Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 29-53

Measurement invariance testing in questionnaires: A comparison of three Multigroup-CFA and IRT-based approaches

Janine Buchholz[1] *& Johannes Hartig*[2]

Abstract

International Large-Scale Assessments aim at comparisons of countries with respect to latent constructs such as attitudes, values and beliefs. Measurement invariance (MI) needs to hold in order for such comparisons to be valid. Several statistical approaches to test for MI have been proposed: While Multigroup Confirmatory Factor Analysis (MGCFA) is particularly popular, a newer, IRT-based approach was introduced for non-cognitive constructs in PISA 2015, thus raising the question of consistency between these approaches. A total of three approaches (MGCFA for ordinal and continuous data, multi-group IRT) were applied to simulated data containing different types and extents of MI violations, and to the empirical non-cognitive PISA 2015 data. Analyses are based on indices of the *magnitude* (i.e., parameter-specific modification indices resulting from MGCFA and group-specific item fit statistics resulting from the IRT approach) and *direction* of local misfit (i.e., standardized parameter change and mean deviation, respectively). Results indicate that all measures were sensitive to (some) MI violations and more consistent in identifying group differences in item difficulty parameters.

Key words: item response theory, item fit, confirmatory factor analysis, modification indices, PISA

[1] *Correspondence concerning this article should be addressed to:* Janine Buchholz, PhD, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt, Germany; email: buchholz@dipf.de

[2] DIPF | Leibniz Institute for Research and Information in Education, Germany

Background

Many international large-scale assessments (ILSAs) such as the Programme for International Student Assessment (PISA), the Programme for International Assessment of Adult Competencies (PIAAC), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS), aim at measuring latent constructs such as competencies, attitudes, values and beliefs in participating countries, and at comparing the resulting scale scores or relationships between them across these countries. In order for such comparisons to be valid, measurement invariance (MI) pertaining to the underlying measurement model needs to hold. A number of different statistical procedures to test for MI have been proposed (for an overview, see Vandenberg, & Lance, 2000; for more recent developments, see Van De Schoot, Schmidt, De Beuckelaer, Lek & Zondervan-Zwijnenburg, 2015), the most popular of which being the multigroup confirmatory factor analysis (Jöreskog, 1971) approach (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Greiff & Scherer, 2018). In its most recent cycle, the Programme for International Student Assessment (PISA 2015) applied another, rather new approach for testing the invariance of IRT-scaled constructs (OECD, 2017). A total of three statistical approaches resulting from these two general frameworks are presented in detail below, and their comparison forms the subject of this paper.

Many, if not all of the ILSAs consist of both a cognitive assessment and a questionnaire, the latter providing auxiliary information regarding the contexts of teaching and learning such as attitudes toward learning, home resources, pedagogical practices, and school resources, to name a few (Kuger, Klieme, Jude, & Kaplan, 2016). While a lot of (media) attention is placed on findings regarding the cognitive assessment alone, context questionnaires are equally important as they contribute to the achievement estimation and allow for the "contextualization" of student performances in participating countries (Rutkowski & Rutkowski, 2010). In fact, a recent literature review on the nature of PISA-related publications demonstrated that the majority of secondary research focused on constructs administered with questionnaires (Hopfenbeck, Lenkeit, El Masri, Cantrell, Ryan, & Baird, 2018). Despite this prominence, the authors reported that while many studies investigated measurement invariance of the cognitive part of the assessment, only few studies did for the PISA questionnaires. This imbalance in favor of cognitive constructs is not restricted to PISA but generalizes to other ILSAs such as TIMSS (Braeken & Blömeke, 2016). According to another recent literature review, only very few studies in cross-cultural psychology focusing on comparisons between cultures tested for MI (e.g., only 16.8% in the *Journal of Cross-Cultural Psychology* (JCCP): Boer, Hanke & He, 2018). This study will therefore focus on MI testing related to constructs administered with questionnaires.

Multigroup confirmatory factor analysis (MGCFA)

Multigroup confirmatory factor analysis (MGCFA; Jöreskog, 1971) presents the traditional and by far most commonly employed approach to MI testing (e.g., Boer et al., 2018; Cieciuch et al., 2014; Greiff & Scherer, 2018). It is based on confirmatory factor

analysis (Jöreskog, 1969) in which a set of observed indicator items X_i is predicted by a latent person variable ξ_j. However, instead of one set of parameters describing the linear relationship between X_i and ξ, the MGCFA model consists of one set of parameters per group, and the equivalence of these parameters between groups is tested in a series of logically ordered and increasingly restrictive models (Byrne, 2012). The model can be represented by

$$X_{ijg} = \lambda_{ig}\xi_j + \alpha_{ig} + \varepsilon_{ijg} \tag{1}$$

in which the response of person *j* in group *g* on item *i*, X_{ijg}, is predicted by the loading (or "slope", λ_{ig}) and intercept (α_{ig}) parameters of item *i* in group *g*, the person's level on the latent construct, ξ_j, as well as an error term ε_{ijg} with $\varepsilon_{ijg} \sim N\left(0, \sigma^2_{\varepsilon_{ig}}\right)$.

Typically, the analysis begins with a model in which the configuration (i.e., the set of items serving as indicators for the latent construct) is specified to be the same across groups, yet all parameters are freely estimated, followed by a model in which all slope parameters (λ_i) are constrained to be equal across groups, followed by a model in which both slope (λ_i) and intercept (α_i) parameters are constrained to be equal across groups. These models represent different levels of measurement invariance, i.e., configural, metric (or "weak", cf. Meredith, 1993), and scalar (or "strong", cf. Meredith, 1993), respectively, and have implications for the interpretation of factor scores, ξ_j. In the presence of metric invariance, associations between factor scores and other variables can be compared across groups, while in the presence of scalar invariance, factor scores themselves may also be compared, in addition to associations among them. The decision on the level of measurement invariance (configural, metric, scalar) is typically based on the degree of change in global model fit between two subsequent models, thus indicating whether the introduction of the respective equality constraints and, thus, the equality of parameters can be assumed. Several rules of thumb for acceptable change in model fit have been proposed for various model fit indices such as CFI, TLI, and RMSEA (e.g., Chen, 2007).

The MGCFA method, however, has proven to be impractical and unreliable in the presence of many groups (Rutkowski & Svetina, 2014), making it unfeasible for operational use in the context of ILSAs in which many groups are rather common. For example, 72 countries participated in PISA 2015, however, the most extreme simulation condition implemented in Rutkowski and Svetina (2014) contained a total of only 20 groups. Even though the authors suggested more lenient criteria in the presence of 20 groups, it can be expected that these are still too conservative for ILSAs such as PISA.

Modification indices. One way to get around the global model fit information is to make use of parameter-specific modification indices (MoIs), thus providing information on *local* model fit. For each parameter constraint, MoIs indicate how much the global model fit would improve in terms of likelihood (transformed into chi-square distributed quantities) if this particular parameter was freely estimated or if this parameter's constraint (e.g., equality between groups) was released (Sörbom, 1989). Modification indices may be used to respecify the model at hand by introducing additional parameters, but this changes the nature of the procedure from being confirmatory to being exploratory. It also

needs to be noted that a model's MoIs are not independent from each other as the release of one parameter can change the fit of the model as a whole. Consequently, the sequence of releasing model constraints might change the final model resulting from such a procedure (e.g., MacCallum, Roznowski, & Necowitz, 1992). In this study, MoIs will be subject to analyses, not to guide model modification but because they provide a valuable indication of local model fit.

Expected parameter change. The Expected Parameter Change (EPC) statistic has been suggested as an alternative way to evaluate model misspecifications. The EPC indicates the estimated change in a restricted model parameter if it were freely estimated, thus providing "a direct estimate of the size of the misspecification for the restricted parameters" (Saris, Satorra, & Sörbom, 1987, p. 120). First introduced by Saris et al. (1987), Chou and Bentler (1993) provided a fully standardized version of the statistic that we will refer to as "SEPC" in the following. According to Whittaker (2012) who conducted a cited reference search in the Social Sciences Citation Index, the majority of empirical studies investigating measurement invariance using the MGCFA approach based their analyses on both MoIs and the EPC or SEPC statistic.

It needs to be noted that the model presented above (eq. 1) assumes the variables to be continuous and follow a normal distribution. However, items used in questionnaires of large-scale assessments typically use ordered categorical, Likert-type items, thus violating this assumption. For PISA 2015, for example, the median number of response categories of the items used for scaling the non-cognitive constructs was four (OECD, 2017; also, see Annex A). Measurement models for ordered categorical data have been extended to the multiple-group case (e.g., Muthén & Asparouhov, 2002), but were found to be less widely discussed than the single group case (Millsap, 2011, p. 126) and are hardly seen in practice. However, we will also include findings resulting from ordinal MGCFAs.

IRT item fit

A rather new approach to investigating measurement invariance is based on item fit of an IRT model (for applications, see Oliveri & von Davier, 2011, 2014; Pokropek, Borgonovi, & McCormick, 2017). While it has been implemented operationally in an ILSA before (PIAAC, see Yamamoto, Khorramdel & von Davier, 2013), PISA 2015 was the first ILSA to use the procedure for both the cognitive assessment *and* the context questionnaires. It was applied to all 58 scales based on response data from questionnaires administered to students, parents, school principals and teachers (OECD, 2017).

In an initial step, a concurrent calibration is conducted in which all item parameters are constrained to be equal across all equally weighted ("senate weighted", cf. Gonzalez, 2012, p. 121) groups (i.e., languages within countries). The calibration is based on the generalized partial credit model (GPCM; Muraki, 1992) which takes the form

$$P\left(X_{ijg}=x\,|\,\theta_j,\alpha_{ig},\beta_{iug}\right)=\frac{\exp\left(\sum_{u=0}^{x}\alpha_{ig}\left(\theta_j-\beta_{iug}\right)\right)}{\sum_{r=0}^{m}\exp\left(\sum_{u=0}^{r}\alpha_{ig}\left(\theta_j-\beta_{iug}\right)\right)} \tag{2}$$

with

$$\sum_{u=0}^{0}\alpha_{ig}\left(\theta_j-\beta_{iug}\right)\equiv 0 \tag{3}$$

in which $P\left(X_{ijg}\right)$ is the probability of person j in group g responding in category x of item i, θ_j is the latent trait of person j, α_{ig} represents the discrimination of item i in group g, and β_{iug} represents the threshold u of item i in group g. Based on this model with equal item parameters across all groups g, a group-specific item-fit statistic (root-mean-square deviance; $RMSD_g$) was calculated and served as a measure for the invariance of item parameters for individual groups.

RMSD. For an item i with $k=0, 1,\ldots K$ response categories, $RMSD_g$ for group g is defined as

$$RMSD_g=\sqrt{\frac{1}{K+1}\sum_{k=0}^{K}\left(P_{obs,gk}\left(\theta\right)-P_{exp,gk}\left(\theta\right)\right)^2 f\left(\theta\right)d\theta}, \tag{4}$$

quantifying the difference between the observed item characteristic curve based on pseudo counts from the E-step of the EM algorithm (ICC, $P_{obs,gk}\left(\theta\right)$) with the model-based ICC ($P_{exp,gk}\left(\theta\right)$; OECD, 2017; Khorramdel, Shin, & von Davier; 2019; M. von Davier, personal communication, November 8, 2019). Good item fit – i.e. an RMSD close to zero – indicates that a group's data can be described well by the joint item parameters, thus pointing at the presence of MI. Bad item fit, in contrast, indicates that data cannot be described well by the joint item parameters, thus pointing at a possible violation of MI. It needs to be noted that RMSD is an overall measure of fit assessing both cross-country comparability *and* the overall goodness-of-fit of that item, so that the presence of bad item fit is not necessarily indicative of a violation of MI (Pokropek et al., 2017). In PISA 2015, equality constraints were released and group-specific item parameters assigned whenever RMSD was above a certain cutoff-value, thus resembling the concept of partial MI known in the context of MGCFA (Byrne, Shavelson, & Muthén, 1989). The efficiency of this procedure in achieving good model fit while maintaining comparable scales has been demonstrated (Oliveri & von Davier, 2011, 2014).

Mean deviation. The mean deviation (MD) provides an alternative, yet related measure of item fit. Just as the RMSD, it quantifies the difference between an item's observed and model-based ICC, but as it is based on the weighted sum of these differences, it quantifies both the magnitude *and* direction of these differences. For the polytomous case, MD is defined as

$$MD_g=\frac{1}{K+1}\sum_{k=0}^{K}\left(P_{obs,gk}\left(\theta\right)-P_{exp,gk}\left(\theta\right)\right)f\left(\theta\right)d\theta \tag{5}$$

According to Khorramdel et al. (2019), "the MD is most sensitive to the deviations of observed item difficulty parameters from the estimated ICC, [while] the RMSD is sensitive to the deviations of both the item difficulty parameters and item slope parameters" (p. 622).

As RMSD and MD have been introduced only recently, they have not been studied extensively. To our knowledge, of the two only the RMSD is subject to current research, both in the one-group and multiple-group case (e.g., Köhler, Robitzsch, and Hartig, in press; Author). For the one-group scenario, Köhler and colleagues (in press) demonstrated that the RMSD depends on sample size and the number of indicator items: it decreases with increasing sample size (due to the reduction of the finite-sample bias) and it increases for higher numbers of indicator items (due to a decrease of accuracy in item difficulty estimation which leads to a larger difference between the expected and the observed IRF and, thus, an increase in the RMSD). Knowing the statistic's null distribution, however, is particularly important as RMSD provides descriptive information rather than a formal test, thus depending heavily on a cutoff criterion that can be well-justified. While Oliveri and von Davier (2011, 2014) recommended a value of RMSD equal to .1 to serve as a cutoff criterion, values of .12 and .3 were used for the cognitive and non-cognitive constructs in PISA 2015, respectively (OECD, 2017). A recent sensitivity study, however, indicated that a value of .3 for polytomous items (typical for questionnaires) might have still been too lenient (Author). For the purpose of this study, we do not rely on the choice of a specific cutoff criterion for the RMSD.

The RMSD can be regarded as a measure of the magnitude of local misfit (similar to MoIs in MGCFA), indicating how well the joint parameters of a specific item (discrimination, threshold) fit the data of a specific group, while MD can be regarded as a measure for the direction of local misfit (similar to SEPC in MGCFA).

Aim of the study

Two general frameworks for MI testing (MGCFA-based, IRT-based) have been introduced above. While one of the two was operationally used in a prominent ILSA and promises practicality in the ILSA context, the other one is considered the method of choice in the research community. While all three approaches provide quantitative measures of the magnitude (RMSD, MoIs) and direction (MD, SEPC) of model misfit, they also differ in four major aspects: (1) the underlying measurement model (IRT vs. CFA), (2) the nature of the analysis (descriptive vs. formal test), (3) the level of analysis (item fit vs. model fit – or, when using MoIs – item fit vs. parameter fit), and (4) the assumptions about the nature of the indicators (ordered categorical vs. continuous). Little is known about the relationship between the approaches: Are their findings consistent or would researchers come to different conclusions about the presence of MI, depending on the statistical approach they were using? The present study therefore aims at investigating the consistency between the approaches in quantifying the degree of MI for a given dataset.

Particular emphasis will be placed on questionnaire data that are typically polytomous in nature. This is to address the observation according to which MI testing in the context of ILSAs is almost exclusively focused on the cognitive parts of the assessment only in which dichotomous items are more typical.

We will first conduct a Monte-Carlo simulation study in which we vary the pattern and extent of the underlying non-invariance and apply the three methods (MGCFA- and GPCM-based, respectively) to these data. In the subsequent empirical application, we will investigate the consistency between the two approaches based on the published PISA 2015 questionnaire data.

Method

To compare the performance of the two approaches in identifying violations of MI (or "non-MI"), we conducted a Monte-Carlo simulation study in which true data follow different patterns of non-MI, and the two approaches' ability in quantifying these is investigated.

Data generation. Response data for five four-category items were generated for a total of N=50,000 simulees across 50 groups with 1,000 simulees each. These responses are based on a normal ogive graded response model of the form

$$\pi_{ixj} = \Phi\left(\tau_{ix} - \lambda_i\theta_j\right) - \Phi\left(\tau_{ix-1} - \lambda_i\theta_j\right) \tag{6}$$

in which π_{ixj} represents the probability of person j responding in category x ($x \in \{1,2,3,4\}$) of item i, θ_j represents the ability of person j, λ_i represents the loading (also "slope" or "discrimination") of item i, and τ_{ix} represents the threshold for category x on item i with $\tau_{i0} = -\infty$ and $\tau_{i4} = +\infty$.

In the basic simulation setup, parameters for all five items i were set to be identical across groups with $\boldsymbol{\tau_i} = (-1,0,1)$ and $\lambda_i = 1$. For the implementation of the different patterns of violations of MI, see the next section ("Simulation design"). Values for the person parameter θ were drawn from a group-specific normal distribution, and groups (g) were allowed to vary in both means and standard deviation:

$$\theta_{jg} \sim N\left(\mu_g, \sigma_g\right) \tag{7}$$

Because of

$$\mu_g \sim N(0, 0.5) \tag{8}$$

$$\sigma_g \sim U(.8, 1.2) \tag{9}$$

the variation within groups was larger than the variation between groups. Distribution parameters μ_g and σ_g as well as values for θ were sampled within replications. Each condition was replicated 1,000 times.

Simulation design. To implement different patterns of non-MI, we shifted the item parameters λ_i and $\boldsymbol{\tau_i}$ for the last item in each group (item 5) by adding a group-specific constant, indicated by L_g and T_g, respectively, depending on the simulation condition:

$$\lambda_{5g} = \lambda_i + L_g \tag{10}$$

$$\boldsymbol{\tau}_{5\boldsymbol{g}} = \boldsymbol{\tau_i} + T_g \tag{11}$$

We manipulated both type (between-replications) and extent (within-replications) of non-MI. Regarding *type*, we implemented three simulation conditions: (1) a baseline condition with equal item parameters across all groups and items, i.e., a condition in which MI holds, (2) a condition in which groups differ with respect to the slope parameter of item 5, λ_5, and (3) a condition in which groups differ with respect to the threshold parameters of item 5, $\boldsymbol{\tau}_5$. In addition, we manipulated the *extent* of non-MI within replications by varying the shift of a group's parameter from the average parameter. Table 1 provides the exact specification of these "shift parameters" (L_g, T_g), thus illustrating the simulation design. Note that they were not sampled from the intervals provided in Table 1, but instead were chosen to be equidistant and unique across the 50 groups, thus taking on as many values as there are groups. With 50 groups and an interval width of 2, ranging between -1 and 1, L_g for the first three groups are $L_1 = -1$, $L_2 = -0.959$, and $L_3 = -0.918$, respectively, and $L_{50} = 1$ for the last group. With 50 groups and an interval width of 4, ranging between -2 and 2, T_g becomes $T_1 = -2$, $T_2 = -1.918$, $T_3 = -1.837$, …, and $T_{50} = 2$, respectively. As a result, group 1 in condition 2 was assigned the item parameters $\lambda_{51} = \lambda_i + L_1 = 1 + (-1) = 0$ and $\boldsymbol{\tau}_{51} = \boldsymbol{\tau_i} + T_1 = (-1,0,1) + 0 = (-1,0,1)$. Group 1 in condition 3, in contrast, received the parameters $\lambda_{51} = \lambda_i + L_1 = 1 + 0 = 1$ and $\boldsymbol{\tau}_{51} = \boldsymbol{\tau_i} + T_1 = (-1,0,1) + (-2) = (-3,-2,-1)$.

The deliberate choice of item parameters allows for a systematic inspection of findings for each level of a violation of MI.

Table 1:
Simulation design with three simulation conditions representing different types of violations of measurement invariance.

	Condition		
	1	2	3
L_g	0	$[-1:1]$	0
T_g	0	0	$[-2:2]$

Analysis. The 3 (between-conditions) * 1,000 (replications) = 3,000 datasets were each analyzed under (a) the data-generating ordinal MGCFA model, (b) the MGCFA model assuming normality, and (c) the multiple-group GPCM model.

For (a), we estimated a MGCFA model for ordinal data in Mplus (version 8; Muthén & Muthén, 1998-2017) with equal item parameters (slopes, thresholds) across all groups using the weighted least square mean and variance adjusted estimator (WLSMV) with the theta parameterization. With equal slope and threshold parameters across groups, this model corresponds to a scalar level of MI. The modification index (MoI) and standardized expected parameter change (SEPC, "StdYX E.P.C." in the Mplus output) statistic for each parameter (slope, threshold), item, group, condition, and replication were extracted. For each item's three threshold parameters, the average MoI and average SEPC were computed and form the basis for analysis.

For (b), we estimated a MGCFA model for continuous data in Mplus (version 8; Muthén & Muthén, 1998-2017) with equal item parameters (slopes, intercepts) across all groups using a maximum likelihood estimator with robust standard errors (MLR). In this model, the items are treated as continuous variables, an assumption actually violated given the categorical data generation (for consequences of such a violation, see Li, 2016). However, we deliberately chose to include this analysis as we believe the estimation with models for continuous variables is more common in empirical applications (e.g. Rutkowski & Svetina, 2014). With equal slope and intercept parameters across groups, this model corresponds to a scalar level of MI. The modification index (MoI) and expected parameter change (SEPC) statistic for each parameter (slope, intercept), item, group, condition, and replication were extracted.

For (c), we followed the operational procedure in PISA 2015 as documented in the Technical Report (OECD, 2017). As such, the data were analyzed under the GPCM (see eq. 2) with equal item parameters (discriminations, thresholds) across groups using mdltm (version 1.965; von Davier, 2005; Khorramdel et al., 2019). RMSD and MD values for each item, group, condition, and replication were extracted.

Results

Table 2 contains the aggregated findings on convergence and global model fit across replications for the three simulation conditions and the three approaches each. With one exception, all analyses converged without any problems; problems with convergence occurred only for the ordinal MGCFA model when groups differed with respect to their thresholds. For the majority of these cases (5.2 out of 6.8%), this was due to a lack of observations in each response category for every group which is a necessary condition for this model. As expected, the baseline condition (condition 1) consistently shows best fit compared with the other two conditions. Among the remaining two conditions, better absolute and relative fit indices are observed when groups differ in slopes (condition 2), and worst fit occurs when groups differ in thresholds (condition 3).

In the following, all results are based on item 5, the item for which groups differed in either their slope or threshold parameter, depending on the simulation condition. Table 3 contains descriptive statistics on indices regarding the magnitude of local misfit resulting from each of the three approaches, aggregated across replications for each of the three

Table 2:
Global model fit across replications by simulation condition.

	Condition		
	1 (baseline)	2 (slope)	3 (thresholds)
MGCFA (ordinal)			
Successful replications (%)	100	100	93.2
CFI	1.000 (0.000)	0.938 (0.004)	0.668 (0.020)
RMSEA	0.003 (0.003)	0.090 (0.002)	0.206 (0.005)
χ^2 (*df*=887)	889 (43)	8141 (336)	38454 (1956)
MGCFA (continuous)			
Successful replications (%)	100	100	100
CFI	1.000 (0.000)	0.921 (0.005)	0.447 (0.024)
RMSEA	0.003 (0.004)	0.081 (0.002)	0.215 (0.003)
χ^2 (*df*=642)	643 (37)	4840 (231)	30233 (699)
GPCM			
Successful replications (%)	100	100	100
AIC	606191 (4729)	611487 (4393)	615238 (4900)
BIC	607231 (4729)	612527 (4393)	616279 (4900)
Log likelihood	-302977 (2365)	-305625 (2197)	-307501 (2450)

Note. First value represents Mean, value in brackets *SD*. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds.

Table 3:
Indices regarding the magnitude of local misfit on item 5 resulting from the different approaches (MGCFA- and GPCM-based) across replications by simulation condition.

	Condition		
	1 (baseline)	2 (slope)	3 (thresholds)
MGCFA (ordinal)			
MoI (slope)	1.399 (2.069)	128.202 (146.339)	921.303 (1679.896)
MoI (threshold)	1.001 (0.908)	44.560 (49.276)	557.796 (818.863)
MGCFA (continuous)			
MoI (slope)	0.803 (1.132)	72.651 (88.312)	117.864 (156.947)
MoI (intercept)	0.989 (1.389)	20.198 (38.681)	376.301 (251.118)
GPCM			
RMSD	0.035 (0.010)	0.114 (0.056)	0.254 (0.102)

Note. First value represents Mean, value in brackets *SD*. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds. MoI: modification index.

conditions. Just as on the global level, best local fit occurred in the baseline condition (condition 1) while worst fit was observed when groups differ in item thresholds (condition 3). Under the ordinal MGCFA model, MoIs are consistently higher for the slope parameter regardless the simulation condition. Under the assumption of continuous indicators, however, the findings relate to the two models' parameters in the expected direction: When groups differed with respect to item slopes (condition 2), higher values (larger misfit) occurred for slope-related MoIs than for intercept-related MoIs. In contrast, when groups differed with respect to item thresholds (condition 3), higher values (larger misfit) occurred for the intercept-related MoIs than for slope-related MoIs. In general, the largest misfit is observed in condition 3 for all indices on the magnitude of local misfit.

In addition to the type of non-MI, the extent of non-MI was manipulated within replications. To see the impact of a group's deviation from the true parameter, Figures 1 and 2 show the indices on the magnitude (MoI, RMSD) and direction of local misfit (SEPC, MD), respectively, conditional on group membership and, thus, conditional on the extent of the MI violation. Under all approaches, indices for the magnitude of local misfit (MoI, RMSD) are generally highest in the condition with group differences in thresholds (condition 3) and lowest in the condition where MI holds (condition 1). Also under all approaches, the two conditions with MI violations (conditions 2 and 3) show a U-shaped pattern, indicating larger misfit for groups with high absolute deviations from the average parameter as indicated by the shift parameters, L_g and T_g. For the ordinal MGCFA model, MoIs relating to the slope parameter are consistently highest, regardless the simulation condition. Under the assumption of continuous indicators, the parameter-specific MoIs reflect the type of MI violation much better: when groups differ in slopes, the slope-related MoIs exceed the intercept-related MoIs (condition 2), and when groups differ in thresholds, the intercept-related MoIs exceed the slope-related MoIs (condition 3). However, for both MGCFA-based approaches, the two parameter-related MoIs appear to be not completely independent from each other either. Under both types of non-MI, not only did the MoI relating to the parameter in question react but also the other MoI. When comparing the two MGCFA-based approaches with the GPCM, a difference in the symmetry of the pattern resulting under condition 2 becomes apparent: While the RMSD is sensitive to both negative and positive deviations of a group's slope from the average slope parameter, the MoIs resulting from both MGCFA-based approaches seem to only pick up negative deviations, i.e., low slopes.

Regarding the indices for the direction of local misfit (Figure 2), the patterns are very similar with those described above: indices are highest when groups differ in their thresholds (condition 3) and lowest when MI holds (condition 1). For the ordinal MGCFA model, the pattern is only in the expected direction when groups differ in slopes, and indistinct when groups differ in thresholds. The pattern, again, is much clearer for the continuous MGCFA model: the slope-related SEPC varies as a function of the shift parameter L_g when groups differ in slopes (condition 2), and the intercept-related SEPC varies as a function of T_g when groups differ in thresholds (condition 3). Under the GPCM, MD is always in the expected direction, but much more pronounced under the implementation of non-MI in condition 3.

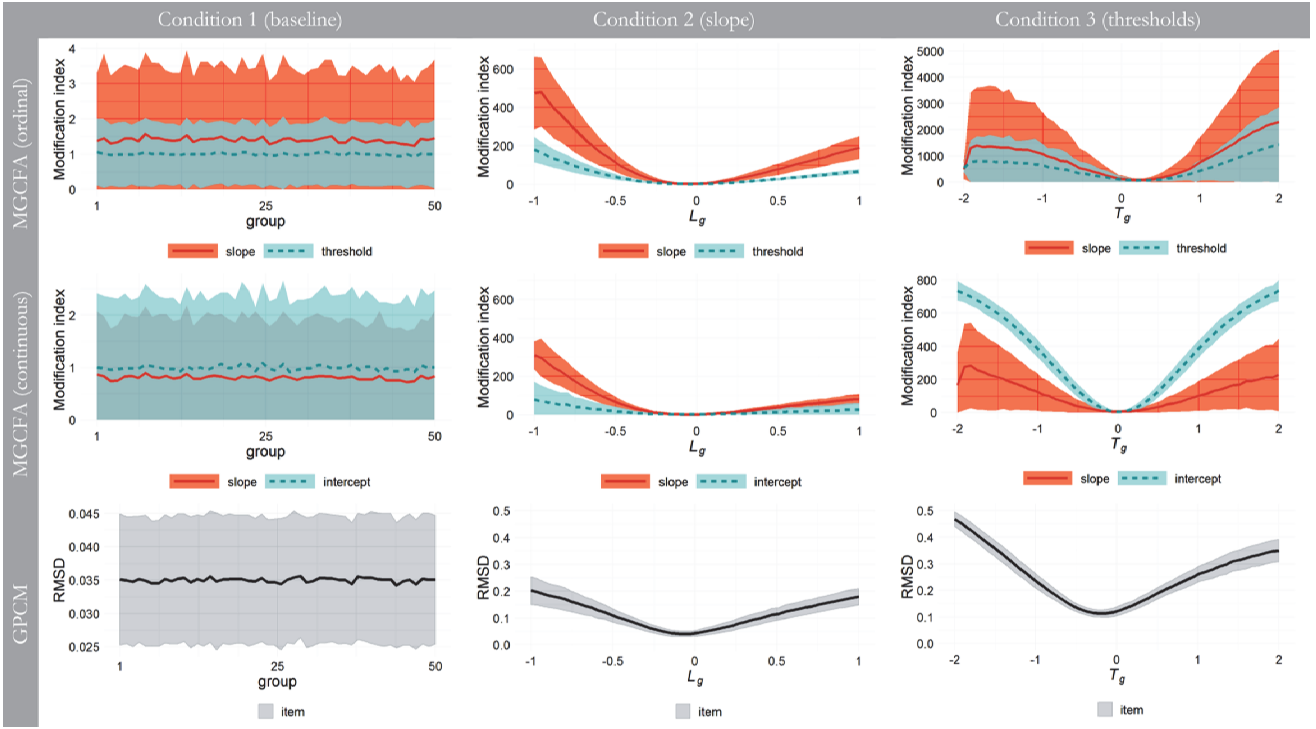

Figure 1:
Indices regarding the magnitude of local misfit (Modification index, RMSD) for each of the approaches (MGCFA- and GPCM-based) conditional on simulation condition. Lines represent mean and ribbon represents ±1 *SD* across 1,000 replications each. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds.

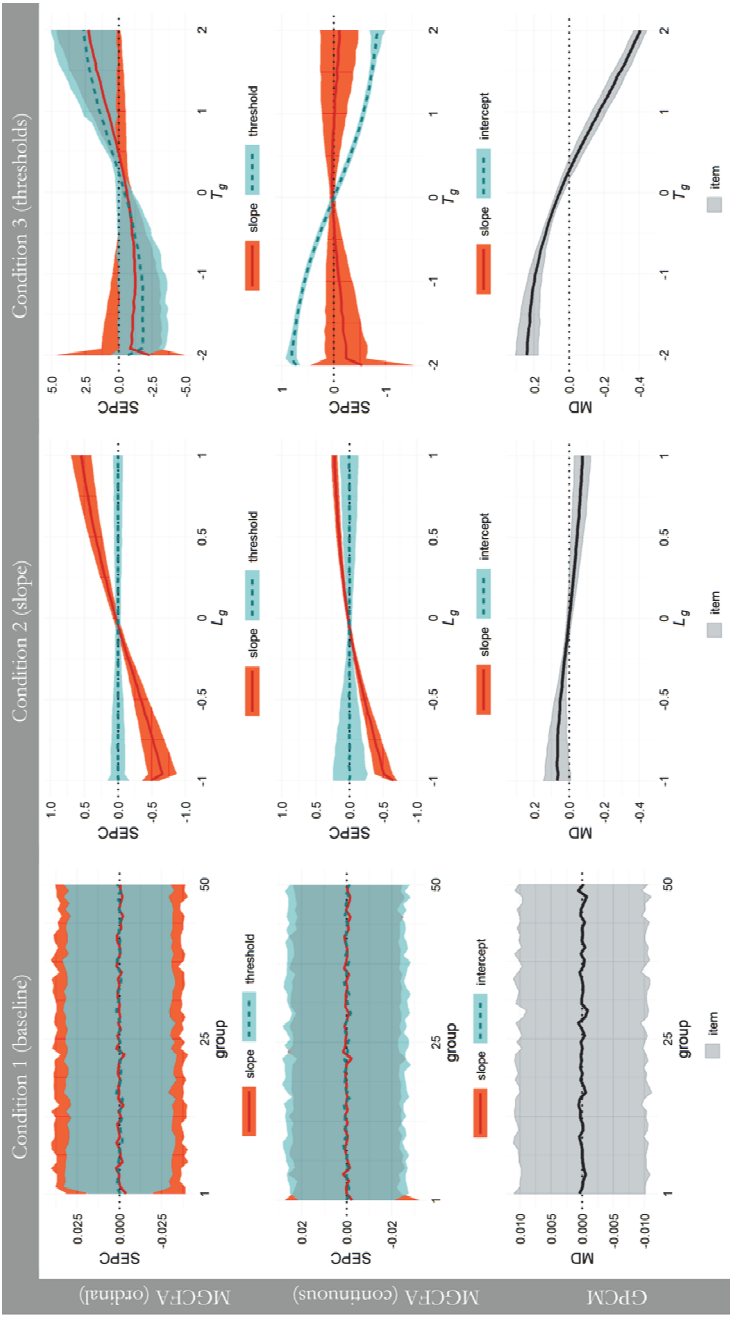

Figure 2:

Indices regarding the direction of local misfit for each of the approaches (MGCFA- and GPCM-based) conditional on simulation condition. Lines represent mean and ribbon represents ±1 *SD* across 1,000 replications each. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds.

While Table 3 as well as Figures 1 and 2 were based on the univariate properties of the indices of the magnitude and direction of local misfit, the next section focuses on the relationship among the indices resulting from the two MGCFA approaches on the one hand and the GPCM on the other hand. For each replication in each condition, the Pearson correlation between all of the indices on the magnitude of misfit was calculated. Table 4 contains descriptive statistics of the correlation coefficients, aggregated by simulation condition.

For both the CFA for ordinal and continuous indicators, the lowest correlations occurred in the condition in which MI holds (condition 1). Yet, the correlations are not zero either and higher for the ordinal CFA model, thus indicating that the parameters are not completely independent which could be explained through the presence of joint sampling variance underlying the data simulation. With respect to the other two conditions, the pattern differs between the two MGCFA approaches. For the continuous CFA model, the highest correlations can be observed between the sets of statistics that could be expected: In the condition with group differences in slope parameters (condition 2), RMSD correlates higher with slope-related MoIs (M = .752) than with intercept-related MoIs (M = .584); in the condition with group differences in threshold parameters (condition 3), RMSD correlates higher with intercept-related MoIs (M = .915) than with slope-related MoIs (M = .521). Under the ordinal CFA model, the pattern is flipped: when groups differ in slopes (condition 2), the average correlation of RMSD with the threshold-related MoI is higher (M = .0813) than with slope-related MoI (M = .783); when groups differ in

Table 4:
Average correlation between the indices regarding the magnitude of local misfit resulting from the different approaches (MGCFA- and GPCM-based) across replications by simulation condition.

	Condition		
	1 (baseline)	2 (slope)	3 (thresholds)
MGCFA (ordinal)			
RMSD with MoI (slope)	0.226 (0.151)	0.783 (0.101)	0.455 (0.270)
RMSD with MoI (threshold)	0.443 (0.121)	0.813 (0.041)	0.588 (0.115)
MoI (threshold) with MoI (slope)	0.627 (0.165)	0.937 (0.091)	0.708 (0.311)
MGCFA (continuous)			
RMSD with MoI (slope)	0.155 (0.143)	0.752 (0.057)	0.521 (0.128)
RMSD with MoI (intercept)	0.214 (0.143)	0.548 (0.114)	0.915 (0.023)
MoI (intercept) with MoI (slope)	0.263 (0.231)	0.642 (0.140)	0.570 (0.118)

Note. First value represents Mean, value in brackets *SD*. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds. MoI: modification index.

thresholds (condition 3), the average correlation of RMSD with slope-related MoI (M = .588) is higher than with threshold-related MoI (M = .455). Table 4 also contains findings for the correlations between the two sets of MoIs. Under both MGCFA approaches and across conditions, they are consistently larger than zero, thus indicating a dependency between these parameters. This is consistent with the observation made on the basis of Figure 1 in which both MoIs were found to react to group differences in either of the two parameters. This finding is surprising given that the parameters of the data generating model are independent from each other. As such, the introduction of group differences in either of the two parameters cannot have influenced the respective other parameter. Even more surprising is the fact that these correlations are much higher for the ordinal, data-generating model.

Parallel to the analyses above, we also calculated the Pearson correlation between all of the indices on the direction of local misfit. Table 5 contains the descriptive statistics of these correlations, aggregated by condition. In contrast to the findings reported above, the two MGCFA-based indices (SEPC) show a very similar pattern for their relationship with the GPCM-based index (MD): for both the ordinal and continuous MGCFA-model, MD correlates higher with slope- than with intercept-related SEPC when groups differ in slopes (condition 2), and MD correlates higher with intercept- than with slope-related SEPC when groups differ in thresholds (condition 3). Also in contrast with the findings reported above, the correlations among the two parameter-related SEPC indices are much lower than the correlations among the MoIs.

Table 5:
Average correlation between the indices regarding the direction of local misfit resulting from the different approaches (MGCFA- and GPCM-based) across replications by simulation condition.

	Condition		
	1 (baseline)	2 (slope)	3 (thresholds)
MGCFA (ordinal)			
MD with SEPC (slope)	-0.425 (0.354)	-0.785 (0.066)	-0.347 (0.733)
MD with SEPC (threshold)	-0.65 (0.096)	-0.537 (0.124)	-0.964 (0.012)
SEPC (threshold) with SEPC (slope)	0.008 (0.588)	0.000 (0.202)	0.313 (0.770)
MGCFA (continuous)			
MD with SEPC (slope)	-0.258 (0.275)	-0.742 (0.116)	-0.169 (0.601)
MD with SEPC (intercept)	0.632 (0.083)	0.579 (0.114)	0.955 (0.016)
SEPC (intercept) with SEPC (slope)	-0.005 (0.402)	-0.002 (0.257)	-0.155 (0.626)

Note. First value represents Mean, value in brackets *SD*. Conditions: (1) baseline condition with equal item parameters across all groups, (2) group differences with respect to item slope, (3) group differences with respect to item thresholds. SEPC: standardized expected parameter change; MD: mean deviation.

Empirical application

In the following, the consistency of findings resulting from the different approaches is investigated using the published data of the PISA 2015 assessment. A total of 58 scales based on data from the student, school principal, parent, and teacher questionnaires have been reported (for an overview, see OECD, 2017, as well as Appendix A) and are subject to this analysis.

Just as in the simulation study, data were estimated under (a) the MGCFA model assuming normality, and (b) the multiple-group GPCM. However, note that because of computational difficulties we refrained from reporting findings resulting from the ordinal MGCFA approach due to extremely skew distributions: many of the scales had too few observations for at least one group on at least one item. Removing such instances from the analysis would have strongly impaired our analysis of the consistency with findings from the GPCM approach. For approaches (a) and (b), we used the same software and model specifications as those detailed in the context of the simulation study above. In addition, we applied senate weights for both analyses to replicate the operational procedure of PISA 2015 (OECD, 2017); however, grouping was based on countries for ease of reporting, not on country-by-language interactions. These analyses yield one RMSD and two MoIs for each item of a scale in each country. The data can thus be described as fully-crossed, with each item being affiliated with both a scale and a country. The consistency between findings from the two approaches can therefore, in theory, be investigated from three different angles: whether the approaches are consistent in (1) identifying problematic items, (2) in identifying problematic scales, and (3) in identifying problematic groups. As such, findings for the consistency between indices of the magnitude and direction of local misfit resulting from either of the two approaches could potentially be (1) presented on the item level, (2) aggregated on the scale level, and (3) aggregated on the country level. However, MoIs indicate the improvement in global model fit in terms of likelihood (transformed into chi-square distributed quantities). As such, the statistic depends on the model's chi-square which in turn depends on the number of variables and cases. As a consequence of its unstandardized nature, MoIs cannot be compared between scales, so analyses of this empirical application are restricted to comparisons within scales.

Consistency on the scale level. For each of the 58 scales, Annex A contains descriptive statistics on CFA model fit of the scalar model, indices of the magnitude and direction of local misfit resulting from either of the two approaches, and coefficients for the correlation between these indices.

Many of the scales exhibit values on the indices of absolute model fit below common cutoff values (e.g. Byrne, 2012), thus indicating that scalar MI might not hold in all cases (or, even worse, that weaker levels of MI or even the measurement model as a whole does not fit the data). Results on local misfit consistently show larger values for MoIs relating to intercepts than those relating to slopes, pointing at group differences with respect to the items' difficulty throughout all of the scales. Most important to the research interest are the correlations among the indices of the magnitude and direction of local misfit resulting from either of the two approaches. Regarding the magnitude of

local misfit, RMSD and intercept-related MoIs correlate at r = .604 (SD = .145), on average, while RMSD and slope-related MoIs correlate at only r = .349 (SD = .150), on average. In addition, the correlations between RMSD and intercept-related MoIs were consistently higher throughout all of the scales. Regarding the direction of local misfit, MD and intercept-related SEPC correlate at r = .643 (SD = .201), on average, while MD and slope-related SEPC correlate at only r = -.149 (SD = .337), on average. For 55 out of 58 scales, the absolute value of the correlation coefficient between MD and intercept-related SEPC was higher than that of MD and the slope-related SEPC.

Discussion

Measurement invariance presents a fundamental prerequisite underlying valid country comparisons with respect to latent constructs. It is surprising to see how rarely MI testing is actually conducted in research practice, especially when it comes to non-cognitive constructs (Boer et al., 2018; Braeken & Blömeke, 2016; Rutkowski & Rutkowski, 2017). This study aimed at a better understanding of the relationship between statistical approaches of particular prominence, thus hoping to reduce potential confusion in selecting either one or the other, and ultimately increase the number of studies in which MI is tested before substantive research questions are addressed.

One of the methods under investigation in this study was a rather new, IRT-based approach that showed great promise in a particularly large ILSA while a second, CFA-based method is known to many applied researchers in the field. In addition, we included a third, CFA-based approach that accounts for the categorical nature of items typically found in questionnaires. Regarding the first two methods, results for both types of indices (pertaining to magnitude and direction of local misfit) in both the simulation study and the empirical application demonstrated a high consistency in identifying group differences in the threshold parameter representing the difficulty of responding in a high category of an item. In contrast, the consistency in identifying group differences in the slope parameter, representing the strength of the relationship between an indicator and the latent construct, was substantially lower. This raises the question as to which approach is superior in identifying the latter type of MI violation. While the truth underlying the empirical data is unknown, findings of the simulation study might help in answering this question. It needs to be noted that the scale of the respective statistics (GPCM-based RMSD and MGCFA-based modification indices; GPCM-based MD and MGCFA-based SEPC) cannot be compared per se. However, the variation of the respective statistic conditional on the degree of MI violations might be consulted (cf. Figures 1 and 2). According to this, RMSD is sensitive to both negative and positive deviations of a group's discrimination parameter from the true parameter, while the linear MGCFA's slope-related modification index appears to almost exclusively react to negative deviations, i.e., when the true slope parameter is close to zero. True slope parameters close to 2, in contrast, would likely be overlooked although this also presents a violation of MI. In this case, a particular indicator "drives" the measurement of the latent construct, thus shifting its meaning toward the content of the particular indicator, and ultimately threatening the comparability of the latent construct between groups. We also confirmed that

MD is “most sensitive” to misfit in the threshold parameter while RMSD is sensitive to misfit in both the discrimination and threshold parameter (cf. Khorramdel et al., 2019, p. 622).

Results for the third approach (ordinal MGCFA) are surprisingly indistinct given that this model served as the data-generating model. While modification index and SEPC proved to be sensitive to group differences in the slope parameter, both indices pointed at problems in both slope and threshold parameters when groups differed in thresholds. This pattern might be explained by the behavior of the model in which the thresholds indicate cut points of an underlying continuous latent response variable. When thresholds are fixed to certain inappropriate values in the scalar model, relaxing the slope parameter can also improve fit to the data as this scales the latent response variable.

As mentioned before, the two general frameworks for statistical approaches to testing measurement invariance comprising this study have developed independently of each other, each rooted in their own tradition, and they differ along a number of dimensions. In order to compare their implications regarding measurement invariance over and above a dichotomous decision (MI holds or doesn’t hold), we identified measures resulting from these approaches that quantify similar information on both the magnitude and direction of local model misfit; however, attention needs to be placed on the conceptual differences between them. While modification indices and SEPC quantify misfit separately for slopes and intercepts, RMSD and MD measure an item’s *overall* goodness-of-fit, indicating misfit relating to the discrimination parameter, the threshold parameters, or both. In addition, modification indices and SEPC exclusively quantify misfit due to constraining parameters to be equal across groups while RMSD and MD quantify misfit that can be caused by general item misfit, by group differences in the item’s parameters, or by a combination of both.

Although we tried to keep the simulation setup realistic, some factors limiting the generalizability of findings need to be discussed. First, we simulated the sample size to be equal with 1,000 cases per group, the number of groups to be 50, the number of response categories to be 4, and the number of indicator items to be 5. The number of response categories and the number of indicator items correspond to the respective median across the 58 questionnaire scales in PISA 2015; sample size and the number of groups, in contrast, were chosen to be more representative of other ILSAs that are typically smaller than PISA. In addition, attention needs to be placed on the type and extent of MI violations that were implemented. With respect to type, we assumed group parameters to vary more or less but symmetrically around a central parameter. Other patterns, however, are plausible, for example when one set of groups differs in their joint parameter from the joint parameter shared by the remaining set of groups. With respect to the extent of MI violations implemented in this study, we chose the most severe values of shift parameters to result in parameters that would still be plausible. However, even more severe violations could be investigated. Finally, it should also be noted that when applying the MGCFA to ordinal response data, we explicitly violated the non-linearity underlying the data generation in the simulation study and the nature of the categorical response data in the empirical application. As mentioned before, the decision of including this model was

based on the observation according to which continuous MFCFA is often applied in practice.

With respect to the empirical study, we aggregated our findings on consistency on the scale level (across countries) to evaluate the two approaches' consistency in identifying problematic scales. It would have been conceptually appealing to also aggregate findings on the country level (across scales) to evaluate the two approaches' consistency in identifying problematic countries. However, modification indices are on the metric of model fit in terms of chi-square and, as such, confounded with properties of the data (numbers of variables and cases). It would be desirable to single out these scale properties to allow for comparisons of modification indices across scales.

References

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*, 713–734. doi:10.1177/0022022117749042

Braeken, J. & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education, 41*, 733–749. doi:10.1080/02602938.2016.1161005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: basic concepts, applications, and programming*. New York: Routledge.

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

Chou, C.-P., & Bentler, P. M. (1993). Invariant standardized estimated parameter change for model modification in covariance structure analysis. *Multivariate Behavioral Research, 28*, 97–110.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: across-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 1–10. doi:10.3389/fpsyg.2014.00982

Gonzalez, E. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. In D. Hastedt & M. von Davier (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 5, pp. 117–134). Hamburg, Germany and Princeton, NJ: IEA-ETS Research Institute.

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

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J. & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. doi:10.1080/00313831.2016.1258726

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133. doi:10.1007/BF02291393

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202. doi:10.1007/BF02289343

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models. Methodology of educational measurement and assessment* (Ch. 30). Cham (CH): Springer.

Köhler, C., Robitzsch, A., & Hartig, J. (in press). A bias corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.) (2016). *Assessing contexts of learning. An international perspective*. New York: Springer International Publishing. doi:10.1007/978-3-319-45357-6

MacCallum, R., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111(3)*, 490–504. doi:10.1037/0033-2909.111.3.490

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

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. doi:10.1002/j.2333-8504.1992.tb01436.x

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes No. 4). Los Angeles: University of California, Los Angeles.

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

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

Oliveri, M. E. & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Oliveri, M. E. & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*, 1–21. doi:10.1080/15305058.2013.825265

Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the cross-country comparability of indicators of socioeconomic resources in PISA. *Applied Measurement in Education, 30(*4). 234–258. doi:10.1080/08957347.2017.1353985

Rutkowski, L. & Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research 62*(3), 354–367. doi: 10.1080/00313831.2016.1261044

Rutkowski, L. & Rutkowski, D. (2010). Getting it better: The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies, 42*(3), 411–430. doi:10.1080/00220272.2010.487546

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

Saris,W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 105–129). San Francisco, CA: Jossey-Bass.

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371–384. doi:10.1007/BF02294623

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K. & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology, 6*, 1064. doi:10.3389/fpsyg.2015.01064

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

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: ETS.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The journal of experimental education, 80*(1), 26–44.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical Report of the Survey of Adult Skills (PIAAC)* (chapter 17). Retrieved from http://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Appendix A:

Scales based on the PISA 2015 questionnaires: properties, MGCFA model fit indices, descriptive statistics (*M*, *SD*) for indices regarding the magnitude of local misfit resulting from the two approaches to measurement invariance testing (MoI, RMSD), and Pearson coefficients for the correlation between them.

	Ques-tion	n_i	n_c	N	n_g	CFI	TLI	RMSEA	$\chi^2(df)$	MoI (inter-cept)	MoI (slope)	RMSD	r (MoI (intercept), RMSD)	r (MoI (slope), RMSD)	r (SEPC (intercept), MD)	r (SEPC (slope), MD)
ENTUSE	IC008	12	5	294882	47	0.530	0.589	0.126	355374 (3550)	199.2 (330.37)	37.72 (103.9)	0.11 (0.04)	0.754	0.515	0.833	-0.044
HOMESCH	IC010	12	5	288923	47	0.648	0.693	0.127	355870 (3550)	189.33 (315.82)	43.37 (111.74)	0.14 (0.04)	0.469	0.141	0.816	-0.395
USESCH	IC011	9	5	289188	47	0.754	0.793	0.108	145894 (2005)	149.82 (248.15)	54.93 (107.77)	0.11 (0.04)	0.453	0.264	0.864	-0.593
INTICT	IC013	6	4	286550	47	0.802	0.842	0.098	53022 (883)	120.36 (186.75)	15.33 (30.27)	0.12 (0.05)	0.740	0.347	0.716	-0.088
COMPICT	IC014	5	4	283366	47	0.854	0.886	0.102	38567 (603)	66.34 (118.44)	33.29 (67.95)	0.11 (0.05)	0.625	0.399	0.577	-0.499
AUTICT	IC015	5	4	283470	47	0.823	0.862	0.114	47963 (603)	132.25 (201.9)	29.41 (48.57)	0.13 (0.05)	0.687	0.273	0.733	-0.216
SOIAICT	IC016	5	4	280508	47	0.916	0.935	0.077	22158 (603)	52.38 (78.04)	16.27 (26.88)	0.11 (0.03)	0.479	0.275	0.665	-0.279
PRESUPP	PA002	10	4	86077	17	0.534	0.596	0.119	64186 (883)	161.92 (252.35)	20.52 (39.36)	0.1 (0.05)	0.689	0.308	0.917	0.202
CURSUPP	PA003	8	5	91014	18	0.578	0.645	0.133	53984 (598)	209.88 (289.3)	39.72 (70.41)	0.12 (0.06)	0.753	0.214	0.523	0.499
EMOSUPP	PA004	4	4	90693	18	0.881	0.907	0.094	6323 (138)	68.13 (149.19)	18.78 (33.98)	0.08 (0.05)	0.704	0.275	0.457	-0.169
PASCHPOL	PA007	6	4	90581	18	0.719	0.771	0.138	32343 (332)	153.5 (279.64)	34.33 (70.46)	0.12 (0.06)	0.854	0.702	0.707	0.427
PQSCHOOL	PA007	7	4	90687	18	0.908	0.924	0.075	13511 (456)	49.48 (72.63)	15.95 (32.91)	0.09 (0.04)	0.721	0.421	0.493	-0.291
PQGENSCI	PA033	5	4	89494	18	0.815	0.852	0.126	18036 (226)	129.29 (179.94)	19.56 (28.92)	0.1 (0.06)	0.860	0.227	0.375	0.213

	Question	n_i	n_c	N	n_g	CFI	TLI	RMSEA	$\chi^2(df)$	MoI (intercept)	MoI (slope)	RMSD	r (MoI (intercept), RMSD)	r (MoI (slope), RMSD)	r (SEPC (intercept), MD)	r (SEPC (slope), MD)
PQENPERC	PA035	7	4	89395	18	0.836	0.864	0.058	8198 (456)	24.67 (35.36)	13.11 (30.85)	0.09 (0.05)	0.637	0.246	0.214	-0.115
PQENVOPT	PA036	7	3	89426	18	0.901	0.918	0.080	14951 (456)	51.86 (72.1)	19.3 (39.56)	0.07 (0.03)	0.408	0.154	0.914	0.39
LEAD	SC009	13	6	15477	69	0.571	0.623	0.123	26791 (6117)	10.46 (14.52)	4.32 (6.63)	0.22 (0.07)	0.466	0.362	0.362	-0.284
LEADCOM	SC009	4	6	15475	69	0.662	0.743	0.132	2685 (546)	8.66 (12)	4.91 (7.72)	0.2 (0.08)	0.347	0.222	0.258	-0.194
LEADINST	SC009	3	6	15445	69	0.716	0.784	0.143	1514 (272)	8.24 (11.12)	3.43 (6.39)	0.16 (0.05)	0.498	0.358	0.369	-0.207
LEADPD	SC009	3	6	15439	69	0.809	0.855	0.133	1356 (272)	6.57 (8.33)	5.78 (7.97)	0.16 (0.05)	0.412	0.298	0.276	-0.204
LEADTCH	SC009	3	6	15428	69	0.696	0.768	0.157	1773 (272)	8.4 (11.8)	2.9 (4.84)	0.18 (0.07)	0.630	0.304	0.435	-0.126
EDUSHORT	SC017	4	4	15504	69	0.501	0.622	0.235	7302 (546)	7.43 (10.47)	5.17 (13.24)	0.15 (0.07)	0.372	0.119	0.804	-0.635
STAFFSHORT	SC017	4	4	15508	69	0.517	0.634	0.164	3862 (546)	8.47 (10.56)	3.47 (5.38)	0.15 (0.06)	0.543	0.408	0.838	0.229
STUBEHA	SC061	5	4	15431	69	0.573	0.669	0.173	6832 (889)	11.09 (14.67)	6.84 (10.68)	0.18 (0.08)	0.573	0.375	0.851	0.568
TEACHBEHA	SC061	5	4	15422	69	0.679	0.751	0.132	4366 (889)	9.57 (13.8)	6.23 (10.15)	0.15 (0.06)	0.586	0.536	0.882	0.585
CULTPOSS	ST011	5	4	466686	68	0.342	0.489	0.157	149051 (876)	400.51 (562)	96.95 (196.96)	0.09 (0.05)	0.751	0.418	0.981	0.404
HEDRES	ST011	7	2	467110	68	0.074	0.247	0.101	124306 (1756)	179.78 (280.99)	94.73 (211.93)	0.07 (0.05)	0.763	0.565	0.931	-0.171
HOMEPOS	ST011	25	6	393200	56	0.000	0.036	0.097	1209849 (18040)	390.74 (609.63)	184.85 (358.32)	0.09 (0.08)	0.361	0.149	0.672	-0.297
ICTRES	ST011	6	4	463012	67	0.284	0.430	0.126	139926 (1263)	245.42 (397.05)	190.4 (363.48)	0.07 (0.05)	0.529	0.282	0.706	-0.195
WEALTH	ST011	12	4	392957	56	0.000	0.062	0.128	487530 (4234)	475.88 (693.79)	283.49 (496.98)	0.08 (0.08)	0.133	0.079	0.404	-0.306

	Question	n_i	n_c	N	n_g	CFI	TLI	RMSEA	$\chi^2(df)$	MoI (intercept)	MoI (slope)	RMSD	r (MoI (intercept), RMSD)	r (MoI (slope), RMSD)	r (SEPC (intercept), MD)	r (SEPC (slope), MD)
BELONG	ST034	6	4	454280	67	0.739	0.793	0.111	107474 (1263)	99.08 (173.08)	34.11 (72.81)	0.12 (0.05)	0.713	0.448	0.501	0.169
COOPERATE	ST082	4	4	396788	55	0.831	0.871	0.080	20502 (434)	105.37 (141.53)	29.13 (53.34)	0.09 (0.03)	0.744	0.483	0.62	-0.501
CPSVALUE	ST082	4	4	397108	55	0.892	0.918	0.087	24110 (434)	96.86 (162.67)	41.86 (66.85)	0.1 (0.04)	0.728	0.370	0.552	-0.223
ENVAWARE	ST092	7	4	424106	67	0.747	0.794	0.130	185532 (1730)	231.94 (335.71)	21.26 (45.21)	0.15 (0.06)	0.635	0.367	0.741	0.102
ENVOPT	ST093	7	3	372543	55	0.902	0.920	0.076	56528 (1418)	65.57 (111.58)	36.32 (75.03)	0.08 (0.04)	0.522	0.424	0.883	-0.848
JOYSCIE	ST094	5	4	436011	68	0.937	0.951	0.083	39719 (876)	84.99 (126.46)	18.03 (38.54)	0.11 (0.04)	0.602	0.303	0.549	-0.365
INTBRSCI	ST095	5	5	365515	55	0.800	0.844	0.117	64949 (707)	82.47 (188.95)	24.6 (53.77)	0.1 (0.04)	0.536	0.474	0.765	-0.411
DISCLISCI	ST097	5	4	417366	68	0.905	0.926	0.096	50775 (876)	94.74 (126.01)	21.45 (63.94)	0.11 (0.04)	0.661	0.450	0.561	-0.07
IBTEACH	ST098	9	4	412439	68	0.684	0.734	0.122	266065 (2908)	254.32 (378.03)	78.71 (134.7)	0.13 (0.06)	0.647	0.466	0.86	-0.625
TEACHSUP	ST100	5	4	412032	68	0.887	0.912	0.106	60418 (876)	149.82 (226.54)	26.46 (55.45)	0.12 (0.05)	0.606	0.329	0.766	-0.543
TDTEACH	ST103	4	4	405589	67	0.814	0.859	0.136	59638 (530)	189.44 (246.23)	44.95 (111.05)	0.11 (0.04)	0.643	0.418	0.785	-0.49
PERFEED	ST104	5	4	403577	67	0.892	0.916	0.105	58656 (863)	80.57 (150.03)	26.97 (65.9)	0.11 (0.04)	0.426	0.191	0.785	-0.624
ADINST	ST107	3	4	339087	54	0.903	0.926	0.106	15178 (212)	121.29 (190.81)	31.3 (71.97)	0.08 (0.03)	0.637	0.485	0.843	-0.501
INSTSCIE	ST113	4	4	428711	68	0.943	0.957	0.081	22649 (538)	28.95 (47.71)	22.51 (34.95)	0.1 (0.03)	0.330	0.130	0.291	-0.314
ANXTEST	ST118	5	4	398783	55	0.819	0.859	0.113	66740 (707)	157.3 (208.99)	34.48 (57.57)	0.11 (0.04)	0.604	0.206	0.653	-0.156
MOTIVAT	ST119	5	4	397921	55	0.681	0.752	0.145	107559 (707)	278.99 (481.23)	172.01 (243.46)	0.14 (0.06)	0.682	0.537	0.665	-0.372

	Question	n_i	n_c	N	n_g	CFI	TLI	RMSEA	$\chi^2(df)$	MoI (intercept)	MoI (slope)	RMSD	r (MoI (intercept), RMSD)	r (MoI (slope), RMSD)	r (SEPC (intercept), MD)	r (SEPC (slope), MD)
EMOSUPS	ST123	4	4	393939	54	0.888	0.915	0.101	32054 (426)	130.74 (235.85)	22.83 (37.51)	0.09 (0.05)	0.775	0.315	0.559	-0.278
SCIEEFF	ST129	8	4	424750	68	0.888	0.908	0.076	85596 (2298)	103.02 (167.5)	27.47 (59.67)	0.12 (0.04)	0.495	0.344	0.659	-0.614
EPIST	ST131	6	4	425482	68	0.876	0.902	0.089	64786 (1282)	47.45 (86.44)	15.64 (35.31)	0.11 (0.04)	0.633	0.435	0.18	-0.286
SCIEACT	ST146	9	4	364223	55	0.784	0.818	0.105	172716 (2349)	54.83 (98.91)	29.05 (54.4)	0.09 (0.04)	0.535	0.301	0.86	-0.661
SATJOB	TC026	4	4	88142	18	0.733	0.791	0.161	17707 (138)	106.26 (211.78)	48.46 (82.62)	0.09 (0.05)	0.788	0.018	0.523	-0.082
SATTEACH	TC026	4	4	88157	18	0.684	0.753	0.163	18118 (138)	209.92 (340.63)	51.74 (97.27)	0.12 (0.06)	0.807	0.487	0.683	0.147
TCEDUSHORT	TC028	4	4	88040	18	0.594	0.682	0.272	50167 (138)	95.68 (137.28)	21.43 (32.36)	0.1 (0.05)	0.503	0.040	0.765	0.092
TCSTAFFSHORT	TC028	4	4	87939	18	0.657	0.732	0.190	24582 (138)	259.04 (294.9)	69.03 (92.13)	0.11 (0.05)	0.680	0.258	0.831	-0.04
COLSCIT	TC031	8	4	29457	18	0.811	0.841	0.119	14490 (598)	68.84 (112.17)	28.59 (56.62)	0.13 (0.06)	0.718	0.539	0.62	-0.6
SETEACH	TC033	4	4	27062	17	0.863	0.893	0.109	2611 (130)	36.91 (51.91)	23.63 (35.44)	0.1 (0.05)	0.611	0.536	0.468	-0.167
SECONT	TC034	4	4	27066	17	0.793	0.837	0.147	4579 (130)	68.19 (90.88)	23.27 (44.74)	0.12 (0.06)	0.703	0.550	0.475	-0.379
EXCHT	TC046	4	6	58625	18	0.578	0.670	0.196	17388 (138)	247.56 (290.19)	96.2 (134.32)	0.14 (0.06)	0.593	0.400	0.604	-0.36
TCLEAD	TC060	5	4	108292	18	0.832	0.866	0.092	11683 (226)	46.97 (85.55)	42.67 (69.75)	0.1 (0.05)	0.692	0.682	0.66	-0.649

Note. Question: Question-ID of indicator items, n_i: number of indicator items, n_c: number of response categories, N: total number of respondents (regardless of country), n_g: number of countries underlying MI testing, MoI (intercept): modification index pertaining to intercept parameter, MoI (slope): modification index pertaining to slope parameter, r: correlation coefficient