METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Multigroup CFA and Alignment Approaches for Testing Measurement Invariance and Factor Score Estimation: Illustration With the Schoolwork-Related Anxiety Survey Across Countries and Gender

Jason C. Immekus[a]

[a] *Department of Educational Leadership, Evaluation, & Organizational Development, University of Louisville, Louisville, KY, USA.*

Methodology, 2021, Vol. 17(1), 22–38, https://doi.org/10.5964/meth.2281

Received: 2019-11-25 • **Accepted:** 2021-02-22 • **Published (VoR):** 2021-03-31

Corresponding Author: Jason C. Immekus, Department of Educational Leadership, Evaluation, & Organizational Development, University of Louisville, 1905 S. 1st St., Louisville, KY 40292, USA. E-mail: jcimme01@louisville.edu

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

Within large-scale international studies, the utility of survey scores to yield meaningful comparative data hinges on the degree to which their item parameters demonstrate measurement invariance (MI) across compared groups (e.g., culture). To-date, methodological challenges have restricted the ability to test the measurement invariance of item parameters of these instruments in the presence of many groups (e.g., countries). This study compares multigroup confirmatory factor analysis (MGCFA) and alignment method to investigate the MI of the schoolwork-related anxiety survey across gender groups within the 35 Organisation for Economic Co-operation and Development (OECD) countries (gender × country) of the Programme for International Student Assessment 2015 study. Subsequently, the predictive validity of MGCFA and alignment-based factor scores for subsequent mathematics achievement are examined. Considerations related to invariance testing of noncognitive instruments with many groups are discussed.

Keywords

multigroup CFA, alignment method, measurement invariance, schoolwork-related anxiety, cross-national measurement

Within large-scale international studies, surveys represent an important approach to the collection of comparative data for understanding individuals' attributes within and

across countries. For instance, the Programme for International Student Achievement (PISA) and the Trends in International Mathematics and Science Study yield cross-national data on students' academic achievement. Across such studies, surveys play a vital role in the operationalization of theoretically meaningful constructs with scores used to test and develop theory, conduct cross-national comparisons, and inform policy. Irrespective of the construct or study context, two key conditions related to score validity are the accuracy of the adaptation of item content and their statistical equivalency across groups (e.g., culture; Greiff & Iliescu, 2017). Whereas test adaptation occurs *a priori* to data collection, evaluation of the measurement invariance of scores occurs post-data collection to judge across group score equivalence. To-date, methodological considerations of traditional measurement invariance (MI) procedures have limited the ability to examine the psychometric properties of obtained scores across many groups (Rutkowski & Svetina, 2014).

MI indicates that an instrument's item parameters are equivalent across groups and, thus, a prerequisite to comparisons based on mean score differences (Vandenberg & Lance, 2000). Scores meet the condition of MI if individuals with similar trait standing have equal likelihood of selecting a particular item response, irrespective of group member (Millsap, 2011). Measurement noninvariance indicates the presence of systematic measurement error that results in the unequal probability of a particular item response across equal ability groups. Despite the importance of MI for test score validity, methodological considerations have restricted its widespread implementation in large-scale international studies (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014).

Multigroup confirmatory factor analysis (MGCFA) is the most commonly used approach to testing item parameter invariance. MGCFA requires the sequential comparison of nested models that differ in terms of the item parameters constrained equal across groups to identify (non)invariant parameters. Typically, comparisons are made across a small number of groups (e.g., ≤ 3) with the aim of establishing at minimum *partial measurement invariance* (Byrne, Shavelson, & Muthén, 1989) to compare latent mean score differences. Two notable challenges associated with this approach include, first, item intercepts rarely demonstrate full (scalar) invariance (Marsh et al., 2018). Second, the number of pairwise comparisons required to identify individual noninvariant parameters becomes overwhelmingly restrictive as the number of groups increases (Rutkowski & Svetina, 2014). Such considerations have resulted in the development of alternative approaches to invariance testing (e.g., Jak, Oort, & Dolan, 2013).

The alignment method (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014) is a procedure to estimate factor means and variances in the presence of noninvariant item parameters. This study builds on alignment research (e.g., Cieciuch, Davidov, & Schmidt, 2018; Lomazzi, 2018; Munck, Barber, & Torney-Purta, 2018), including its comparison with MGCFA (e.g., Coromina & Bartolomé Peral, 2020), by examining the MI of the schoolwork-related anxiety survey administered within the PISA 2015 study across

gender groups within the 35 Organisation for Economic Co-operation and Development (OECD) countries (i.e., country by gender). Test anxiety is a key construct in education and has received much attention in the literature (e.g., Cizek & Burg, 2006), including cross-cultural research (Bodas & Ollendick, 2005). Specifically, test anxiety is a determinant of poor test performance (Chapell et al., 2005; Von der Embse & Hasson, 2012) and school dropout (Cizek & Burg, 2006). Gender comparisons have reported that females have higher test anxiety compared to males (Hembree, 1988; Segool, Carlson, Goforth, Von der Embse, & Barterian, 2013). Within the psychometric literature, test anxiety measures have been found to function similarly across gender groups with females reporting higher latent means (e.g., Everson, Millsap, & Rodriguez, 1991; Lowe, 2015). To-date, there is little empirical evidence regarding the functioning of measures of test anxiety across gender groups within and across countries.

In response, MGCFA and the alignment method are used to test item parameter invariance and group-specific factor mean estimation across gender groups within and across countries, resulting in the analysis of 70 groups (i.e., 35 countries × 2 genders). Consideration of gender within countries serves to examine the comparability of these methods for invariance testing and identify factors that may explain noninvariance of international survey data beyond the country level. Further, this study examines the predictive validity of MGCFA and alignment- based factor scores for students' mathematics achievement.

Multiple Group Confirmatory Factor Analysis

The following multiple-group factor analysis model identifies the measurement parameters of focus in invariance studies

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\eta_{ig} + \epsilon_{ipg} \tag{1}$$

where y_{ipg} is the response of individual i to the observed variable (e.g., item) p ($p = 1, \ldots, P$) in group g ($g = 1, \ldots, G$), ν_{pg} is the intercept for variable p for group g, λ_{pg} is the factor loading of variable p for the latent variable, η_{ig}, for group g, and ϵ_{ipg} is the error variance. It is assumed that $\epsilon_{ipg} \sim N\left(0, \theta_{\text{pg}}\right)$ and $\eta_{ig} \sim N\left(\alpha_g, \psi_g\right)$.

Within MGCFA, invariance testing proceeds through the comparison of nested models that differ according to the equality constraints imposed on model parameters to determine their across group equivalence. An instrument's factor structure can demonstrate three types of invariance: *Configural*, *metric*, and *scalar* (Vandenberg & Lance, 2000). Each level corresponds to Meredith's (1993) categorization of an instrument's factor structure as meeting *weak*, *strong*, or *strict* factorial invariance. Configural invariance is the least restrictive and requires that a common measurement model accounts for the relationship between indicators and factors across groups (Horn & McArdle, 1992). Within this model, item parameters are unconstrained and freely estimated across groups,

with the exception of those required for model identification (e.g., factor variance[s] fixed to 1.0). Subsequently, metric (weak) invariance is met with invariant factor loadings. Subsequently, scalar (strong) invariance requires the additional condition of equal intercepts in addition to factor loadings, and a prerequisite for latent mean comparisons. Strict factorial invariance is the most restrictive level of invariance that also requires the additional condition of equal error variances. Partial measurement invariance occurs when the parameters of two indicators (e.g., items) are equal across groups (Byrne et al., 1989).

MGCFA begins with a baseline model in which to begin the process of identifying invariant item parameters. One approach begins with the configural model, followed by the specification of models with equality constraints imposed on specific parameters to test metric and, subsequently, scalar invariance (see Stark, Chernyshenko, & Drasgow, 2006). Contrary, the baseline model can be a fully invariant measurement model with all item intercepts and factor loadings constrained equal across groups. In this case, modification indices are used to identify individual item parameters to set free across groups to attain a well-fitting model with invariant parameters constrained equal, and noninvariant parameters freely estimated, across groups (Sörbom, 1989). Across approaches, a likelihood ratio (LR) chi-square test, based on the statistical significance of the difference between chi-square values between nested models, is used to identify noninvariant parameters (Stark et al., 2006; Vandenberg & Lance, 2000). Due to the sensitivity of the LR test (e.g., sample size), identification of noninvariant parameters can also be based on the magnitude of the difference between fit indexes (e.g., Root Mean Square Error of Approximation [RMSEA]) of nested models (see Chen, 2007; Rutkowski & Svetina, 2014). Saris, Satorra, and van der Veld (2009) and Van der Veld and Saris (2018) have advanced alternative approaches to evaluate model misspecification to address concerns regarding the use of global fit indices and modification indices in multiple group analyses.

Concerns have been raise regarding the impracticality of MGCFA to testing MI with many groups (Muthén & Asparouhov, 2014; Rutkowski & Svetina, 2014). Specifically, as the number of compared groups and/or scale items increases, so does the number of pairwise comparisons needed to identify noninvariant parameters. For instance, if metric invariance is not met, individually testing the invariance of each factor loading across the 35 OECD countries would require 595 pairwise comparisons ($35 \times 34/2$). Thus, invariance testing becomes increasingly challenging and overwhelming as the number of groups and/or indicators increases. Furthermore, in cross-national research in which factor means may be of most interest, attaining the prerequisite condition of scalar invariance is rarely met (e.g., Davidov et al., 2014; Rutkowski & Svetina, 2014). Additionally, in consideration of the reliance of modification indices and subsequent pairwise comparisons required to identify noninvariant parameters, Marsh et al. (2018) emphasize critiques raised regarding the use of statistically-driven, stepwise procedures for hypothesis testing that may increase the likelihood of chance findings. Taken together, MGCFA

may be a less than optimal procedure for invariance testing when many groups are included in the comparison

Alignment Method

The alignment method is an approach to estimate group-specific factor means and variances in the presence of measurement noninvariance (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014). The model assumes is that it is possible to reduce the amount and size of noninvariant parameters. This is done through a two-step procedure that estimates group-factor means α_g and variances ψ_g in which the amount of MI among item parameters is reduced (Asparouhov & Muthén, 2014).

Similar to MGCFA, the alignment method begins with the specification of a statistically well-fitting configural model, referred to model M0 (Asparouhov & Muthén, 2014). The model represents the factor structure with the best overall model-data fit across groups with item intercepts and factor loadings freely estimated, and factor means (α_g) and variances $\left(\psi_g\right)$ fixed to 0 and 1 for identification purposes.

Alignment optimization is the second step and focuses on the estimation of the group-specific factor means and variances that reduce the amount of measurement noninvariance of the item parameters across groups. Specifically, α_g and ψ_g are those that minimize the total loss function F

$$F = \sum_{p} \sum_{g1 < g2} w_{g1,g2} f\left(\lambda_{pg1} - \lambda_{pg2}\right) + \sum_{p} \sum_{g1 < g2} w_{g1,g2} f\left(\nu_{pg1} - \nu_{pg2}\right) \tag{2}$$

where p is the number of items, g_1 and g_2 are group 1 and group 2 for each pair of compared groups, λ_{pg1} and λ_{pg2} are the factor loadings for variable p for each of the two groups, and ν_{pg1} and ν_{pg2} are the intercepts for variable p across the groups. The difference between the two group loadings and intercepts is scaled via the component loss function,

$$f(x) = \sqrt{x^2 + 0.01} \tag{3}$$

(Jennrich, 2006). Based on the component loss function, F is minimized when a few item parameters report large noninvariance and most are approximately invariant (Asparouhov & Muthén, 2014). Thus, alignment optimization provides the basis for the selection of group-specific factor means and variances that yield intercept and factor loadings that are approximately invariant, based on a model that reports the same model-data fit as the configural model.

The corresponding weight factor, $w_{g1,g2}$, is

$$w_{g1,g2} = \sqrt{N_{g1} N_{g2}} \tag{4}$$

where N_{g1} and N_{g2} are the sample sizes for Group 1 and Group 2, respectively. Consequently, groups comprised of larger sample sizes will contribute more to the total loss function.

There are two approaches to alignment optimization. The first is the *fixed* which constrains a pre-selected group mean and variance to 1 and 0. Contrary, within the *free* optimization, group means are freely estimated with one group variance set to 1. Asparouhov and Muthén (2014) present the alignment method within maximum likelihood (ML) or Bayesian estimation and identify advantages of each approach to invariance testing. Notably, applied studies using free optimization have reported poor model identification (e.g., Byrne & Van de Vijver, 2017; Cieciuch et al., 2018; Marsh et al., 2018) and, thus, have used the resultant *Mplus* output to select the group with the factor mean closest to 0 for the referent group in the subsequent analysis using the fixed approach. Alignment analyses with noncognitive surveys have used the fixed approach with ML estimation (e.g., Marsh et al., 2018; Munck et al., 2018).

Beyond estimating group-specific factor means, the procedure (as conducted in *Mplus*) implements an automatic ad hoc procedure that provides specific information regarding the noninvariance of each item parameter. Specifically, a list of the groups for which a particular item parameter is (non)invariant is reported. This is based on a multistep procedure in which each group's intercept and factor loading is compared to the parameter's average value, based on an *invariant set* (Asparouhov & Muthén, 2014). The invariant set is a sub-set of groups from all groups in the analysis with a parameter value that is statistically equivalent. The contribution of each item to simplicity function F is also reported, with smaller values indicative of higher invariance. Correspondingly, an R^2 reported for each item parameter reports the amount of across group parameter variation in the configural (M0) model accounted for by the variation in the factor means and variances across groups, with values ranging between 0 and 1 (values closer to 1 indicative of higher invariance). Collectively, the information can serve as a guide to subsequent decisions related to item functioning and development.

Muthén and Asparouhov (2014) offer general recommendations to assist researchers with inspecting the trustworthiness of results. For example, 25% or less of the item parameters should report noninvariance, which has been questioned because it does not consider the degree and location of noninvariance (Kim, Cao, Wang, & Nguyen, 2017). Also, a Monte Carlo study using the real data estimated item parameters as start values in the simulation offers a way to examine parameter recovery accuracy and compare factor mean ordering based on the real and simulated data (correlation of 0.98 or higher suggested).

There are several attractive features of the alignment method for invariance testing. First, the procedure is automatic and, thus, reduces the tediousness associated with inspecting modification indices to conduct numerous statistically driven hypothesis tests. Second, model-data fit after alignment optimization is the same as the configural model.

Third, it is applicable to measurement instruments comprised of a small number of items and data collected in a complex sampling framework. Furthermore, the method does not require that the data be normally distributed, an uncommon attribute of item-level data. In addition to information regarding the degree of invariance of item parameters, group-specific factor means and variances are obtained with a factor structure that is not meet scalar invariance, and applicable to group sizes ranging from 2 to 100 (Asparouhov & Muthén, 2014).

Recent methodological and applied studies shed light on the contributions of the alignment method for invariance testing of large-scale international surveys. Specifically, simulation studies have supported the method's recovery accuracy of item parameters and factor means under various conditions (e.g., Asparouhov & Muthén, 2014; Flake & McCoach, 2018; Muthén & Asparouhov, 2018; Pokropek, Davidov, & Schmidt, 2019) and detection rates (Kim et al., 2017). For example, Flake and McCoach (2018) found that the method performed well with parameter recovery even in conditions with high levels of noninvariance. For latent mean estimation, Pokropek et al.'s (2019) simulation study reports that partial MI models are able to recover latent means in the presence of many (known) noninvariant items, approximate MI models may be work well when parameters may be roughly equivalent, and the alignment method may serve applicable with a few noninvariant items. A growing number of applied studies demonstrate its potential to advance our understanding of the measurement properties of scales used within cross-cultural research (e.g., Byrne & Van de Vijver, 2017; Marsh et al., 2018; Munck et al., 2018). Currently, there is a literature gap with applied studies comparing MGCFA and alignment approaches for invariance testing of large-scale international surveys in education.

Study Purpose

This study extends the literature on the use of MGCFA and alignment optimization in large-scale, international educational survey research through its application to the schoolwork-related anxiety measure administered within PISA 2015 for country by gender, with 70 compared groups. For alignment optimization, a Monte Carlo study is used to examine the accuracy of factor means. The present study builds on the existing literature (e.g., Coromina & Bartolomé Peral, 2020; Munck et al., 2018) by comparing MGCFA and alignment optimization procedures for invariance testing and, correspondingly, the predictive validity of factor scores for mathematics achievement.

Method

Participants

Data were based on nationally representative samples of 15-year-old students (N = 237,241) enrolled in educational institutions across the 35 OECD countries in the PISA 2015 study. Student selection was based on a two-stage stratified sampling design in which schools within countries were selected first and, subsequently, students were sampled within schools (OECD, 2017a). Additional PISA study information is available on its website (http://www.oecd.org/pisa/). Within this study, groups were created by combining country and gender for a total of 70 groups (i.e., 35 countries × 2 gender groups).

Instrumentation

The schoolwork-related anxiety scale includes five selected-response items designed to operationalize "[T]he anxiety related to school tasks and tests, along with the pressure to get higher marks and the concern about receiving poor grades" (OECD, 2017b, p. 84). Item content is, Item 1: "I often worry that it will be difficult for me taking a test;" Item 2: "I worry that I will get poor <grades> at school;" Item 3: "Even if I am well-prepared for a test I feel very anxious;" Item 4: "I get very tense when I study for a test;" and, Item 5: "I get nervous when I don't know how to solve a task at school." Responses are reported on a four-point scale (1 = *Strongly agree*; 2 = *Disagree*; 3 = *Agree*; 4 = *Strongly agree*). The median reliability estimate based on coefficient omega across countries was 0.94 (Range: 0.78 [Japan] - 0.99 [Germany]).

Data Analysis

As a first step, MGCFA was used to determine the level of invariance (e.g., metric) among the item parameters. Criteria for metric and scalar invariance included nonsignificant chi-square difference statistic ($p > .05$) and ΔRMSEA ≤ .03 for metric (≤ .01 for scalar) and ΔCFI (Comparative Fit Index) ≤ |0.02| for metric (≤ |.01| for scalar), as per Rutkowski and Svetina (2014).

For the alignment analysis, the free option was selected with parameter estimation based on the robust maximum likelihood (MLR) estimator using *Mplus* (Version 8.3; Muthén & Muthén, 1998-2017), reported to yield accurate parameter estimates for indicators with at least 4 categories (Beauducel & Herzberg, 2006; DiStefano, 2002). In the event of poor model identification, the group with the factor mean closest to zero was selected as the referent group using the fixed approach. PISA study design features were incorporated into the analysis by accounting for stratum and student weights. Missing data was handled using the full information maximum likelihood procedure implemented in *Mplus*. Model-data fit of the configural model was based on the following fit indexes:

RMSEA (Steiger, 1990), CFI (Bentler, 1990), with RMSEA values < 0.05 and CFI values > 0.95 indicative of acceptable fit (Hu & Bentler, 1999). Subsequently, Monte Carlo simulation was conducted to investigate the trustworthiness of alignment optimization results with sample sizes of 250, 500, 1,000, 2,000, and 3,000, based on 100 replications (Asparouhov & Muthén, 2014).

Subsequently, the predictive validity of factor scores for mathematics achievement was examined using a random intercepts two-level multilevel model (MLM). As PISA reports 10 plausible values (PVs) of mathematics performance, separate analyses were conducted for each PV in which regression coefficients and standard errors were averaged across analyses.

Results

A single-factor model reported acceptable model-data fit (see Table 1). As expected, LR chi-square difference tests were statistically significant (*ps* < .01) across compared models (e.g., metric vs. configural), whereas ΔRMSEA and ΔCFI were 0.016 and 0.021 between the metric and scalar invariance models, respectively. Modifications indices indicated Items 2 and 5 were most invariant across groups (negligible changes across groups) and, thus, were constrained equal across groups to obtain a model that demonstrated partial MI, $\chi^2(764) = 5{,}795.14$, RMSEA = 0.43 (90% CI [0.042, 0.044]), CFI = 0.98, SRMR = 0.033.

Table 1

Model-Data Fit of Multiple Group Measurement Invariance Testing of Schoolwork Related Anxiety Factor Structure

Model	Number of free parameters	χ^2	*df*	RMSEA	RMSEA 90% CI	CFI	SRMR	ΔRMSEA	ΔCFI	ΔSRMR
Configural	1,050	1,888.58*	350	0.035	0.034-0.037	0.994	0.011			
Metric	774	2,755.57*	626	0.031	0.030-0.032	0.992	0.023	0.004	0.002	0.012
Scalar	498	8,099.41*	902	0.047	0.046-0.048	0.971	0.039	0.016	-0.021	0.016
Partial MI	636	5,610.34	764	0.042	0.041-0.043	0.981	0.032	0.011	-0.011	0.009

Note. MI = Measurement invariance; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual.
*$p < .01$.

Alignment optimization using the free approach produced an error message that the model may be poorly identified. Correspondingly, the Finland-Females group had the factor mean closest to 0, which was selected as the referent group in the subsequent analysis with the fixed approach. Table 2 reports key alignment optimization results. Column 1 reports the fit function for each item intercept and factor loading, as well as the total for each item parameter. Among the factor loadings, Item 2 contributed the most amount of noninvariance, followed by Item 4. Among intercepts, Item 4 reported

the highest degree of invariance, whereas Item 2 had the most amount of noninvariance. Comparatively, the overall fit function for the factor loadings was lower than that for the intercepts, indicating a higher degree of invariance among the loadings compared to the intercepts. The overall fit function of the individual items (obtained by summing the fit functions of factor loadings and intercepts) indicated that Item 2 (fit function = -2,454.19) was the least invariant item and Item 3 (fit function = 1,975.28) was the most invariant (see Table S1 in Supplementary Materials for complete alignment results).

Table 2

Alignment Analysis Results of the Measurement Invariance of the Motivation to Achieve PISA 2015 Items

Item parameter	Fit Function Contribution	R^2	Number of groups with approximate MI	*M*	*SD*	Min		Max	
						Estimate	Group	Estimate	Group
Loading									
ST118Q01	-918.74	0.42	47	1.00	0.09	0.77	Mexico (F)	1.28	Finland (M)
ST118Q02	-1,032.56	0.28	49	1.00	0.13	0.57	Spain (F)	1.33	Finland (F)
ST118Q03	-867.17	0.61	46	1.00	0.06	0.85	Finland (F)	1.19	Spain (M)
ST118Q04	-988.81	0.14	42	1.00	0.11	0.72	Japan (M)	1.21	Iceland (F)
ST118Q05	-922.82	0.51	58	1.00	0.08	0.77	Japan (M)	1.13	Chile (M)
Sum	-4,730.10								
Intercept									
ST118Q01	-1,186.58	0.73	16	-0.34	0.20	-0.83	Greece (F)	0.04	Portugal (F)
ST118Q02	-1,421.63	0.72	18	-0.33	0.29	-1.02	Greece (F)	0.44	Spain (F)
ST118Q03	-1,108.11	0.90	15	-0.35	0.15	-0.68	Turkey (F)	0.04	Finland (M)
ST118Q04	-1,112.41	0.83	29	-0.38	0.19	-1.11	Austria (F)	-0.03	Norway (F)
ST118Q05	-1,259.18	0.85	14	-0.36	0.21	-0.70	Japan (F)	0.24	Greece (F)
Sum	-6,087.91								
Total	-10,661.82								

Note. MI = Measurement invariance; (M) = male; (F) = female.

Consequently, the parameters of each item demonstrated some degree of noninvariance, with no item parameter demonstrating approximate MI across groups. Specifically, the number of countries and gender groups with approximately MI among factor loadings ranged from 42 (Item 4) to 58 (Item 5), whereas among intercepts the number ranged from 42 (Item 4) to 58 (Item 8). Column 2 reports R^2 for each item, which indicates the degree of parameter variation across the groups within the configural model (M0) that is explained by across group factor mean and variance variation (values closer to 1.00 indicative of higher invariance). Overall, 30.86% and 73.71% of the factor loadings and intercepts were noninvariant, which resulted in 52.29% of the parameters being non-invariant, exceeding Muthén and Asparouhov's (2014) recommendation of 25%. Notably, Muthén and Asparouhov (2014) report the percent of noninvariant item parameters is one indicator of the trustworthiness of the results. As reported in column 4, the average

factor loading across groups was 1.00, whereas intercepts ranged from -0.33 (Item 1) to -0.38 (Item 5).

Columns 8 and 9 report the groups with the lowest and highest parameter values for each item. Specifically, Item 1 reported the weakest relationship to the schoolwork-related anxiety factor for Mexico-Females (loading = 0.77), whereas its relationship was the strongest for Finland-Males (loading = 1.28). For Item 3, the weakest relationship was for Finland-Females and the strongest for Spain-Males. Similarly, Items 4 and 5 reported the weakest relation to the anxiety factor among Japan-Males, whereas Items 4 and 5 were most strongly related to the factor for Iceland-Females and Chile-Males, respectively. Among intercepts, Greece-Females had the lowest reported levels of schoolwork-related anxiety for both Items 1 and 2, whereas the highest anxiety levels for these items were among Portugal-Females and Spain-Females. Interestingly, whereas Greece-Females reported the lowest anxiety for Items 1 and 2, they reported the highest intercept for Item 5 that assessed anxiety related to getting nervous when not knowing how to solve a task at school.

Alignment-based factor means indicated that Portugal-Females (M = 1.42), Italy – Females (M = 1.38), and Spain Females (M = 1.191) had the highest levels of schoolwork-related anxiety. Groups with the lowest factor means included Finland-Males (M = -0.446), Switzerland-Males (M = -0.586), and Netherlands-Males (M = -0.676). The correlation between the alignment and partial invariance model factor means was 0.97 (see Table S2 for rank-ordering of alignment-based factor means and Table S3 for comparison of rank-order of alignment and MGCFA factor means in Supplementary Materials). For the partial invariance model, Portugal-Females (M = 0.542), Japan-Females (M = 0.517), and Turkey-Females (M = 0.445) reported the highest factor means, whereas the lowest scores were for Belgium-Males (M = -0.36), Germany-Females (M = -0.556), and Germany-Males (M = -1.083). The partial invariance model and alignment procedures identified 45.71% and 48.57% of the groups in the upper half of the distribution comprised of female groups. Figure S1 in the Supplementary Materials shows the rank ordering across procedures in which increased dispersion is observed seen among the groups with the lowest factor means (e.g., Germany-Males).

Based on the Monte Carlo simulations, correlations between the alignment optimization population and estimated factor means were 0.99 across conditions[1] , except for the group size of 250 which was 0.98, thus meeting the criteria of 0.98 (Muthén & Asparouhov, 2014). Therefore, despite the presence of large measurement noninvariance of item parameters, the factor means were well- recovered.

For the predictive validity of scores (intraclass correlation coefficient = 0.37), across the 10 mathematics achievement PVs, the average regression coefficient for the align-

1) In each Monte Carlo simulation, the number of instances of nonconvergence for specific groups in each sample size condition was: 1 for 500 and 2,000; 3 for 250 and 3,000; and 5 for 1,000.

ment-based factor score was -12.46 (*SE* = 0.15; minimum = -12.89 and maximum = -12.27) and −35.09 (*SE* = 0.43; minimum = -35.39 and maximum = -0.44) for factor scores based on partial MI.

Discussion

Only recently have methodological and applied studies emerged on the utility of the alignment method to invariance testing, including its comparison to alternative approaches with many groups. Within this literature, there is limited application of the method in large-scale, international educational studies, and less illustrations of its use to invariance testing of country by gender comparisons with more than 70 groups. Further, there is limited research regarding the relationship of partial invariance and alignment-based scores to external variables (e.g., Pokropek et al., 2019), a key source of validity evidence.

Within PISA, cross-country scale comparability included strict translation procedures, whereas construct validity was based on within country reliability, and, for item parameter invariance, the root mean square deviance item-fit statistic (RMSD; i.e., difference between model-based and observed item characteristic curves) for country-by-language combinations (OECD, 2017a). Notably, the OECD 2015 Technical Reports indicate validity studies are ongoing and suggests research to compare their approach to invariance testing to other psychometric methods to determine the comparability of cross-country scores (see Note 4, p. 343).

Aligned with previous research, item factor loadings demonstrated higher amounts of invariance compared to intercepts. While the amount of measurement noninvariance exceeded Muthén and Asparouhov's (2014) recommendation (< 25%), this is perhaps not surprising considering the methodological restrictions of pursuing such analyses across multiple groups. Marsh et al. (2018) note that this criteria may be too simplistic, sample-dependent, and not generalizable and, thus, suggest research to determine an "index" to assist with decision-making. Therefore, empirical results provide relevant information for continued scale development. Specifically, Item 3 was the most invariant item and dealt with students feeling anxious for a test even when well-prepared. Contrary, alignment optimization identified Item 2 as the least invariant and addressed students' worry about poor school grades. Notably, among factor loadings, noninvariance was observed more across countries than within country gender groups, indicating that the relationships between the items and latent trait varied according to cultural differences, not gender. Contrary, there was less of a trend for item intercepts. Correspondingly, although MGCFA did not support complete scalar invariance, Items 2 and 5 reported the least amount of noninvariance and were constrained equal across all groups to obtain a partial invariance model for factor mean estimation. Notably, Pokropek et al. (2019) report that MGCFA is able to recover factor means reasonably well in the presence of

varying degrees of approximate MI for 5-item scales. Substantively, these results may offer evidence to guide investigations into linguistic and contextual factors that may influence individuals' responses to self-report surveys.

Within international studies, survey scores provide an important source of information to identify factors that may be associated with intended study outcomes. Whereas Monte Carlo simulation results indicated the trustworthiness of estimate factor means under varying sample sizes, a subsequent question of this study was their predictive validity. Compared to partial invariance scores, alignment-based scores reported a lower relationship to mathematics achievement. Along these lines, Pokropek et al. (2019) report that regression coefficients can be well-recovered in the presence of varying degrees of noninvariance using alternative procedures. This is an emerging area of research with additional investigations warranted to examine the degree levels of noninvariance may influence item parameter and factor mean estimation, including their relationships with external variables. Nonetheless, study findings further substantiate the negative relationship between test anxiety and students' academic achievement (e.g., Von der Embse & Hasson, 2012). As research on the utility of alignment optimization in international studies increases, the extent measurement noninvariance influences the relationship between factor analytic scores to external variables is an area of future research.

Whereas the alignment procedure has largely been used as an exploratory method, Marsh et al. (2018) present its use as a confirmatory-based approach which expands its use in applied research. Notably, the alignment method is one recently proposed approach to invariance testing in the presence of many groups and Muthén and Asparouhov (2018) compare and contrast the alignment method to a multilevel approach, which treats groups as random instead of fixed. As a new approach to invariance testing, there are unique areas of research to further understanding its methodological contributions to scale development and validation, as well as how it compares to other methods (e.g., Coromina, & Bartolomé Peral, 2020; Pokropek et al., 2019) and using different estimation procedures (e.g., Bayesian; Muthén & Asparouhov, 2018).

Funding: The author has no funding to report.

Acknowledgments: The author has no support to report.

Competing Interests: The author has declared that no competing interests exist.

Supplementary Materials

For this article the following Supplementary Materials are available via the PsychArchives repository (for access see Index of Supplementary Materials below):

- Table S1 for approximate measurement invariance of item parameters (intercepts, factor loadings).
- Table S2 for alignment-based factor mean rank-ordering.
- Table S3 for alignment and MGCFA factor mean rank-ordering.
- Figure S1 shows the rank ordering of alignment and MGCFA factor means.

Index of Supplementary Materials

Immekus, J. C. (2021). *Supplementary materials to: Multigroup CFA and alignment approaches for testing measurement invariance and factor score estimation: Illustration with the schoolwork-related anxiety survey across countries and gender* [Tables S1-S3 and Figure S1]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.4719

References

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

Bodas, J., & Ollendick, T. H. (2005). Test anxiety: A cross-cultural perspective. *Clinical Child and Family Psychology Review, 8*, 65-88. https://doi.org/10.1007/s10567-005-2342-x

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and measn structure: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M., & Van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29*(4), 539-551. https://doi.org/10.7334/psicothema2017.178

Chapell, M. S., Blanding, Z. B., Silerstein, M. E., Takahashi, M. N. B., Newman, B., Gubi, A., & McCain, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274. https://doi.org/10.1037/0022-0663.97.2.268

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

Cieciuch, J., Davidov, E., & Schmidt, P. (2018). Alignment optimization: Estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and applications* (2nd ed., pp. 571-592). New York, NY, USA: Routledge.

Cizek, G., & Burg, S. S. (2006). *Addressing test anxiety in a high-stakes environment.* Thousand Oaks, CA, USA: Sage.

Coromina, L., & Bartolomé Peral, E. (2020). Comparing alignment and multiple group CFA for analysing political trust in Europe during the crisis. *Methodology, 16*(1), 21-40. https://doi.org/10.5964/meth.2791

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327-346. https://doi.org/10.1207/S15328007SEM0903_2

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51*(1), 243-251. https://doi.org/10.1177/0013164491511024

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polyotomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling, 25*(1), 56-70. https://doi.org/10.1080/10705511.2017.1374187

Greiff, S., & Iliescu, D. (2017). A test is much more than just the test itself: Some thoughts on adaptation and equivalence. *European Journal of Psychological Assessment, 33*(3), 145-148. https://doi.org/10.1027/1015-5759/a000428

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77. https://doi.org/10.3102/00346543058001047

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144. https://doi.org/10.1080/03610739208253916

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

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling, 20*(2), 265-282. https://doi.org/10.1080/10705511.2013.769392

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika, 71*, 173-191. https://doi.org/10.1007/s11336-003-1136-B

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

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, data, analyses, 12*(1), 77-104. https://doi.org/10.12758/mda.2017.09

Lowe, P. A. (2015). Should test anxiety be measured differently for males and females? Examination of measurement bias across gender on measures of test anxiety for middle and high school, and

college students. *Journal of Psychoeducational Assessment, 33*(3), 238-246. https://doi.org/10.1177/0734282914549428

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

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, USA: Routledge.

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research, 47*(4), 687-728. https://doi.org/10.1177/0049124117729691

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, Article 978. https://doi.org/10.3389/fpsyg.2014.00978

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

Muthén, L. K., & Muthén, B. O. (1998-2017). *MPLUS user's guide* (8th ed.) Los Angeles, CA, USA: Muthén & Muthén.

OECD. (2017a). *PISA 2015 Technical report*. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

OECD (2017b). *PISA 2015 results (Volume III): Students' well-being*. Paris, France: PISA, OECD Publishing. https://doi.org/10.1787/9789264273856-en

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling, 26*(5), 724-744. https://doi.org/10.1080/10705511.2018.1561293

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561-582. https://doi.org/10.1080/10705510903203433

Segool, N. K., Carlson, J. S., Goforth, A. N., Von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*(5), 489-499. https://doi.org/10.1002/pits.21689

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384. https://doi.org/10.1007/BF02294623

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology, 91*(6), 1292-1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Van der Veld, W. M., & Saris, W. E. (2018). Measurement equivalence testing 2.0. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and application* (pp. 245-282). New York, NY, USA: Routledge.

Von der Embse, N., & Hasson, R. (2012). Test anxiety and high-stakes test performance between school settings: Implications for Educators. *Preventing School Failure, 56*(3), 180-187. https://doi.org/10.1080/1045988X.2011.633285