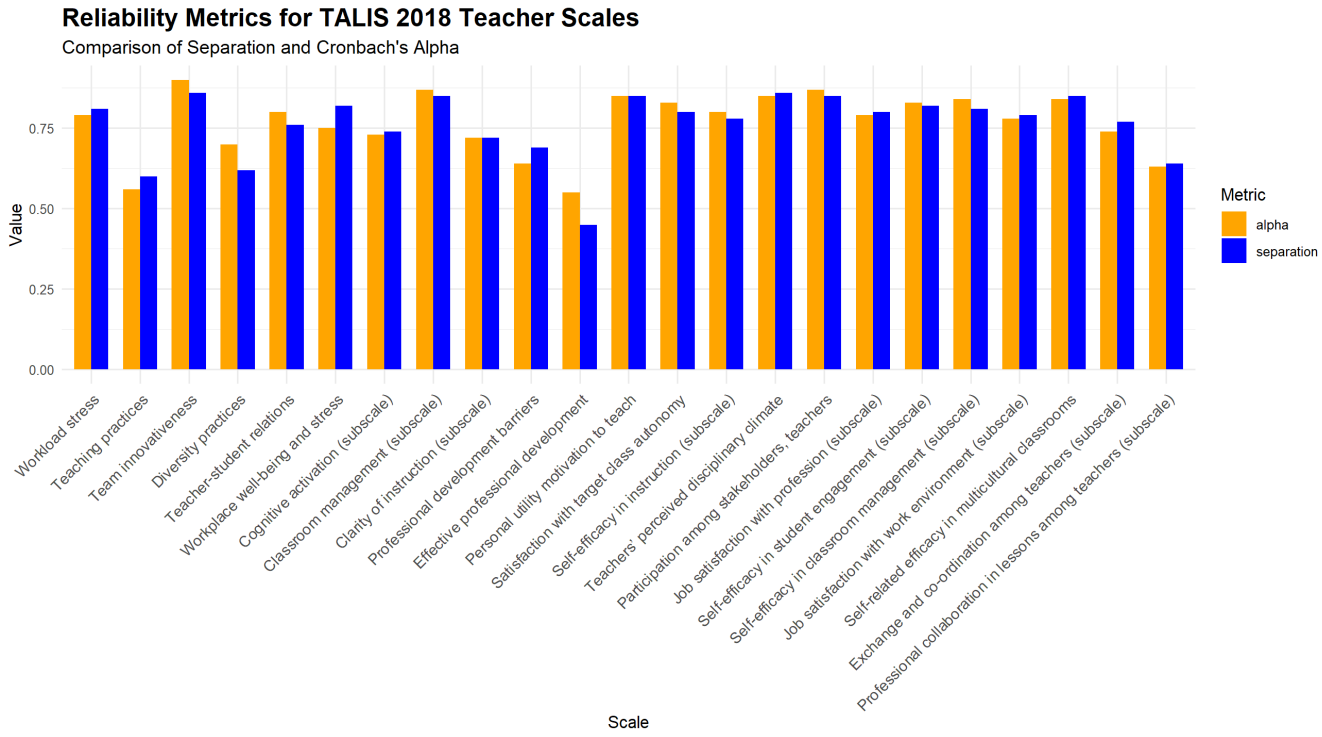

Figure 1. The reliability results for teacher scales

Figure 2 showcases the reliability metrics for principal scales from TALIS 2018, comparing Cronbach's Alpha (orange bars) and Separation Reliability (blue bars). The results indicate consistent reliability across most scales, with values typically above 0.7 for both metrics, suggesting strong internal consistency.

Some scales, such as "Diversity beliefs" and "Organizational innovativeness," exhibit particularly high reliability, with both metrics closely aligned and surpassing 0.8. However, a few scales, such as "Diversity policies, school" and "Diversity practices, school," display relatively lower reliability values, though they still remain within acceptable ranges. The alignment of the two metrics across scales underscores the robustness of these measures for principal-related constructs in TALIS 2018.

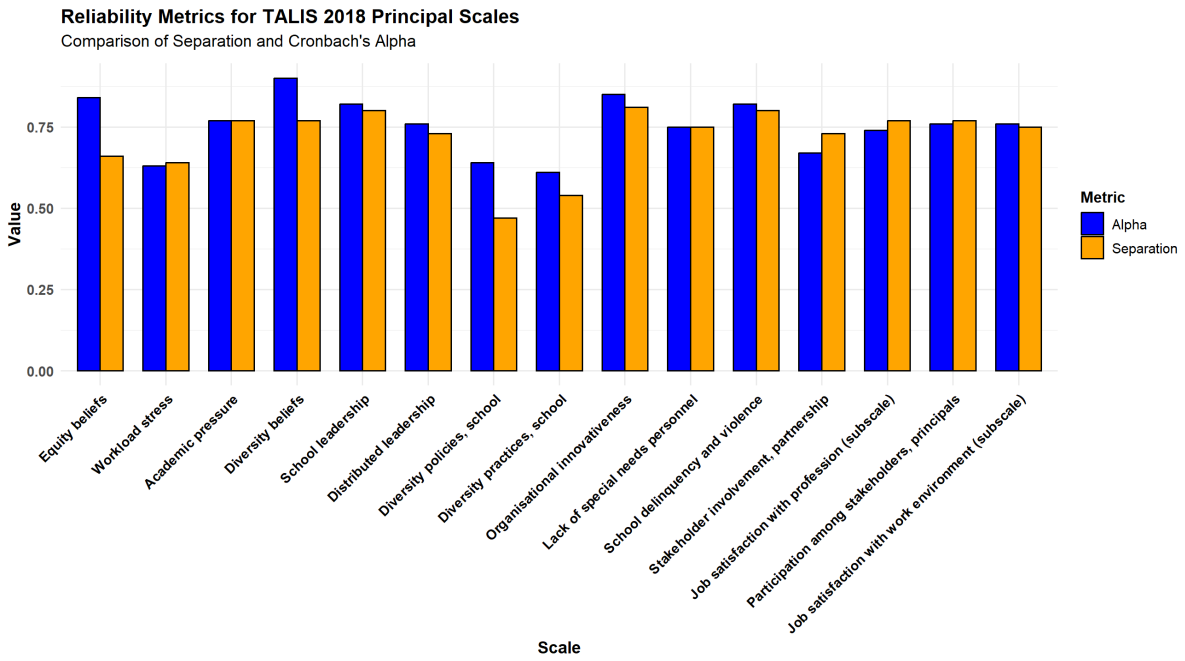

Figure 2. The reliability results for principal scales

3. The results of Measurement Invariance Testing

Figure 3 displays the pooled data fit indices (CFI, RMSEA, and SRMR) for the teacher scales in TALIS 2018, with thresholds set at 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 blue line (CFI) shows that the majority of scales meet or exceed the threshold of 0.90, indicating an acceptable fit for most teacher scales. The green line (SRMR) remains below the threshold of 0.08 across all scales, demonstrating adequate fit. The red line (RMSEA), however, shows variability, with some scales nearing or exceeding the threshold of 0.10, suggesting potential areas where the model fit is less robust. Overall, the results indicate that most teacher scales achieve acceptable model fit, with some requiring further investigation to improve their alignment.

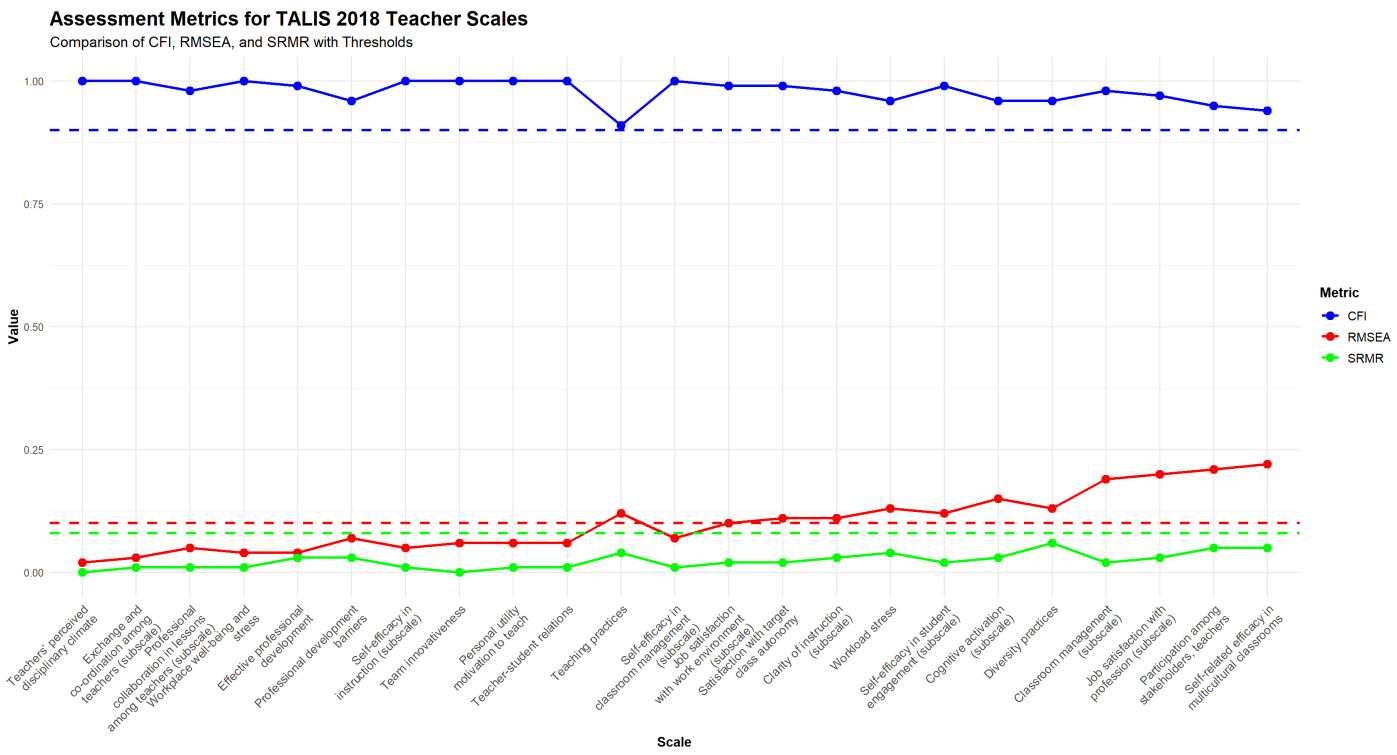

Figure 3. Pooled data Fit Indices for Teacher Scales

Figure 4 presents the pooled fit indices (CFI, RMSEA, and SRMR) for the principal scales in TALIS 2018, with thresholds set at 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 blue line (CFI) demonstrates that all principal scales consistently meet or exceed the threshold of 0.90, indicating good model fit across the board for this metric. The green line (SRMR) remains consistently below the 0.08 threshold for all scales, confirming an acceptable fit with respect to this measure. The red line (RMSEA), however, shows variability, with some scales slightly exceeding the threshold of 0.10. This indicates potential model fit issues for a few scales regarding RMSEA, although the majority remain within acceptable limits.

Overall, the results suggest strong evidence of acceptable model fit for most principal scales, with minimal scales requiring additional evaluation to refine model alignment, particularly when considering RMSEA.

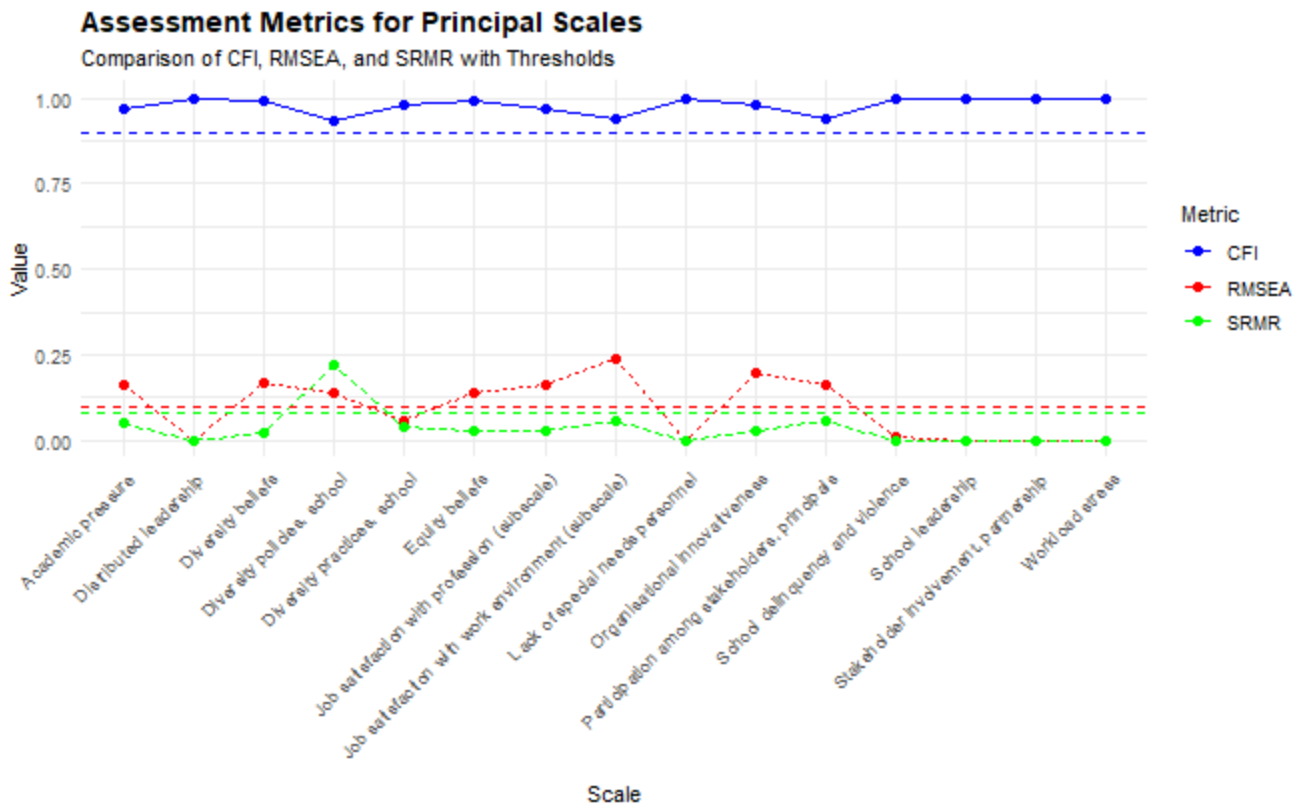

Figure 4. Pooled data Fit Indices for Principal Scales

In the Appendix, Table 2 summarizes the scalar-level invariance results for the teacher scales in TALIS 2018. Scalar invariance is necessary to ensure that mean scores are comparable across countries. Using a stringent RMSEA threshold of 0.055, only a few teacher scales demonstrate comparability across countries.

Specifically, the scales Self-efficacy in classroom management (subscale), Self-efficacy in instruction (subscale), Teacher-student relations, and Team innovativeness satisfy the RMSEA threshold, as indicated by the "yes" result in the RMSEA test. For these scales, mean comparisons across countries can be confidently made. However, the majority of teacher scales either exceed the RMSEA threshold or fail to converge, meaning their mean scores cannot be assumed to be comparable across countries without further refinement.

In the Appendix, Table 3 presents the scalar-level invariance results for the principal scales in TALIS 2018. To ensure valid cross-country comparisons of mean scores, a strict RMSEA threshold of 0.055 was applied.

The results reveal that none of the principal scales meet the threshold for scalar invariance. While some scales, such as Lack of special needs personnel, come close to the threshold, they still fall short of meeting the criteria. Additionally, many scales fail to converge during testing, further limiting their utility for cross-country mean comparisons. These results highlight the need for caution when interpreting mean scores of principal scales across countries.

4. The results of Alignment Optimization

Figure 5 presents the alignment test results for teacher scales in TALIS 2018. Each scale's invariance parameter (inv_par) is plotted, with a dashed vertical line indicating the threshold for acceptable alignment (inv_par = 0.75). Scales with green dots ("yes") meet the alignment threshold, signifying that their latent means are comparable across groups. In contrast, the majority of scales (red dots, "no") fail to reach the threshold, indicating limited cross-country comparability. This highlights challenges in ensuring the comparability of teacher-reported constructs in international contexts. Only one scale successfully passes the alignment test, emphasizing the need for methodological refinement or additional validation.

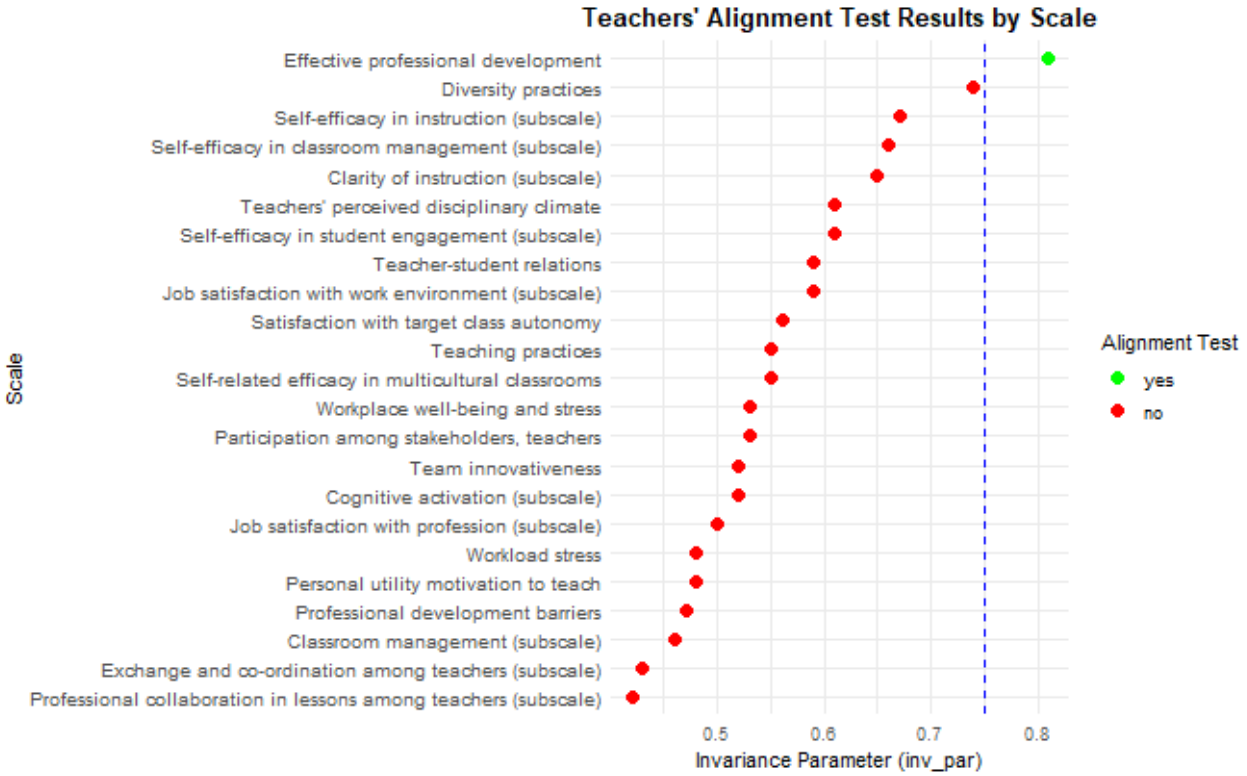

Figure 5. Alignment Results with Teacher Scales

Figure 6 displays the alignment test results for principal scales in TALIS 2018. Each scale is plotted with its corresponding invariance parameter (inv_par), and the dashed vertical line marks the threshold for alignment (inv_par = 0.75). All scales surpass the alignment threshold, as indicated by the green dots ("yes"), demonstrating cross-country comparability of latent mean scores for all principal-related constructs. These results suggest strong consistency and robustness in the measurement of principal-reported scales across different national contexts.

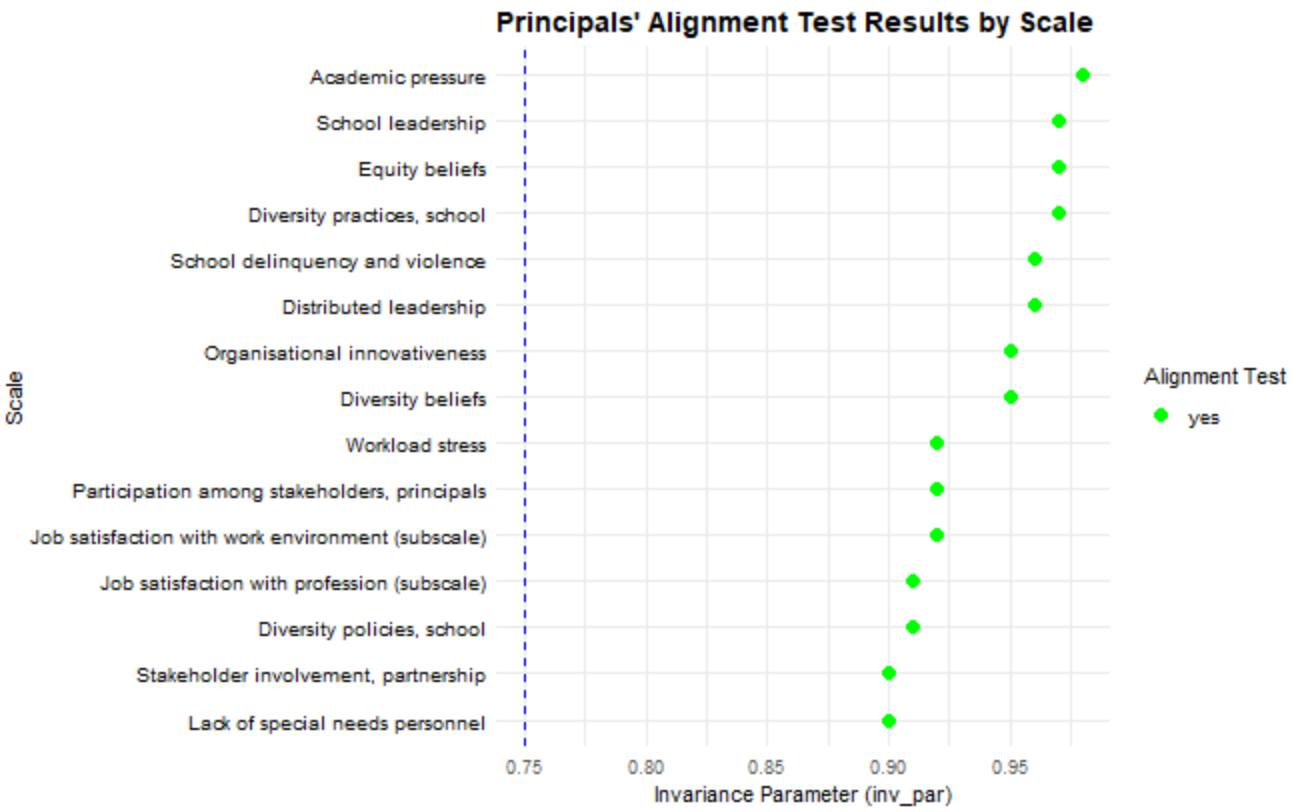

Figure 6. Alignment Results with Principals Scales

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 results highlighted 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 failure to achieve scalar invariance indicates that mean comparisons across countries for most scales are likely to be biased, as differing cultural and contextual interpretations of survey items may distort the constructs being measured.

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. 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.

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 et al., 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.

In conclusion, while alignment optimization proved effective for principal scales, its limited success with teacher scales highlights the ongoing need for methodological refinement and careful interpretation of results. This method is particularly beneficial in identifying the most non-invariant items and the specific groups contributing to non-invariance, both for factor loadings and intercepts. By acknowledging these limitations and leveraging advanced techniques, researchers can better support evidence-based educational policymaking on a global scale.

Future research should prioritise 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.

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

Carrasco, D., Eryilmaz, N., & Sandoval-Hernandez, A. (2024). *rd3c3: R library for advanced psychometric analysis*. Retrieved from 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

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

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.

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 de Educación, 84*(1), 109-133.

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.

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

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.

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 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

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.

Martin, M. O., Von Davier, M., & Mullis, I. V. S. (2020). Methods and Procedures: TIMSS 2019 Technical Report. *TIMSS & PIRLS International Study Center.*

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

Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York, NY: Routledge.

Padgett, R. N. (2023). A tutorial on cross-wave measurement invariance testing with item factor analysis. *Practical Assessment, Research & Evaluation, 28*(13), 1–25. https://openpublishing.library.umass.edu/pare/article/id/1884/

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., & 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

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,* 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 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

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: Manzi, J., García, M.R., Taut, S. (eds) *Validity of Educational Assessments in Chile and Latin America*. Springer, Cham. 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

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

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

Table 1. 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
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

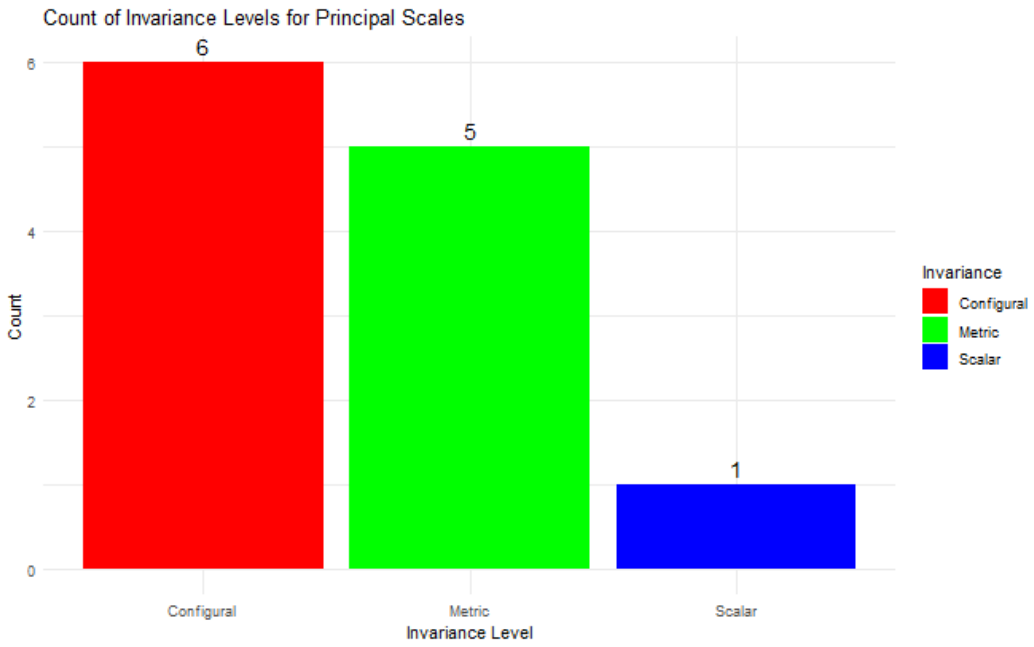

Figure X. Measurement Invariance Level of Principal Scales in TALIS 2018 Report

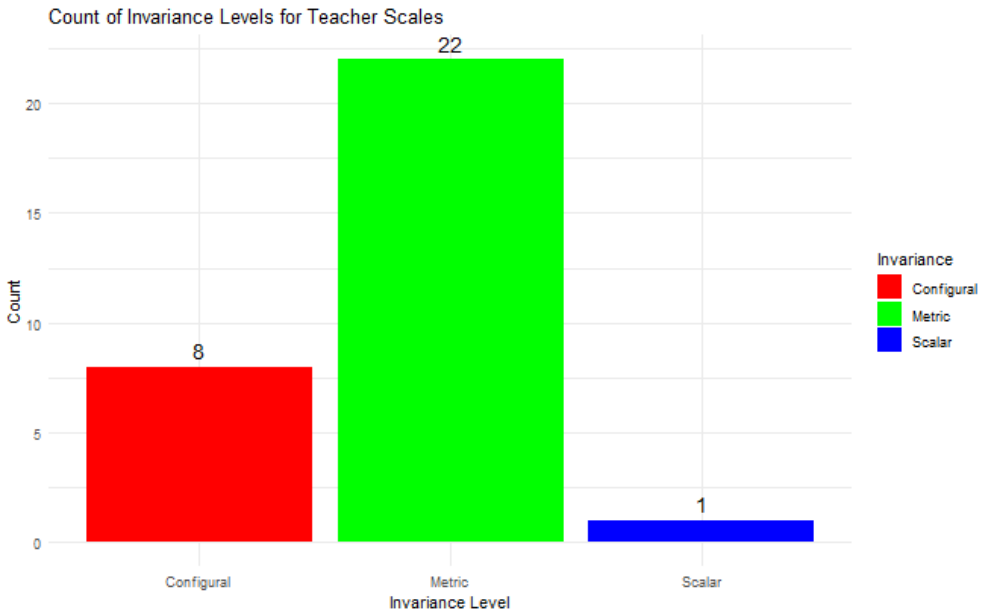

Figure Y. Measurement Invariance Level of Teacher Scales in TALIS 2018 Report

Table 2. 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

Table 3. 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				
Equity beliefs	scalar				

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