# A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance

Pokropek, Artur ; Davidov, Eldad ; Schmidt, Peter

Abstract: Several structural equation modeling (SEM) strategies were developed for assessing measurement invariance (MI) across groups relaxing the assumptions of strict MI to partial, approximate, and partial approximate MI. Nonetheless, applied researchers still do not know if and under what conditions these strategies might provide results that allow for valid comparisons across groups in large-scale comparative surveys. We perform a comprehensive Monte Carlo simulation study to assess the conditions under which various SEM methods are appropriate to estimate latent means and path coefficients and their differences across groups. We find that while SEM path coefficients are relatively robust to violations of full MI and can be rather effectively recovered, recovering latent means and their group rankings might be difficult. Our results suggest that, contrary to some previous recommendations, partial invariance may rather effectively recover both path coefficients and latent means even when the majority of items are noninvariant. Although it is more difficult to recover latent means using approximate and partial approximate MI methods, it is possible under specific conditions and using appropriate models. These models also have the advantage of providing accurate standard errors. Alignment is recommended for recovering latent means in cases where there are only a few noninvariant parameters across groups.

DOI: https://doi.org/10.1080/10705511.2018.1561293

# A Monte Carlo Simulation Study to Assess the Appropriateness of Traditional and Newer Approaches to test for Measurement Invariance

Artur Pokropek
*Institute of Philosophy and Sociology, Polish Academy of Sciences*

Eldad Davidov
*University of Cologne, Faculty of Management, Economics and Social Sciences, The Institute of Sociology and Social Psychology;*
*University of Zurich, Department of Sociology, and URPP Social Networks*

Peter Schmidt
*Department of Political Science, Justus Liebig University Giessen*

Correspondence should be sent to Dr hab. Artur Pokropek

This is an electronic version of an article published online in ***Structural Equation Modeling*** on 28-Jan-2019. This journal is available online at: www.tandfonline.com. The definitive publisher-authenticated version of this article is available online under

https://doi.org/10.1080/10705511.2018.1561293

# APPROACHES TO TEST FOR MEASUREMENT INVARIANCE

Note: This work has been prepared under the project Scales Comparability in Large-Scale Cross-country Surveys, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934). Eldad Davidov would like to thank the University of Zurich Research Priority Programme “Social Networks” for their support. The work of Peter Schmidt was supported by the Alexander von Humboldt Polish Honorary Research Fellowship granted by the Foundation for Polish Science for the international cooperation of Peter Schmidt with Jan Cieciuch. The authors would like to thank Lisa Trierweiler for the English proof of the manuscript.

The second and third authors are ordered alphabetically.

# A Monte Carlo Simulation Study to Assess the Appropriateness of Traditional and Newer Approaches to test for Measurement Invariance

Comparative analysis may take different forms: It may involve comparisons across national groups, cultural groups, time points, or samples collected using different modes, just to name a few possibilities. In these types of studies, measurement invariance (MI, also often called measurement equivalence) is a necessary condition to allow meaningful comparisons of means or associations such as covariances and unstandardized regression coefficients across groups (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Meredith, 1993; Millsap 2011;Steenkamp & Baumgartner, 1998). However, researchers often experience difficulties in achieving sufficient levels of invariance, especially when scalar invariance should be reached, when the number of groups is large or when cultural differences are significant (Marsh et al., 2017).

Several authors have thus argued that measurement parameters need not be equal across groups for all indicators. Valid comparisons across groups of means or associations can also be made if only a subset of indicators functions equivalently; this situation is called *partial equivalence* (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). However, some other studies provided contradictory evidence based on newer simulations suggesting that partial equivalence may not always be sufficient for meaningful cross-group comparisons (Brown, 2015; De Beuckelaer & Swinnen, 2018; Steinmetz, 2018; Vandenberg & Lance, 2000). Thus, according to the state of the art of the literature, which is till now rather limited and based only on a few different simulations with a restricted number of conditions (see Table 1), researchers are often unsure whether partial invariance is, in fact, sufficient for meaningful comparisons or not.

Recently, new approaches and methods for modeling MI as *approximate* (rather than exact) measurement invariance (Asparouhov & Muthén, 2014, 2017) were developed and introduced. Although very promising for applied researchers, the adequacy of these methodologies was examined in only a small number of simulation studies (Muthén & Asparouhov, 2013; van de Schoot et al., 2013).

On top of that, partial and approximate MI settings might be combined (Muthén & Asparouhov, 2013). It could well be the case that some items may be approximately invariant across groups whereas others are not (e.g., Zercher, Schmidt, Cieciuch, & Davidov, 2015). However, applied researchers still do not know if and under what conditions partial, approximate, and

partial approximate MI provide results that allow for valid comparisons across groups in large-scale comparative surveys. Obviously, new studies with further simulations that cover additional conditions and new methods are needed to provide more informative recommendations for applied researchers on whether they may rely on partial, exact, approximate, or partial approximate measurement equivalence when full equivalence is not given (Davidov et al., 2014).

The main aim of this paper is to contribute to this ongoing research by testing under various conditions whether partial, approximate, and approximate partial MI is sufficient or not for meaningful comparisons. For the simulations we use real-life conditions that are often encountered by researchers dealing with large-scale international survey data, such as the European Social Survey (ESS), the International Social Survey Program (ISSP), the Program for International Student Achievement (PISA) Studies, or the World Value Study (WVS), just to name a few. By doing so, we hope to be in a better position to answer the question of how many noninvariant items may be present in the model without risking noncomparability under different plausible real-life conditions present in large-scale survey data. The ultimate goal of this paper is to provide useful guidance for applied researchers who encounter different types of MI in their analyses as to whether they may carry out meaningful comparisons or not.

## Models for Multigroup Analysis and MI

In this paper, we are focusing on five types of multigroup confirmatory factor analysis (CFA) models to test for MI, namely: multigroup confirmatory factor analysis (MG-CFA, designed for full exact MI), partial multigroup confirmatory factor analysis (PMG-CFA, designed for partial exact MI), multigroup Bayesian SEM (MG-BSEM, designed for full approximate MI), partial multigroup Bayesian SEM (PMG-BSEM, designed for partial approximate MI)[1], and MG-CFA with alignment optimization (AMG-CFA, designed for partial approximate MI). These models are supplemented by a structural part (with additional dependent variables) to also examine the recovery of path coefficients. Although it does not exhaust all the possibilities—we do not consider, for example, multilevel CFA and CFA mixture modeling—the chosen set is characterized by a comparable parametrization with similar possibilities of parameter constraints that allow addressing partial MI. Moreover, the methods in the chosen set are arguably most often used by applied researchers in the context of cross-group analysis.

### Multigroup Confirmatory Factor Analysis Model (MG-CFA)

CFA and its multigroup extension (Jöreskog, 1971) is the most used approach for performing cross-group comparisons based on constructed scales in cross-cultural surveys. In a CFA, observed items are indicators of constructs

1 We follow the convention introduced by Muthén and Asparouhov (2012) and use the term BSEM for models designed for dealing with approximate MI even when these models include only a measurement but no structural part.

that were targeted to be measured. The CFA model assumes that the relation between a latent factor and observed items can be expressed using item related parameters such as a factor loading and an item intercept. A classical CFA and its multigroup extension (MG-CFA), that allow for estimating group means, variances, and group-specific parameters, assume that the observed indicator $Y_{ig}$ is continuous and the relation between the latent trait $\theta$ and observed indicators $Y_{ig}$ is described by a linear equation (for a simple one-dimensional case):

$$y_{ig} = \tau_{ig} + \lambda_{ig}\eta_g + \epsilon_{ig} \quad (1)$$

where $\tau_{ig}$ describes the intercept while $\lambda_{ig}$ indicates the factor loading of the item in group *g*. The index *i* denotes the item number, *g* the group belonging, and $\epsilon_{ig}$ denotes a random error. Factor loadings may be interpreted as slopes in a regression analysis, and they provide information on how the predicted value of an indicator differs with a change of one unit in the value of the latent variable, while the intercepts provide information about the expected mean value of $Y_{ig}$ when $\eta_g = 0$ .

The exact MI test constrains factor loadings and item intercepts to be equal across all groups: $\tau_{ig} = \tau_{ig'}$ and $\lambda_{ig} = \lambda_{ig'}$. When all factor loadings and item intercepts are constrained to be exactly the same across groups and the model fits the data, it implies that exact scalar invariance is given and mean estimates of latent factors are comparable (Meredith, 1993; Millsap, 2011). However, exact scalar equivalence is rarely achieved when using real survey data (Davidov et al., 2014).

**Partially Invariant Multigroup Confirmatory Factor Analysis Model (PMG-CFA)**

In the *partial equivalence approach,* some item parameters are constrained to be equal across groups whereas others are estimated freely while relaxing the assumptions of the classical full exact invariance analysis using multigroup CFA models (Byrne et al., 1989). The approach allows the comparison of latent means and their associations with other constructs across groups. However, there is no consensus on how many items with equal factor loadings and intercepts are required to achieve unbiased estimates of latent traits. Byrne and colleagues (1989) argued that the partial invariance model requires at least two factor loadings and intercepts to be equal across groups, with one of them being the so-called anchor item. Some other researchers (e.g., Reise, Widaman, & Pugh, 1993) indicated that a *majority* of the items should be invariant to achieve meaningful comparisons. However, the number of simulation studies to support this claim is very limited. Steinmetz (2013) showed that for four and six items in small sample situations ($N = 100$ and $N = 300$), half of the items has to be invariant to achieve meaningful comparisons. Literature that focused on cognitive testing using IRT models provided evidence that 20-25% of invariant items (linking or anchor items in this jargon) is sufficient to

compare two populations (Hambleton, Swaminathan, & Rogers, 1991; Kolen & Brennan, 2004). However, one needs to keep in mind that this requirement applies to a specific situation where the number of items and the sample sizes are very large. Thus, these studies did not consider conditions that are particularly common when analyzing large-scale international survey data, such as a rather small number of items per latent variable or sample sizes of approximately 1,000 to 2,000 respondents.

Both MG-CFA and PMG-CFA might be easily incorporated into a full SEM framework by adding relations between the latent variable and other variables (Jöreskog, 1971). The measurement part (represented by the CFA model) and the path part (represented by the relation to another variable) are then estimated jointly in the SEM framework. The impact of the presence or absence of MI on estimated path coefficients in SEM modeling was seldom considered by researchers (Guenole & Brown, 2014). In our simulation study, we consider multigroup SEM models, where the independent (exogenous) variable is represented by a latent variable measured by multiple indicators in a CFA model and the dependent variables are assumed to be manifest (measured each by a single indicator without controlling for measurement error (see Figure 1).

### Multigroup Bayesian SEM (MG-BSEM) and Partial MG-BSEM (PMG-BSEM)

As previously indicated, the classical exact MI analysis based on the frequentist approach (see Muthén & Asparouhov, 2012; van de Schoot et al., 2013) constrains factor loadings and item parameters to equality across all groups: $\tau_{ig} = \tau_{ig'}$ and $\lambda_{ig} = \lambda_{ig'}$. In the approximate invariance approach proposed by Muthén and Asparouhov (2012, 2013), these constraints are relaxed by assuming that item related parameters are approximately equal: $\tau_{ig} \approx \tau_{ig'}$ and $\lambda_{ig} \approx \lambda_{ig'}$. This is done by introducing cross-group variation between item parameters, similar to a multilevel CFA, by using zero-mean small-variance informative priors for the parameters and a Bayesian analysis (Muthén & Asparouhov, 2013). Similar to an MG-CFA model, an MG-BSEM model might address the problem of noninvariance of some measurement parameters. An MG-BSEM model could accommodate both full and partial approximate MI by allowing for “wiggle room” (van de Schoot et al., 2013) for some or all parameters while allowing full noninvariance for some others. In fact, BSEM might be even more flexible, allowing for exact MI for some parameters, exact measurement noninvariance for some other parameters, and approximate MI for the remaining parameters. In this simulation study, we focus only on two model specifications. The first one is MG-BSEM where all parameters were assumed to be approximately measurement invariant and the second is PMG-BSEM where some of the parameters are freely estimated (assuming exact measurement noninvariance for them) while for other parameters approximate MI is imposed.

The number of past studies conducting *simulations* considering BSEM MI

analysis is very small (see Table 1). Below, we briefly describe the results of some of them. Muthén and Asparouhov (2013) conducted a simulation study that included 10 groups, one factor, and 6 items with a sample size of 500. This study introduced various levels of bias for loadings and intercepts on different items and generated conditions that were more similar to the partial noninvariance than to the approximate measurement situation. They tested five models: (1) exact MI (MG-CFA model); (2) approximate MI (MG-BSEM); (3) a combined model with exact MI for invariant items and approximate MI (BSEM) for other items; (4) a combined model with free noninvariant parameters and approximate invariance for moderately noninvariant parameters (PMG-BSEM); and finally, (5) a model with exact MI for the invariant parameters while allowing noninvariant parameters to be freely estimated (PMG-CFA). The authors found that the worst bias was generated by using an MG-BSEM model that imposed small noninvariance while ignoring large noninvariant parameters. The MG-BSEM performed even worse than the MG-CFA model that constrained exact equality on noninvariant parameters. This work concluded that PMG-CFA provided the most accurate estimates and that the PMG-BSEM model was the second-best model.

A second simulation study (van de Schoot et al., 2013) investigated seven groups and differently varying intercepts across groups. The study suggested that the MG-BSEM model behaves relatively well in the approximate measurement invariant situation but should not be applied when partial MI is present. In their simulation study with two groups and a medium-sized sample size of 435 observations, Chiorri, Day, and Malmberg (2014) concluded that PMG-BSEM and PMG-CFA models give similar results when partial noninvariance is present for both loadings and intercepts.

In sum, based on the presented simulation studies, the simple PMG-CFA model provided the best results when partial MI was present. Furthermore, MG-BSEM provided biased results when it was applied on partially invariant models and was less robust than MG-CFA when misspecifications were present. However, it remains to be answered whether the PMG-BSEM model outperforms the PMG-CFA model in the presence of partial noninvariance under specific conditions in which some items are not invariant and others are only approximately invariant. In addition, as evident in Table 1, the results presented are based on a very limited number of simulated conditions. This makes it difficult to generalize them to other situations.

### MG-CFA with Alignment Optimization (AMG-CFA)

Another method that could account for partial invariance is the alignment optimization procedure (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2017). This procedure replaces the cross-group equality constraints with a technique similar to the rotation in EFA (exploratory factor analysis). An algorithm estimates a solution that minimizes overall differences between groups' parameters using a simplicity function, which is optimized at a few large noninvariant parameters and many approximately invariant parameters. To date, several studies have applied this method to large-scale

survey data (see, e.g., Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Seddig, Maskileyson, & Davidov, in press; Munck, Barber, & Torney-Purta, 2017).

Asparouhov and Muthén (2014) provided support for the AMG-CFA model on the basis of a simulation and also using real data. In their simulation study, they varied the sample size (100 or 1,000 per group), the number of groups (2, 3, 15, or 60), and the extent of noninvariance (0%, 10%, or 20% noninvariant items). Their results demonstrated that known population parameters were accurately estimated even when there was substantial noninvariance, particularly when sample sizes were large. In conclusion, the authors recommended to rely on the means estimated using the alignment procedure when there are less than 25% noninvariant parameters, because under these circumstances the alignment procedure can provide a good recovery of factor means and factor variances. Moreover, according to Asparouhov and Muthén (2014), the alignment method can work very well even with a small number of indicators. A newer study by Flake and McCoach (2018), using a two-factor model with seven items per factor, confirmed these conclusions.

Table 1 summarizes the results of previous simulation studies on the parameter recovery in partial invariance and/or approximate MI conditions, which we briefly discussed above. In this summary we did not include any studies that examined the capabilities of detection of MI (Kim, Cao, Wang, & Nguyen, 2017; Meade & Lautenschlager, 2004; Yoon & Millsap, 2007; see also Saris, Satorra, & van der Veld, 2009, for methods for the detection of model misspecifications in SEM). In other words, we did not summarize, in our overview, the results of studies which tried to identify which specific measurement parameters prevented reaching measurement invariance. Instead, we focused on those studies that examined under which conditions partial or approximate MI may be sufficient to estimate latent means and associations of latent variables with other theoretical constructs of interest with confidence.

As evidenced in the table, the literature on this topic is very limited. We were able to identify only seven studies which examined the accuracy of latent means estimations under different types of MI. Surprisingly, although PMG-CFA has been applied for the past three decades, the number of simulation studies that focused on the recovery of the latent means under partial invariance is very limited. Most of the previous recommendations were based on educated guesses (Byrne et al., 1989; Reise et al., 1993; Steenkamp & Baumgartner, 1998) rather than on systematic investigations supported by empirical evidence. Only with the recent development of alignment and BSEM methodologies have new simulation studies emerged but still with some substantial gaps:

(1) The efficacy of the alignment method seems to be documented, but only for an exact partial measurement invariance (PMI) condition and not for approximate MI and partial approximate MI conditions;

(2) We have very limited information on how the MG-BSEM model behaves

under different conditions;
(3) A limited number of simulations designed to examine the appropriateness of newly developed methods (BSEM and alignment) have typically been tested only on a small number of groups and/or small sample sizes;
(4) Very little is also known about the recovery of parameters other than means and standard deviations (e.g., regression coefficients) under approximate and partial approximate MI in multigroup modeling.
In the next section, we will try to address these gaps by providing simulation studies that assess the recovery of various parameters of interest under the different conditions of MI described above.

## Simulation Study

### Conditions for Simulations

In this study, we focused on conditions that simulated different measurement (non)invariance scenarios as depicted in Table 2. They entailed 24 groups with 1,500 observations each. These reflect a common lower bound of sample sizes and number of groups spectrum that is widespread in cross-country surveys.[2]

Simulation *conditions* (Bandalos & Gagne, 2014) included *partial exact invariance* (referred to as PMI in Table 2), *partial approximate invariance* (referred to as AMI in Table 2), and *a combination of both* (referred to as PMI + AMI in Table 2). We analyzed conditions with various numbers of items per scale (3 to 5).[3] For partial exact invariance conditions, we considered situations in which 25%, 50%, 75%, and 100% of the compared groups were

2 The ESS requires a sample size of at least 1,500 per country, the ISSP between 1,000 and 1,400, while the WVS targets 1,200 and the Eurobarometer at least 1,000 respondents. More recently, some specific surveys have presented larger sample sizes as is the case with PISA requiring a minimum of 5,400 respondents per country, the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) requiring a minimum of 4,000, or the World Mental Health Initiative Survey (WMHIS) requiring an average of 4,000 respondents. Also, the number of groups in international surveys varies. Depending on the wave, the Eurobarometer studies vary in size between 13 and 39 countries, the ESS varies between 22 to 31 countries, the ISSP between 7 and 37 countries, while large-scale educational assessments like the PISA survey include up to 72 countries. However, one should keep in mind that if the focus of research is on comparisons both across countries and time points simultaneously, the number of groups may quickly increase to 90 and more (Marsh et al., 2017; Zercher et al., 2015).

3 In most surveys (especially sociologically oriented ones), the number of items per scale is rather limited. For instance, in PIAAC, apart from cognitive testing, 18 noncognitive scales were included. The number of items per

affected by noninvariance. We also considered conditions that varied the number of noninvarint items per group between 2 and 4. The number of noninvariant items in the scale ranged from 1 to *n*-1, where *n* represented the number of items in a particular scale.[4] For the approximate MI conditions, we allowed wiggle room for the parameter differences using 0.001, 0.005, 0.010, and 0.050 as variances (see Table 2).

All models were estimated without and with criterion variables to assess the recovery of both latent means and path coefficients of the SEM model.[5] While examining the recovery of the latent mean, data were simulated using the measurement model only. For examining the recovery of the path coefficients, the true model contained the measurement as well the criterion variables.

**Data Generating Procedures**

Data were generated using a CFA model for continuous data with a 5-point Likert scale for each item, because it is one of the most commonly used type of scales in survey research (Leung, 2011). We did this by discretizing continuous indicators into five categories that mimic a 5-point Likert scale. By doing so, we wanted to apply our simulations to the common practice in

scale varies in the PIAAC between 2 and 7, with an average of 4.6 items per scale.

4 It should be noted that, in practice, a situation where more than half of the items in a scale is noninvariant is rather extreme. In real settings, it is difficult to detect noninvariant items if the share of such items is larger than half and even more difficult to detect if the sign of the bias for all noninvariant items is the same. However, there are situations in which such a condition is worth examining, for instance, in the case of planned missing designs (Graham, Taylor, Olchowski, & Cumsille, 2006; Pokropek, 2011) where some of the items are missing or noninvariant by design and other items are considered as invariant. In this case one may quickly run into a problem if the items which are not missing and assumed to be invariant are in fact partially or fully noninvariant. A good example of this situation is the index of home possession included in the PISA Study, where some of the items are purposely noninvariant to reflect the country-specific level of household conditions (OECD, 2014). Such designs in real surveys provide us with the rationale for studying conditions with more than half of the items being noninvariant.

5 We consider path coefficients between our latent variable and criterion manifest variables as a special case of the estimation of path coefficients between latent variables and another set of latent variables. We preferred using criterion manifest variables rather than another set of latent variables to keep the already quite complex simulations as simple as possible.

large-scale surveys of treating 5-point Likert scales as continuous regardless of the fact that the data are collected as ordered-categorical information. This common simplification practice was justified in various previous studies showing that when the number of categories is at least four, parameters estimated using maximum likelihood are accurate (Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Johnson & Creech, 1983; Muthén & Kaplan, 1985).

The data for the simulations were generated in a five-step procedure. Although we are presenting a large number of simulation conditions, the core of the data generation procedure does not change from one condition to another. The overview of the procedure is presented in Figure S1 (in online supplemental material) and described below.

Step 1: Generating latent variables and criterion variables. In the first step, we first generated variables reflecting the latent trait (F) in each group and two criterion variables ($C_1$ and $C_2$) that were used later to assess the performance of the models.[6] Generation of random variables representing the *true latent trait* for each group consisted of two phases. First, we sampled means and standard deviations for each group from normal distributions $N(0,0.3)$ and $N(1,0.1)$, respectively.[7] This was done for all but the first group where the mean was set to zero and the standard deviation to 1 (Little, Slegers, & Card, 2006). This procedure resulted in a realistic condition in which means and variances of the latent traits differ across countries. Second, the C variables were generated from standard normal distributions in such a way that the regression coefficient $C_1 \rightarrow F$ was set to 0.3 in each group and the regression coefficient $C_2 \rightarrow F$ was set to 0.1 in each group. In Figure 1, both Fs and Cs are depicted by rectangles as they were generated and directly observed.

Step 2: Generating item parameters. In the second step, we generated parameters for each item measuring the latent trait F. The procedure of generating item parameters was as follows. First, item parameters, that is, factor loadings ($\lambda_i$) and intercepts ($\tau_i$), were sampled. Factor loadings were sampled from a uniform distribution bounded by 0.65 and 0.85, and intercepts from a standard normal distribution with a mean of zero and a standard deviation of 0.5. The error term $\epsilon_i$ was set to $\left|1-\lambda_i^2\right|$. For our simulation settings, these parameters resulted in scales with relatively high reliabilities. Cronbach's alphas were around 0.80 for scales with 3 items, 0.85 for scales with 4 items, and 0.87 for scales with 5 items.

Step 3: Generating data where MI holds. Using parameters from Step 2, we generated data that fulfilled the assumption of exact MI. Factor indicators

6 These criterion manifest variables should not be confused with the latent variable manifest items.

7 The distributions we used were chosen after first examining distributions and cross-country differences of latent means obtained from MG-CFA models measuring political trust, openness to experience, social engagement, and attitudes toward immigrants in the 7th ESS round from 2014-2015.

were randomly generated with the sampled item parameters. Finally, the continuous factor indicators that were produced were discretized into five categories using the threshold values –1.30, –0.47, 0.47, 1.30 (a similar approach using these thresholds was presented in Sass, Schmitt, & Marsh, 2014). The rationale for using these thresholds was to obtain distributions of categorical indicators that were approximately normal. In such situations, models for continuous data may be applied to categorical data as a reasonable simplification given the large sample size (defined as 1,500 in our case).

Step 4: Adding noninvariance bias. In the fourth step, we added a noninvariance bias in specific affected groups. First, items were independently sampled from each group. The number of sampled items per group differed for different simulation conditions. In each replication, a random assignment of noninvariance was repeated so that no particular pattern of noninvariance was present. Next, the bias was added to those selected items. When an item was selected, we added a bias of +0.2 or -0.2 for both factor loadings and intercepts so that noninvariant items in our study were *always* noninvariant (and invariant items were always invariant) in respect to *both* the factor loadings and the intercepts. The sign of the bias was determined randomly, independently for each item and for each type item parameter. After adding bias, the error terms were updated using the rule described in step 2. These biases reflected conditions of medium item noninvariance with a random direction (Saris et al., 2009).

In AMI conditions, bias was added to all parameters of all items (factor loadings and intercepts), except in the PMI + AMI condition, using random draws from a standard normal distribution with a mean of zero and a variance that was dependent on the simulation conditions. For instance, for simulating AMI at the 0.05 level (i.e., a situation where the distribution of the differences between parameters had a mean of zero and a variance of 0.05), the bias for each item was drawn from a standard normal distribution with a mean of zero and a variance of 0.025 (as the variance of the differences of two random variables equals the sum of the variance of the two random variables, assuming a covariance of zero between their error terms). For conditions that combined PMI and AMI, first the PMI bias was generated with items that are allowed to be different across groups and freely estimated, and then AMI bias was applied to the rest of the items.

Step 5: Generating data with noninvariant items. In the last step, data were generated using parameters obtained from step 4. By doing so, we ended up with a data structure that was affected by noninvariance but otherwise with the same data characteristics as those of the data structure generated in step 3 where MI holds. Similar to step 3, a CFA was used for generating continuous data. The continuous factor indicators were discretized into five categories with threshold values of –1.30, –0.47, 0.47, and 1.30. Next, based on these data, different CFA models were run to test for partial, exact, and approximate invariance, and their performance was investigated.

For the data generation process we used the software package Mplus 7.4 (Muthén & Muthén, 2015) combined with a self-written program in R. All model parameters were sampled in R, while data for each replication were generated using Mplus 7.4.

## Estimation Procedures

We estimated all models using Mplus 7.4. In all tested models we fixed the latent mean to zero and its standard deviation to 1 in the first group. Identification constraints of the partial invariance models required that (a) the parameters (factor loading and intercept) of at least one item were constrained to be equal across at least two groups, and (b) all groups were linked together with equality constrains. Such identification strategies are common in educational testing under the name of multiple matrix designs (Gonzalez & Rutkowski 2010). For simple example, with three groups and two items, the parameters of the first item in group 1 are constrained to be equal to those in group 2. The parameters of the second item in group 2 are constrained to be equal to those in group 3. Although groups 1 and 3 are not directly linked with constraints, they are linked via group 2. This identification strategy was more likely to reflect real data than a strategy which assumes a single item whose parameters are constrained to be equal across all groups. However, our identification strategy did not exclude such a situation in certain conditions in which there may have been one or more items with equal parameters across all groups.

For MG-CFA, PMG-CFA, and AMG-CFA, we used maximum likelihood estimation with robust standard errors (MLR) with default Mplus settings, that is, a maximum of 1,000 iterations and a convergence criterion of 0.00005 for MG-CFA and PMG-CFA. For AMG-CFA, MLR estimation with the EM algorithm was used, with a maximum of 500 iterations and a convergence of 0.000001. For the BSEM models we applied Bayesian estimation using Gibbs sampling with a convergence criterion of 0.05, a minimum of 5,000 iterations, and a maximum of 200,000 iterations using two chains (the default Mplus settings with an increased maximum

number of iterations). For each estimation we used starting values that reflected true parameter values from the data generating phases. This is a common practice used in most simulation studies which allows speeding up the estimation procedure by lowering the number of iterations that are necessary to achieve convergence while not influencing the results (see, e.g., Muthén & Muthén, 2002).

The estimation of regression coefficients was performed in all methods (except for AMG-CFA) in one step with the measurement parameters simultaneously. As alignment does not allow including path coefficients, we used, for alignment, a two-step procedure. In the first step, we estimated a measurement model using AMG-CFA. In the second step, we estimated the SEM model where the parameters for the measurement model were fixed and equal to the parameters from the first step, while the regression coefficients were freely estimated. The approach is similar to the strategy proposed by Marsh et al. (2017). The only difference is that in the Marsh et al. (2017) approach, the largest noninvariant parameters are reestimated as partially noninvariant parameters in a multigroup SEM model, whereas we keep all the parameters from the alignment model fixed. In other words, in step 2 we do not fix any parameter to be exactly equal across groups, but simply use the aligned parameters from step 1 as fixed parameters.

## Performance Measures of Parameter Recovery

While examining the results of the simulations under different conditions of PMI, we focused on two questions: First, to what extent were the models able to recover the true latent means and provide consistent rankings of the groups; second, to what extent were the models able to accurately recover the path coefficient between the latent variable and the two criterion variables.

To answer the first question, we used three statistics:

(1) According to the recommendation of Muthén and Asparouhov (2013, 2014), a correlation of at least 0.98 (and preferably 0.99) between the true mean values and their estimates indicates a reasonably good recovery of the mean rankings. For simplicity when reporting these correlations, we refer to the term *mean correlations*. Correlations reported in the following tables which are higher than 0.98 are indicated in bold. Such a correlation in the

table indicates whether we could expect to recover the true overall mean ranking using a particular method in a particular condition[8].
(2) A good model would provide not only a reasonable recovery of the mean ranking but also a reasonably precise point estimate of the latent means. The overall accuracy of the parameter estimates referring to the group mean was measured by the root mean square error (RMSE). RMSE is defined as:

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\theta_r - \theta\right)^2} \quad (2)$$

where $\theta_r$ is the estimated parameter for the replication $r$, $\theta$ is the true value of the estimated parameter, and $R$ is the number of replications, which was set for all conditions to 400[9]. The interpretation of RMSE values is straightforward when compared to each other. However, providing practical recommendations on how large RMSE may be is less easy. RMSE values are presented in the same metric as our investigated parameters. For instance, an RMSE value of 0.06 may be interpreted as the average absolute difference between the true and the estimated mean parameter. As the standard deviation of the true latent means in all of our simulations was set to 0.3, an RMSE value of 0.06 implies that the average absolute difference between true and estimated means was as large as 20% of the standard deviation of the true means.

(3) Finally, we examined whether the interval estimation would provide correct inferences. The interval estimations are directly related to the standard errors of latent means computed by each tested approach. Instead of reporting direct information on the recovery of standard errors, we focused here on the more intuitive information of how many times the 95% intervals of the estimated latent means contained true values of the parameters. This was assessed by the coverage of the true means with a 95% coefficient interval (CI) generated using standard errors of the estimated means. In the tables below we indicated a good mean recovery, ranging between 0.9 and 1.0, in bold.

---

8 Statistics that summarize simulations are based only on estimated parameters in 23 out of the 24 groups (means and standard deviation of the first group were always fixed). For computing the correlations of mean rankings, RMSE, and CI95%, we performed computations for each replication, and then we averaged out the results across all replications.

9 Although the number 400 may not seem to be very large, various successful past simulation studies dealing with complex models used even smaller numbers of replications. For instance, 100 replications per condition were used in Nylund, Asparouhov, and Muthén (2007) and in Meade and Lautenschlager (2004). Recently, Kim et al. (2017) used 100 replications per condition to analyze some similar models to those in our study.

For answering the second question—to what extent we were able to recover the relation between the latent variable and two criterion variables—we used SEM models where two of our criterion variables were set as dependent variables and the latent factor as a predictor variable (see Figure 1). The true values of regression coefficients in the data generating procedure were set at 0.1 and 0.3, respectively. As the implications of the results for the two path coefficients were virtually the same, in the results section we report only the findings for the recovery of path coefficients with a true value equal to 0.3 (the findings related to the recovery of the path coefficients with a true value equal to 0.1 can be obtained from the first author upon request). We examined the recovery of the regression coefficients using three statistics: the average path coefficient estimate across groups and replications, the RMSE, and the 95% CI coverage of the regression coefficient.

(1) The average path coefficient revealed the biases of estimates. To simplify the interpretation of this result, we set an arbitrary, but substantively reasonable and commonly used threshold (see Hoogland & Boomsma, 1998; Kaplan, 1988) of 5% for a maximum bias tolerated by us. In other words, whenever the average bias of the estimate of the unstandardized coefficient for a given model and condition was higher or lower than 5%, we flagged it as a substantial bias.

(2) RMSE provided information on the overall accuracy of estimation. Similar to the assessment of the latent means recovery, we used the threshold of 0.06. As RMSE is defined using the metric of the estimated parameter, in the case of recovery of path coefficients it may be interpreted as the average difference between estimated and true values of the path coefficient within a certain condition. In other words, an RMSE value of 0.06 for a true path coefficient of 0.3 reveals that the average error of estimation for this parameter was 20% ([0.06/0.3]*100).

(3) Finally, similar to the latent means interval, estimation of unstandardized path coefficients was assessed by the coverage of the true unstandardized path coefficients with a 95% CI generated using standard errors of the estimated models with a threshold between 0.9 and 1.0.

While examining the recovery of unstandardized path coefficients, we are indirectly investigating the recovery accuracy of standard deviations of the latent factor means. After all, a recovery of unstandardized path coefficients depends heavily on a correct recovery of the standard deviation of the latent factor means. We decided to examine path coefficients rather than standard deviations of latent factors as the latter are not of direct interest for substantive researchers in most practical situations. Substantive research tends to focus on testing hypotheses related to path coefficients.

As often the case in simulation studies, our quality criteria for the recovery of regression coefficients are, above all, bias and efficiency (Bandalos & Gagne, 2014). Other researchers might have chosen more or less strict criteria. Indeed, in many social science applications, only the signs and the significance of estimates of associations between variables might be of interest. On the other hand, if one would like to rank groups based on

unstandardized regression coefficients (e.g., when using indexes of inequality in many large-scale assessments studies like PISA), higher precision needs to be achieved. We believe that our selection of quality criteria is well suited for typical studies using large-scale cross-country data.[10]

## Results

Tables 3 to 8 present the results on the recovery of latent means and path coefficients for the simulation conditions PMI (Tables 3 and 4), AMI (Tables 5 and 6), and PMI + AMI (Tables 7 and 8). Mean correlations (> 0.98 in bold), RMSE (> 0.06 in bold), and 95% CI coverage of latent mean estimations (> 0.9 in bold) were reported. As mentioned earlier, the simulation conditions for PMI and PMI + AMI were defined by:

1. the percent of affected groups, that is, the share of groups in which noninvariance was simulated,
2. the number of items per scale, and
3. the number of noninvariant items in the affected group (third column).

The simulation conditions for AMI on data which were generated to be approximately measurement invariant were defined by

a. the number of items per scale, and
b. the variability of the priors.

As mentioned earlier, the simulation study for PMI was designed in such a way that a dataset without noninvariant items was generated first. Based on this data, the classical CFA model was estimated and was used as a reference model (Model 1 MG-CFA [MI] in Tables 3 and 4). This model provided us a reference for the mean correlations, RMSE, and CI values for a given scale length for the condition under which all items were invariant. The results in the next five columns, MG-CFA, PMG-CFA, MG-BSEM, PMG-BSEM, and AMG-CFA (Models 2-6), refer to conditions where noninvariance was introduced into the model according to the strategy outlined earlier. The MG-CFA column provides the results for a model where noninvariant items were simply ignored and all item parameters were constrained to be equal across all groups using a scalar invariance model. The PMG-CFA column provides information for a model where all the noninvariant item parameters were allowed to be freely estimated and all equality constraints on these noninvariant items were relaxed. This is an idealized situation because when real survey data are analyzed, information about the noninvariance of items is not given in advance, and it needs to be detected by the researcher. Such a detection of noninvariance is prone to additional errors that might influence the accuracy of the estimation; therefore, results presented here should be treated as an upper bound that is possible but very difficult to achieve in survey research. The MG-BSEM column provides the statistics on recovery of means for a model where information about noninvariant items

10 Researchers who need to rely on more or less accurate parameter estimates in their cross-group studies should interpret our findings in the simulations accordingly, that is, as too liberal or too strict for their purposes.

was ignored and approximate MI was imposed with small prior variances defining differences between item parameters (both factor loadings and intercepts) at 0.001. The PMG-BSEM model refers to a model where approximate MI was imposed with small prior variances defining differences between item parameters (factor loadings and intercepts) at 0.001. However, in contrast to the MG-BSEM model, parameters for noninvariant items were freely estimated using noninformative priors. The AMG-CFA model presents results for the multigroup alignment optimization model (using the fixed version). Below we discuss the results in each of these simulations. The AMI conditions were examined for only three models: MG-CFA, MG-BSEM, and AMG-CFA. They were estimated only for these three models because the other two models, PMG-CFA and PMG-BSEM, require a free estimation of noninvariant parameters, which is not part of the modeling strategy in AMI.

**PMI**

Recovery of the means. Table 3 presents the results of the simulation conditions for PMI. Analyzing the results of the reference model (Model 1: MG-CFA [MI]) that was estimated on data with full invariance, we concluded that only the 5-item scale provided accurate estimates of means that successfully recovered the group rankings across all simulation conditions under the situation of imposing full MI. Although the correlation between true means and their estimates for all scales was higher than 0.99 and the RMSE measures lower than 0.06 for the 5- and 4-item scales, coverage of the estimates located between the 0.9 and 1.0 interval was achieved only for the 5-item scale. When we treat categorical indicators as approximation of continuous data, as commonly done in research practice, models with less than five items per scale lack sufficient information to correctly estimate standard errors.

The MG-CFA (Model 2) and the MG-BSEM (Model 4) models do not allow modeling the PMI that is present in the data. We examined them to see how such a misspecification influences the parameter estimates. It appears that these models do not fulfill any of the three criteria under all of the conditions. The statistics in these two models demonstrate that ignoring partial MI and imposing full scalar exact or approximate MI in this case might cause substantial biases in the recovery of latent means in cross-group analyses. The mean correlations were similar for both models. MG-BSEM performed worse in terms of RMSE and the 95% CI coverage, showing high sensitivity to item noninvariance misspecification. MG-BSEM models were inclined to stretch the latent mean scale (i.e., the estimated group means at the lower end were underestimated and the estimated group means at the upper end of the distribution were overestimated) when partial invariance was present but ignored. Fortunately, in this method, model fit based on posterior predictive p-values (PPP) and 5% CI for the difference in the chi-square statistic for the observed and simulated data (Muthen & Asparouhov, 2012) proved to detect misfit with power over 95% for all conditions presented in Table 2.

The PMG-CFA (Model 3) and PMG-BSEM models provided the most accurate results although all three criteria were satisfied only for the 5-item scale. Interestingly, even when 75% of the groups were affected by partial noninvariance and four noninvariant items were present per scale or when 100% of the groups were affected and three noninvariant items were present, the two exact and approximate partial invariance models reached a good recovery of the latent means as demonstrated in the high correlations between the true and estimated means, the low RMSE, and the reasonable 95% CI coverage.

Whereas the exact and approximate partial invariance models performed best in the various conditions, the alignment method (Model 6) was the third best performing method and only slightly better than that of the MG-CFA model. However, it should be noted that alignment was the only method in our comparison that automatically detected noninvariant items and accommodated the model accordingly. Thus, when analyzing survey data, this method may outperform other methods in the unbiased estimation of latent means, particularly if the detection of noninvariant items using other approaches lacks power and/or accuracy. The results are also in agreement with previous studies postulating that the alignment optimization procedure performs well unless there is a majority of noninvariant parameters (Muthén & Asparouhov, 2013). Asparouhov and Muthén (2013) and Flake and McCoach (2018) suggested that alignment performs well when not more than 25% of the parameters are noninvariant. As Table 3 demonstrates, in typical cross-country survey analyses with one noninvariant item per group and where up to 50% of the groups are affected by noninvariance, alignment will provide accurate estimations. However, the table also suggests that when noninvariance patterns are more severe, even alignment would not always produce reliable estimates.

Recovery of the path coefficients. Next, we turn to the SEM models used to examine the recovery of the path coefficients. Table 4 lists the statistics on the quality of the recovery of the path coefficients that were modeled from the latent variable to one of the criterion variables (with unstandardized paths constrained to 0.3 in the data generation model). Observing the results presented in Table 4, it becomes evident that all models except for Model 4 (the MG-BSEM) can produce, under certain conditions, precise estimates of the path coefficient. The bias in the MG-BSEM model was strongest when the number of noninvariant items and the number of groups affected were largest. The other models differed in the overall accuracy of the recovery of the path coefficient as measured by the RMSE and the CI. The average value of RMSE for data with perfect MI used in Model 1 was as small as 0.007. PMG-CFA produced a very similar average RMSE, while the other models performed worse but still very well in terms of RMSE under different conditions. The most significant differences in the performance of the models emerged in the CI estimations. Model 3 (PMG-CFA) provided an accurate 95% CI coverage across practically all conditions. Second best was Model 5 (PMG-BSEM). Model 4 (MG-BSEM) performed worst

in terms of the CI, thus displaying a high sensitivity to model misspecification. Finally, Model 6 (AMG-CFA) was not as effective as Models 3 (PMG-CFA) or 5 (PMG-BSEM), but it still provided a reasonably good recovery of path coefficients in terms of the CI, even when a substantial number of items was noninvariant, under the condition that the number of groups affected by noninvariance was rather low (25%).

**AMI**

Recovery of the means. In the previous section we tested different models that reflected the PMI situation. Table 5 presents the quality of recovery of latent means from simulations where data were generated to be approximately measurement invariant (see Table 2).

Only the 5-item scale could fulfill the three benchmarks we set for this analysis when the variability was as low as 0.001. In most conditions neither the 3- nor the 4-item scales provided reasonable statistics for all three criteria, although they did provide high mean correlations in several cases. However, it should be noted that with such a low level of AMI as 0.001 that we applied for the 5-item scale, there is no need to use complex methods such as AMI, because classical MG-CFA appears to be sufficiently robust, as evidenced in Table 5. It is worth noting that the MG-BSEM model performed particularly well in terms of the interval estimation. For a 5-item scale, it gave an acceptable 95% CI coverage for all tested levels of AMI, while other methods substantially underperformed in this aspect. As Table 5 reveals, the MG-BSEM approach did not increase the precision of estimates beyond the other two methods, as it performed similarly in terms of the mean correlations. It was less precise in terms of the RMSE. However, it resulted in good interval estimations (that were closest to the desired 95% level), outperforming the other methods in this respect. [11]

Recovery of the path coefficients. Table 6 presents the recovery of path coefficients where data were generated according to the AMI conditions. Overall, the presence of AMI does not substantially bias the estimates of the path coefficients (unless the AMI was larger than 0.010). In such conditions, ignoring AMI and applying simple MG-CFA could provide reasonably accurate results. When approximate invariance was largest (AMI = 0.05), AMG-CFA performed best, and somewhat better than MG-CFA, with slight upward bias in terms of recovering point estimates. The MG-BSEM method,[12] on the other hand, produced most accurate standard errors but slightly downward biased path coefficient estimates (bias increased with the increase of AMI). To sum

11 For the MG-BSEM condition, we often experienced convergence problems. For the 3-item scale and AMI = 0.01, 98% of the models converged, while for AMI = 0.05, none of the estimations was successful (indicated by "NA" in the table). For the 4-item scale and AMI = 0.01, approximately 84% of the models converged, while for AMI = 0.05, 60% of the estimations reached convergence. For the 5-item scale and AMI = 0.01 or 0.05, 70% of the models converged. For all conditions which are not mentioned, convergence rate was 100%.

up, low to moderate levels of AMI did not substantially bias the estimation of regression coefficients in multigroup settings in any model, but they affected the efficiency of the estimation and the correctness of standard errors. For moderate levels of AMI (0.005-0.010), all approaches provided accurate point estimates, but only BSEM provided a reasonably good interval estimation (especially for longer scales).

**PMI + AMI**

Recovery of the means. In this section, we present results of the simulations for conditions that combined PMI and AMI. In other words, in the simulated conditions we allowed for large differences in some parameters (i.e., no MI for these parameters) and small differences in others parameters (i.e., approximate invariance for these parameters). Table 7 reports the statistics for the recovery of group means, and Table 8 reports the statistics for the recovery of path coefficients. Both tables refer to simulations involving 5-item scales. Since none of the models with three and four items could recover the correct means and most of the path coefficients correctly, we focus here on 5-item scales. In these simulations we focused on two levels of variability for the parameters in the AMI model: 0.005 and 0.010. Higher levels of AMI did not result in a reasonable recovery of latent means and path coefficients, and therefore we omitted these conditions from further investigation. We also omitted conditions with AMI as small as 0.001, because such AMI levels did not introduce significant bias that could alter the results, and all models behaved practically as if there was no parameter variance. Thus, we focused on AMI levels of 0.005 and 0.010.[13]

As Table 7 demonstrates, the MG-CFA model performed quite poorly. The statistics in the second PMG-CFA model improved considerably in terms of the mean correlations and RMSE, providing relatively decent recovery of the rankings, at least for AMI = 0.005 and when not more than 50% of the groups were affected. However, even in relatively favorable conditions (a small number of noninvariant items and AMI = 0.005), the 95% CI coverage hardly exceeded 80%.

Results in the third MG-BSEM model were similar to those in the MG-CFA model. However, the fourth PMG-BSEM model provided better interval coverage, although it was still below the desired 95% average level. The recovery was reasonable when no more than 50% of the groups were affected, AMI was as small as 0.005, and not more than three items were noninvariant. When AMI was larger (0.01) and 50% of the groups were affected, it was possible to recover the means rather well when only one

12 Adding the structural part to the measurement model resolved the convergence problems of the BSEM model. The convergence rate was then higher than 90% even in the most demanding situations.

13 We do not present statistics for the condition with four biased items with 100% of the groups being affected, because this model was not identified.

item was noninvariant.

The alignment approach provided a reasonable recovery of the rankings for quite limited conditions—when 75% (or less) of the groups were affected, one item was noninvariant, and AMI was as small as 0.005—although the 95% interval coverage was similar to that in the first MG-CFA model, where the problem of MI was simply ignored.

In summary, when both PMI and at least a moderate AMI were present, it was more difficult to recover latent group means. The classical MG-CFA approach produced rather inaccurate point estimates with very inaccurate interval estimations. The alignment method was not much help as, similar to PMG-CFA, it was not able to produce correct interval estimations. In a similar vein, even though MG-BSEM accounted for AMI, it was very sensitive to PMI thus producing rather imprecise estimates, especially when more than one item was noninvariant. PMG-BSEM was able to produce reasonable mean rankings when AMI = 0.005, even with the presence of considerable partial noninvariance. PMG-BSEM produced reasonable mean rankings with a higher degree of AMI only when partial noninvariance was limited to a small number of affected groups and items.

Recovery of the path coefficients. Finally, Table 8 presents the statistics for the recovery of path coefficients when both PMI and MI were applied. The first, second, and fifth models, MG-CFA, PMG-CFA, and AMG-CFA, produced rather accurate point estimates; however, these models failed to provide correct interval estimations. BSEM models, on the other hand, produced more bias. Especially the MG-BSEM model resulted in a large downward bias in the estimation of the path coefficient. With a large number of noninvariant items (more than two), the average estimate of the path coefficient for the MG-BSEM model was around 0.25 (compared to the true value of 0.3). The PMG-BSEM resulted in an upward bias. On the other hand, PMG-BSEM models provided a good interval estimation close to 95% under many conditions, including those with a large amount of noninvariance.

## Summary and Conclusions

In this article we aimed at testing, under various conditions, whether PMI, AMI, and the combination of both (partial approximate MI, i.e., PMI + AMI) may be sufficient for meaningful comparisons of latent means and regression coefficients using select state-of-the-art methods.

We designed and conducted a large-scale Monte Carlo simulation study that explored 156 conditions (78 for MG-CFA and 78 for multigroup SEM) using five different strategies of estimation. We selected conditions that simulated realistic situations encountered when performing secondary data analysis using international surveys. This led us to examine results of simulations for 804 conditions (see Table 2) that gave rise to more than 300,000 individual estimations. We examined the performance of these models and tried to provide useful guidance for applied researchers facing partial and approximate MI.

The simulation study that we performed showed that even large deviations from strict MI may allow precise estimations and meaningful comparisons of

both means and path coefficients. Using appropriate models, both group rankings as well regression coefficients of SEM models could be correctly recovered under specific conditions.

Our results showed that for a proper estimation of group means, the number of items per scale constituted a crucial factor. Specifically, 3- and 4-item scales provided too little information to fully recover the group mean rankings with acceptable precision, especially under unfavorable conditions. However, the number of scale items was not as crucial for the correct recovery of regression coefficients, although it did play an important role. Under the assumption that the noninvariant items were known, *partial invariance* may be handled very effectively using partial invariant MG-CFA. Indeed, a good recovery of factor means with partial invariance models is expected under partial invariance because the partial invariance model is a data generating model for the PMI conditions. However, it is worth noting that partial invariance models were robust also in conditions with a large number of free, noninvariant parameters. Our simulations showed that correctly specifying partial invariance models with a large number of noninvariant items (by identifying the noninvariant items and freeing their equality constraints) can after all provide results which are almost as good as those of MG-CFA models under conditions of full invariance. In other words, the simulations demonstrated that even when 80% of the total item pool was noninvariant, partial invariant MG-CFA models gave a good recovery of both path coefficients and latent means with a high correlation between true and estimated means, a small RMSE, and a reasonable 95% CI coverage for the 5-item scale. In our conditions, it was sufficient to have only one anchored (invariant) item to provide results that were very close to the true scores. Practical advice for researchers dealing with partial invariance modeling is, therefore, to maximize the detection rate of noninvariant items (risking even a false detection) and free the parameters of noninvariant items. Thus, PMG-CFA and PMG-BSEM models performed best but required the additional step of detecting the noninvariant items. The alignment procedure was the only approach tested by us that automatically detected noninvariance. This approach, however, required considerably fewer noninvariant items in our conditions (no more than 20% in the total item pool) to provide reasonably accurate parameter estimates.

For AMI we tested four levels of variation: 0.001, 0.005, 0.010, and 0.050 (both for factor loadings and intercepts). The lowest level (0.001) did not result in biased estimates and, in practical terms, such a small level of approximate invariance will not affect the results of models that were not designed to deal with AMI. In other words, MG-CFA could also recover the means satisfactorily under such a condition. Analysis using the highest variability of 0.05 in the data showed that recovery of mean ranking and latent means was very difficult to meet in our conditions, and also the recovery of path coefficients in the structural part of the model did not satisfy all three adopted criteria (mostly interval estimation) using any of the methods. However, it should also be noted that whereas the statistical

criteria were not satisfied under these liberal AMI conditions, substantive findings may still be meaningful and robust from a social scientific point of view, given that many substantive studies in the social sciences are interested only in the sign of associations rather than in their precise size or ranking (whereas for mean estimations, substantive researchers are often interested also in the ranking). With the moderate AMI levels in the data of 0.005–0.010, all approaches delivered accurate point estimates, but only BSEM provided a good estimation of standard errors although it was characterized by some convergence problems. Similar results were obtained in the simulations to recover regression coefficients.

Finally, the hardest situation to cope with from a practical point of view was the condition which combined PMI with AMI. In this situation, a correct estimation of latent means and an accurate recovery of the mean rankings was very difficult to achieve, and it was practically applicable only to the PMG-BSEM model that included at least five items, under the conditions of AMI = 0.005 combined with up to 40% noninvariant items, or AMI = 0.010 combined with up to 20% noninvariant items. As in previous models, partial approximate invariance was nonetheless sufficient to recover regression coefficients under these various conditions.

Although we are presenting one of the largest (if not the largest) simulation study on parameter recovery in multigroup modeling under different types of measurement noninvariance, it should be noted that the study still suffers from the absence of certain relevant conditions. We chose conditions that may typically apply to large-scale comparative research orienting on some common international surveys (such as the ESS or the WVS), considering that this type of analysis is currently the leading edge of modern cross-cultural research, but has been severely underrepresented in previous simulation research thus far. We did not consider conditions that are characterized by the largest international surveys (such as PISA or PIRLS) with even larger sample sizes and an even larger number of groups. However, the conditions employed in our research may still provide valid guidance also when analyzing larger sample sizes and a larger number of groups than those included in our simulations.

We limited our simulations to a single size of bias for partial MI with a random sign (positive or negative). Although one could choose a different bias size, we believe that we chose a realistic noninvariance bias. Indeed, robustness checks suggested that lowering the sample size to 1,000 and/or increasing the partial noninvariance bias from 0.2 to 0.3 essentially did not alter the results. For generating data, we used highly reliable scales with factor loadings between 0.65 and 0.85—in real survey data, factor loadings may range between 0.35 and 1.0. Future research should address these and other conditions not considered in the current study, in order to better understand the hazards and possibilities of cross-group analysis.

Furthermore, convergence rates for BSEM models with relatively high priors (0.01 and 0.05) were low. However, we performed robustness checks

analyzing different convergence criteria,[14] which provided virtually the same conclusions as those presented here. With more strict convergence criteria we achieved better convergence rates but the computational time grew exponentially, stretching the time of model estimation into weeks. Developing faster and more reliable algorithms in terms of convergence may be an important task for future research.

Finally, it should be noted that measurement invariance testing is a necessary but not a sufficient condition for comparability. It may well be the case that the scales under investigation display equal parameters but bear a different meaning across groups, so that comparisons between groups such as countries or cultures may never be ideal apples to apples comparisons.[15] In sum, SEM path coefficients are relatively robust to violations of MI and can be rather effectively recovered. This is not surprising given that regression coefficients between latent variables and other theoretical constructs of interest are not influenced by intercept noninvariance of continuous indicators (Widaman & Reise, 1997). However, recovering mean group rankings seems to be more difficult, and researchers need to consider the data at hand. PMI models may be rather effective to recover both path coefficients and latent means when many or even most items are noninvariant. Approximate MI models are appropriate to recover latent means when many parameters are not exactly equal but are approximately equal. They have the advantage of providing accurate standard errors. Finally, the alignment procedure is recommended for recovering latent means in cases where there are only few noninvariant parameters.

---

14 When convergence rate was particularly low in three of 804 conditions, we changed the convergence criteria in Mplus (the so-called proportional scale reduction, PSR) from 0.05 to 0.01 and the minimum number of iterations from 200,000 to 2,000,000 while estimating the models using 30 additional replications.

15 We would also like to remind readers here that while our study examined the recovery of latent means and regression coefficients, the findings of the study are not applicable for studies trying to recover *observed* group means (see Millsap, 2011).

References

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

Asparouhov, T., & Muthén, B. (2017). Prior-posterior predictive p-values. Mplus Web Notes: No. 22. Version, 2. April 27, 2017. Retrieved from http://www.statmodel.com/examples/webnotes/webnote22.pdf

Bandalos, D. L., & Gagne B. (2014). Simulation models in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92-108). New York: Guilford Press.

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. doi:10.1207/s15328007sem1302_2

Brown, T.A. 2015. *Confirmatory factor analysis for applied research.* New York: Guilford Press.

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*(3), 456-466. doi:10.1037/0033-2909.105.3.456

Chiorri, C., Day, T., & Malmberg, L. E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Frontiers in Psychology, 5*, 983. doi:10.3389/fpsyg.2014.00983

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*,

982. https://doi.org/10.3389/fpsyg.2014.00982

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

De Beuckelaer, A., & Swinnen, G. (2018). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Methods and applications in cross-cultural analysis (2nd ed.)* (pp. 127-156). New York: Routledge.

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

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*(2), 309-326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

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

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, *3*, 125-156.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323-343. doi:10.1037/1082-989X.11.4.323

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research, 26(3),* 329–367. https://doi.org/10.1177/0049124198026003003

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48(3),* 398-407. Retrieved from https://www.jstor.org/stable/2095231

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426. https://doi.org/10.1007/BF02291366

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23(1),* 69–86. http://dx.doi.org/10.1207/s15327906mbr2301_4

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412-421.

https://doi.org/10.1080/01488376.2011.580697

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*(1), 59-72. https://doi.org/10.1207/s15328007sem1301_3

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2017). 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.* Advance online publication. doi:10.1037/met0000113

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*(1), 60-72. https://doi.org/10.1207/S15328007SEM1101_5

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

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

Munck, I., Barber, C., & Torney-Purta, J. (2017). 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*. https://doi.org/10.1177/0049124117729691

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313-335. doi:10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17 January 11, 2013. Retrieved from

https://www.statmodel.com/examples/webnotes/webnote17.pdf

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

Muthén, B., & Asparouhov, T. (2017). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124117701488

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. http://dx.doi.org/10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2015). Mplus user's guide (7th ed.) Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*(4), 535-569. https://doi.org/10.1080/10705510701575396

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Press.

Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *ASK. Research & Methods,* 20 (1), 81-105. Retrieved from https://kb.osu.edu/dspace/bitstream/handle/1811/69576/ASK_2011_81_105.pdf

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item

response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. doi:10.1037/0033-2909.114.3.552

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

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

Seddig, D., Maskileyson, D., & Davidov, E. (in press). The comparability of measures in the ageism module of the fourth round of the European Social Survey, 2008-2009. *Survey Research Methods.*

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90. https://www.jstor.org/stable/10.1086/209528

Steinmetz, H. (2018). Estimation and comparison of latent means across cultures. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and applications* (2nd ed.) (pp. 95-126). New York: Routledge.

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(1), 1-12. doi:10.1027/1614-2241/a000049

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770.

doi:10.3389/fpsyg.2013.00770

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*, 4-69. doi:10.1177/109442810031002

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10222-009

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14*(3), 435-463. https://doi.org/10.1080/10705510701301677

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology, 6*, 733. https://doi.org/10.3389/fpsyg.2015.00733

Table 1
*Comparison of Existing Simulation Studies on Partial and Approximate MI*

| Study | MI | Models | Design | Main Recommendations |
|---|---|---|---|---|
| Steinmetz (2013) | PMI | PMG-CFA; observed scores | 2 groups; 4 and 6 indicators; sample sizes 100 and 300; intercept DIF = 0.3; loading DIF = 0.2 | Half of the items has to be invariant to achieve meaningful comparisons using PMG-CFA. |
| Muthén & Asparouhov (2013) | PMI | AMG-CFA, PMG-CFA, PMG-BSEM | 10 groups; 6 items; sample size 500; intercept DIF = 0.2; loading DIF = 0.2; BSEM priors: 0.01, 0.05, 0.10 | PMG-CFA provided the most accurate estimates and PMG-BSEM model was the second-best model. |
| Asparouhov & Muthén (2014) | PMI | AMG-CFA (ML estimation) | 2, 3, 15, and 60 groups; 5 indicators; sample size 100 and 1,000; large PI DIF on some indicators (up to 20% noninvariance items) | Estimates were unbiased; a combination of small sample size and a large amount of noninvariance may lead to biased estimates. |
| | PMI | AMG-CFA (Bayes vs ML estimation | 2 groups; sample size 300, 1,000, 2,000, 5,000, and 10,000; 20% noninvariance of items | The Bayes estimator gives slightly more accurate standard errors than the ML estimator. ML standard errors are overestimated for small sample sizes. |
| | PMI | AMG-CFA (ML estimation) | 26 groups; sample size 100, 200, 500, and 2,000; 4 indicators; large PI DIF | Good recovery of measurement parameters as well as factor means and factor variances |
| van de Schoot et al. (2013) | PMI + AMI | MG-CFA, PMG-CFA, MG-BSEM PMG-BSEM | 7 groups; sample size 500; 4 indicators, large PI DIF on some indicators | MG-BSEM works relatively well in an AMI situation but should not be applied when partial MI is present |
| Chiorri et al. (2014) | PMI + AMI | 2D-PMG-BSEM 2D-PMG-CFA | 2 groups; sample size 43; 7 indicators per factor | PMG-BSEM and PMG-CFA models give similar results when partial noninvariance is present in the data. |
| Asparouhov & Muthén (2014) | PMI | AMG-CFA | 2, 3, 15, and 60 groups; sample size 100 and 1,000 per group; extent of noninvariance 0%, 10%, 20% items. | With fewer than 25% noninvariant items AMG-CFA can provide a good recovery of factor means and factor variances. |
| Flake & McCoach (2018) | PMI | 2D-AMG-CFA (MLR estimation for categorical data) | 3,9, and 15 groups; sample size 500; 7 indicators per factor; 2 correlated factors; 0%, 14%, 29%, 43% of the items with noninvariant loading (with a bias of 0.1, 0.25, and 0.4) or threshold (with a bias of 0.2, 0.5, and 0.8); 4 categories per item. | Across all of the different estimates there were generally only substantial issues in conditions with more than 29% noninvariance parameters of a medium or large magnitude. Threshold noninvariance produced the largest bias for latent means and variances. |

*Note.* MI: type of measurement invariance tested, PMI: partial measurement invariance; AMI approximate measurement invariance. PMG-CFA: partial invariance multigroup CFA; AMG-CFA: partial invariance multigroup with alignment optimization; PMG-BSEM: partial invariance multigroup BSEM; MG-BSEM: multigroup BSEM.

Table 2

*Conditions for the Simulations*

| No. of items per scale | No. of noninvariant items (partial noninvariance) per group | No. of groups affected by noninvariance | Wiggle room for parameter differences (approximate invariance | Type of MI |
|---|---|---|---|---|
| 3 | 1,2 | 25% | 0 | PMI only |
| 3 | 1,2 | 50% | 0 | PMI only |
| 3 | 1,2 | 75% | 0 | PMI only |
| 3 | 1,2 | 100% | 0 | PMI only |
| 4 | 1,2,3 | 25% | 0 | PMI only |
| 4 | 1,2,3 | 50% | 0 | PMI only |
| 4 | 1,2,3 | 75% | 0 | PMI only |
| 4 | 1,2,3 | 100% | 0 | PMI only |
| 5 | 1,2,3,4 | 25% | 0 | PMI only |
| 5 | 1,2,3,4 | 50% | 0 | PMI only |
| 5 | 1,2,3,4 | 75% | 0 | PMI only |
| 5 | 1,2,3,4 | 100% | 0 | PMI only |
| 3,4,5 | 0 | 0 | 0.001 | AMI only |
| 3,4,5 | 0 | 0 | 0.005 | AMI only |
| 3,4,5 | 0 | 0 | 0.010 | AMI only |
| 3,4,5 | 0 | 0 | 0.050 | AMI only |
| 5 | 1,2,3,4 | 25% | 0.005 | PMI +AMI |
| 5 | 1,2,3,4 | 25% | 0.010 | PMI +AMI |
| 5 | 1,2,3,4 | 50% | 0.005 | PMI +AMI |
| 5 | 1,2,3,4 | 50% | 0.010 | PMI +AMI |
| 5 | 1,2,3,4 | 75% | 0.005 | PMI +AMI |
| 5 | 1,2,3,4 | 75% | 0.010 | PMI +AMI |
| 5 | 1,2,3 | 100% | 0.005 | PMI +AMI |
| 5 | 1,2,3 | 100% | 0.010 | PMI +AMI |

*Note*. Applied to MG-CFA (no MI), MG-CFA, PMG-CFA, MG-BSEM, PMG-BSEM, AMG-CFA (for abbreviations, see the note to Table 1). Each of these six measurement models was also examined as an SEM model with an external dependent variable. PMI: partial measurement invariance; AMI: approximate measurement invariance. In all simulation studies the number of groups was 24, and the sample size was 1,500; standard deviation of the true latent means: 0.3; the number of response categories for the items was 5; true values of regression coefficients to two criterion variables were 0.1 and 0.3; the number of replications was 400.

Table 3. *Recovery of Group Means Under Partial MI*

| Groups affected | Scale | Noninvariant | 1. MG-CFA (MI) | | | 2. MG-CFA | | | 3. PMG-CFA | | | 4. MG-BSEM | | | 5. PMG-BSEM | | | 6. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 |
| 25% | 3-items | 1 | **0.993** | 0.075 | 0.713 | **0.982** | 0.085 | 0.686 | **0.993** | 0.074 | 0.709 | **0.981** | 0.107 | 0.615 | **0.993** | 0.083 | 0.726 | **0.989** | 0.077 | 0.697 |
| | | 2 | **0.993** | 0.075 | 0.713 | 0.967 | 0.100 | 0.649 | **0.990** | 0.077 | 0.710 | 0.968 | 0.132 | 0.562 | **0.991** | 0.085 | 0.752 | 0.956 | 0.109 | 0.626 |
| | 4-items | 1 | **0.991** | **0.055** | 0.862 | **0.985** | 0.064 | 0.817 | **0.991** | **0.055** | 0.861 | **0.985** | 0.083 | 0.662 | **0.991** | 0.061 | 0.818 | **0.990** | **0.057** | 0.851 |
| | | 2 | **0.991** | **0.055** | 0.862 | 0.977 | 0.074 | 0.773 | **0.990** | **0.056** | 0.861 | 0.978 | 0.106 | 0.582 | **0.990** | 0.061 | 0.820 | 0.979 | 0.074 | 0.759 |
| | | 3 | **0.991** | **0.055** | 0.862 | 0.967 | 0.083 | 0.749 | **0.988** | **0.058** | 0.861 | 0.971 | 0.131 | 0.503 | **0.988** | 0.062 | 0.843 | 0.951 | 0.101 | 0.683 |
| | 5-items | 1 | **0.994** | **0.040** | **0.948** | **0.989** | **0.049** | 0.889 | **0.993** | **0.041** | **0.944** | **0.990** | 0.063 | 0.795 | **0.994** | **0.039** | **0.950** | **0.993** | **0.044** | **0.928** |
| | | 2 | **0.994** | **0.040** | **0.948** | **0.984** | **0.057** | 0.855 | **0.993** | **0.041** | **0.940** | **0.985** | 0.084 | 0.682 | **0.994** | **0.039** | **0.948** | **0.989** | **0.051** | 0.883 |
| | | 3 | **0.994** | **0.040** | **0.948** | 0.979 | 0.064 | 0.835 | **0.993** | **0.042** | **0.940** | **0.981** | 0.110 | 0.579 | **0.993** | **0.039** | **0.950** | 0.976 | 0.069 | 0.820 |
| | | 4 | **0.994** | **0.040** | **0.948** | 0.971 | 0.073 | 0.810 | **0.991** | **0.045** | **0.935** | 0.976 | 0.132 | 0.508 | **0.992** | **0.042** | **0.956** | 0.951 | 0.096 | 0.756 |
| 50% | 3-items | 1 | **0.993** | 0.075 | 0.713 | 0.970 | 0.096 | 0.595 | **0.992** | 0.075 | 0.704 | 0.970 | 0.126 | 0.545 | **0.993** | 0.081 | 0.745 | **0.983** | 0.085 | 0.638 |
| | | 2 | **0.993** | 0.075 | 0.713 | 0.946 | 0.119 | 0.553 | **0.988** | 0.079 | 0.718 | 0.948 | 0.181 | 0.431 | **0.989** | 0.085 | 0.780 | 0.929 | 0.133 | 0.512 |
| | 4-items | 1 | **0.991** | **0.055** | 0.862 | 0.978 | 0.073 | 0.751 | **0.990** | **0.055** | 0.861 | 0.978 | 0.102 | 0.582 | **0.991** | **0.058** | 0.840 | **0.987** | 0.061 | 0.816 |
| | | 2 | **0.991** | **0.055** | 0.862 | 0.962 | 0.090 | 0.683 | **0.988** | **0.058** | 0.858 | 0.963 | 0.150 | 0.454 | **0.989** | 0.061 | 0.841 | 0.964 | 0.091 | 0.651 |
| | | 3 | **0.991** | **0.055** | 0.862 | 0.946 | 0.106 | 0.647 | **0.985** | 0.062 | 0.855 | 0.951 | 0.208 | 0.347 | **0.985** | 0.064 | 0.872 | 0.926 | 0.126 | 0.554 |
| | 5-items | 1 | **0.994** | **0.040** | **0.948** | **0.985** | **0.056** | 0.835 | **0.993** | **0.041** | **0.942** | **0.985** | 0.084 | 0.675 | **0.994** | **0.038** | **0.958** | **0.989** | **0.048** | **0.905** |
| | | 2 | **0.994** | **0.040** | **0.948** | 0.975 | 0.069 | 0.764 | **0.993** | **0.041** | **0.940** | 0.977 | 0.122 | 0.545 | **0.994** | **0.038** | **0.955** | **0.984** | 0.061 | 0.816 |
| | | 3 | **0.994** | **0.040** | **0.948** | 0.964 | 0.083 | 0.720 | **0.992** | **0.043** | **0.939** | 0.968 | 0.177 | 0.411 | **0.993** | **0.041** | **0.953** | 0.962 | 0.091 | 0.674 |
| | | 4 | **0.994** | **0.040** | **0.948** | 0.953 | 0.095 | 0.677 | **0.988** | **0.050** | **0.921** | 0.960 | 0.230 | 0.342 | **0.989** | **0.049** | **0.953** | 0.928 | 0.121 | 0.574 |
| 75% | 3-items | 1 | **0.993** | 0.075 | 0.713 | 0.960 | 0.107 | 0.538 | **0.991** | 0.076 | 0.719 | 0.958 | 0.157 | 0.465 | **0.992** | 0.082 | 0.757 | 0.972 | 0.101 | 0.570 |
| | | 2 | **0.993** | 0.075 | 0.713 | 0.924 | 0.137 | 0.460 | **0.985** | 0.082 | 0.718 | 0.925 | 0.240 | 0.355 | **0.986** | 0.088 | 0.810 | 0.908 | 0.151 | 0.421 |
| | 4-items | 1 | **0.991** | **0.055** | 0.862 | 0.971 | 0.080 | 0.674 | **0.990** | **0.056** | 0.853 | 0.971 | 0.127 | 0.490 | **0.990** | **0.059** | 0.838 | **0.984** | 0.068 | 0.750 |
| | | 2 | **0.991** | **0.055** | 0.862 | 0.949 | 0.103 | 0.582 | **0.988** | **0.058** | 0.852 | 0.951 | 0.200 | 0.360 | **0.988** | **0.059** | 0.858 | 0.951 | 0.109 | 0.531 |
| | | 3 | **0.991** | **0.055** | 0.862 | 0.924 | 0.124 | 0.510 | **0.981** | 0.067 | 0.833 | 0.931 | 0.302 | 0.259 | **0.982** | 0.067 | 0.883 | 0.907 | 0.147 | 0.399 |
| | 5-items | 1 | **0.994** | **0.040** | **0.948** | **0.981** | 0.062 | 0.782 | **0.993** | **0.041** | **0.941** | **0.981** | 0.101 | 0.598 | **0.994** | **0.038** | **0.953** | **0.991** | **0.050** | 0.881 |
| | | 2 | **0.994** | **0.040** | **0.948** | 0.966 | 0.080 | 0.677 | **0.993** | **0.042** | **0.937** | 0.968 | 0.168 | 0.433 | **0.993** | **0.039** | **0.955** | 0.976 | 0.076 | 0.703 |
| | | 3 | **0.994** | **0.040** | **0.948** | 0.951 | 0.097 | 0.607 | **0.991** | **0.046** | **0.926** | 0.956 | 0.262 | 0.313 | **0.991** | **0.044** | **0.947** | 0.948 | 0.109 | 0.524 |
| | | 4 | **0.994** | **0.040** | **0.948** | 0.937 | 0.113 | 0.542 | **0.985** | **0.055** | **0.905** | 0.946 | 0.347 | 0.262 | **0.985** | **0.057** | **0.944** | 0.919 | 0.136 | 0.419 |
| 100% | 3-items | 1 | **0.993** | 0.075 | 0.713 | 0.951 | 0.144 | 0.395 | **0.991** | 0.076 | 0.743 | 0.948 | 0.209 | 0.358 | **0.991** | 0.089 | 0.756 | 0.955 | 0.136 | 0.445 |
| | | 2 | **0.993** | 0.075 | 0.713 | 0.908 | 0.198 | 0.312 | NA | NA | NA | 0.909 | 0.367 | 0.239 | NA | NA | NA | 0.894 | 0.191 | 0.352 |
| | 4-items | 1 | **0.991** | **0.055** | 0.862 | 0.966 | 0.111 | 0.483 | **0.990** | **0.056** | 0.852 | 0.965 | 0.168 | 0.382 | **0.990** | **0.059** | 0.835 | **0.980** | 0.090 | 0.625 |
| | | 2 | **0.991** | **0.055** | 0.862 | 0.941 | 0.147 | 0.388 | **0.988** | 0.062 | 0.832 | 0.943 | 0.289 | 0.258 | **0.988** | 0.064 | 0.846 | 0.938 | 0.145 | 0.387 |
| | | 3 | **0.991** | **0.055** | 0.862 | 0.913 | 0.179 | 0.324 | NA | NA | NA | 0.920 | 0.436 | 0.191 | NA | NA | NA | 0.908 | 0.170 | 0.340 |
| | 5-items | 1 | **0.994** | **0.040** | **0.948** | 0.977 | 0.089 | 0.573 | **0.993** | **0.041** | **0.944** | 0.978 | 0.135 | 0.470 | **0.994** | **0.039** | **0.959** | **0.989** | 0.061 | 0.786 |
| | | 2 | **0.994** | **0.040** | **0.948** | 0.958 | 0.118 | 0.460 | **0.992** | **0.045** | **0.931** | 0.961 | 0.240 | 0.323 | **0.992** | **0.044** | **0.950** | 0.965 | 0.110 | 0.480 |
| | | 3 | **0.994** | **0.040** | **0.948** | 0.940 | 0.145 | 0.381 | **0.989** | **0.051** | **0.917** | 0.945 | 0.368 | 0.237 | **0.990** | **0.053** | **0.938** | 0.937 | 0.141 | 0.373 |
| | | 4 | **0.994** | **0.040** | **0.948** | 0.922 | 0.171 | 0.326 | NA | NA | NA | 0.932 | 0.529 | 0.190 | NA | NA | NA | 0.923 | 0.153 | 0.365 |

*Note*. Cor – correlation between the true and estimated means (> 0.98 in bold); RMSE – root mean square error (> 0.06 in bold); CI95 – 95% CI coverage (> 0.9 in bold); NA-model not identified. The first is a reference which is based on data without noninvariant items and an MG-CFA model. The other columns 2-6 impose different model specifications on data with noninvariant items according to the different conditions listed above. For abbreviations of the Models 1-6, see the note to Table 1.

Table 4. *Recovery of Path Coefficients Under Partial MI*

| Groups affected | Scale | Noninvariant | 1. MG-CFA (MI) | | | 2. MG-CFA | | | 3. PMG-CFA | | | 4. MG-BSEM | | | 5. PMG-BSEM | | | 6. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 |
| 25% | 3-items | 1 | **0.303** | **0.008** | **0.920** | **0.304** | **0.009** | 0.899 | **0.302** | **0.008** | **0.918** | **0.293** | **0.012** | 0.870 | **0.301** | **0.010** | 0.895 | **0.305** | **0.011** | **0.916** |
| | | 2 | **0.302** | **0.008** | **0.924** | **0.306** | **0.010** | 0.878 | **0.302** | **0.008** | **0.921** | **0.287** | **0.016** | 0.839 | **0.304** | **0.010** | **0.902** | **0.307** | **0.012** | 0.868 |
| | 4-items | 1 | **0.303** | **0.008** | **0.914** | **0.304** | **0.008** | **0.903** | **0.303** | **0.007** | **0.915** | **0.289** | **0.013** | 0.880 | **0.299** | **0.009** | **0.919** | **0.307** | **0.011** | **0.924** |
| | | 2 | **0.302** | **0.007** | **0.921** | **0.303** | **0.008** | 0.890 | **0.303** | **0.007** | **0.916** | 0.282 | **0.019** | 0.837 | **0.301** | **0.008** | **0.926** | **0.305** | **0.010** | **0.904** |
| | | 3 | **0.303** | **0.007** | **0.916** | **0.304** | **0.008** | 0.872 | **0.303** | **0.007** | **0.915** | 0.275 | **0.026** | 0.766 | **0.304** | **0.009** | **0.927** | **0.305** | **0.010** | 0.861 |
| | 5-items | 1 | **0.303** | **0.007** | **0.915** | **0.302** | **0.007** | **0.909** | **0.302** | **0.007** | **0.917** | **0.288** | **0.014** | 0.894 | **0.298** | **0.010** | **0.933** | **0.306** | **0.010** | **0.931** |
| | | 2 | **0.302** | **0.007** | **0.918** | **0.303** | **0.007** | 0.897 | **0.303** | **0.007** | **0.919** | 0.280 | **0.020** | 0.842 | **0.301** | **0.009** | **0.938** | **0.306** | **0.010** | **0.921** |
| | | 3 | **0.303** | **0.008** | **0.914** | **0.302** | **0.007** | 0.886 | **0.303** | **0.007** | **0.914** | 0.271 | **0.030** | 0.749 | **0.303** | **0.009** | **0.941** | **0.305** | **0.009** | **0.901** |
| | | 4 | **0.303** | **0.007** | **0.919** | **0.302** | **0.007** | 0.871 | **0.303** | **0.007** | **0.915** | 0.262 | **0.038** | 0.644 | **0.305** | **0.009** | **0.939** | **0.306** | **0.010** | 0.847 |
| 50% | 3-items | 1 | **0.303** | **0.008** | **0.925** | **0.306** | **0.009** | 0.880 | **0.303** | **0.008** | **0.919** | **0.287** | **0.016** | 0.831 | **0.304** | **0.011** | 0.892 | **0.304** | **0.011** | **0.908** |
| | | 2 | **0.304** | **0.008** | **0.920** | **0.309** | **0.012** | 0.822 | **0.303** | **0.008** | **0.917** | 0.276 | **0.025** | 0.743 | **0.310** | **0.012** | **0.901** | **0.307** | **0.013** | 0.805 |
| | 4-items | 1 | **0.303** | **0.007** | **0.915** | **0.303** | **0.008** | 0.890 | **0.302** | **0.007** | **0.917** | 0.281 | **0.020** | 0.831 | **0.300** | **0.009** | **0.920** | **0.306** | **0.011** | **0.917** |
| | | 2 | **0.303** | **0.007** | **0.918** | **0.305** | **0.009** | 0.863 | **0.303** | **0.008** | **0.912** | 0.267 | **0.033** | 0.691 | **0.304** | **0.009** | **0.927** | **0.304** | **0.010** | 0.881 |
| | | 3 | **0.303** | **0.008** | **0.920** | **0.306** | **0.010** | 0.823 | **0.303** | **0.007** | **0.911** | 0.251 | **0.049** | 0.492 | **0.309** | **0.011** | **0.923** | **0.303** | **0.011** | 0.799 |
| | 5-items | 1 | **0.303** | **0.007** | **0.923** | **0.303** | **0.007** | **0.901** | **0.303** | **0.007** | **0.917** | 0.280 | **0.020** | 0.845 | **0.302** | **0.009** | **0.937** | **0.306** | **0.010** | **0.927** |
| | | 2 | **0.303** | **0.007** | **0.918** | **0.302** | **0.007** | 0.882 | **0.302** | **0.007** | **0.916** | 0.262 | **0.038** | 0.641 | **0.304** | **0.009** | **0.935** | **0.305** | **0.009** | **0.910** |
| | | 3 | **0.302** | **0.007** | **0.916** | **0.302** | **0.008** | 0.856 | **0.303** | **0.007** | **0.914** | 0.246 | **0.054** | 0.411 | **0.308** | **0.010** | **0.936** | **0.304** | **0.009** | 0.867 |
| | | 4 | **0.303** | **0.007** | **0.918** | **0.302** | **0.009** | 0.829 | **0.302** | **0.007** | **0.913** | 0.229 | 0.071 | 0.216 | **0.312** | **0.013** | **0.932** | **0.304** | **0.012** | 0.795 |
| 75% | 3-items | 1 | **0.303** | **0.008** | **0.915** | **0.308** | **0.010** | 0.860 | **0.303** | **0.008** | **0.917** | 0.283 | **0.019** | 0.797 | **0.308** | **0.012** | **0.900** | **0.305** | **0.011** | 0.885 |
| | | 2 | **0.302** | **0.008** | **0.920** | **0.312** | **0.014** | 0.791 | **0.302** | **0.008** | **0.922** | 0.264 | **0.036** | 0.619 | **0.314** | **0.015** | **0.900** | **0.307** | **0.014** | 0.767 |
| | 4-items | 1 | **0.303** | **0.007** | **0.914** | **0.302** | **0.008** | 0.878 | **0.301** | **0.007** | **0.914** | 0.273 | **0.027** | 0.756 | **0.301** | **0.009** | **0.925** | **0.304** | **0.009** | **0.921** |
| | | 2 | **0.303** | **0.008** | **0.918** | **0.304** | **0.009** | 0.830 | **0.301** | **0.007** | **0.916** | 0.252 | **0.048** | 0.493 | **0.307** | **0.010** | **0.924** | **0.302** | **0.010** | 0.853 |
| | | 3 | **0.303** | **0.008** | **0.918** | **0.306** | **0.010** | 0.779 | **0.302** | **0.007** | **0.915** | 0.230 | 0.070 | 0.261 | 0.315 | **0.016** | **0.916** | **0.301** | **0.012** | 0.749 |
| | 5-items | 1 | **0.303** | **0.008** | **0.912** | **0.302** | **0.007** | 0.893 | **0.302** | **0.007** | **0.917** | 0.271 | **0.029** | 0.758 | **0.302** | **0.009** | **0.938** | **0.305** | **0.009** | 0.925 |
| | | 2 | **0.302** | **0.007** | **0.925** | **0.301** | **0.008** | 0.863 | **0.302** | **0.007** | **0.922** | 0.246 | **0.054** | 0.409 | **0.307** | **0.010** | **0.937** | **0.304** | **0.009** | 0.898 |
| | | 3 | **0.302** | **0.007** | **0.915** | **0.303** | **0.009** | 0.814 | **0.303** | **0.007** | **0.911** | 0.224 | 0.076 | 0.180 | **0.313** | **0.014** | **0.925** | **0.304** | **0.010** | 0.832 |
| | | 4 | **0.303** | **0.007** | **0.915** | **0.302** | **0.009** | 0.785 | **0.303** | **0.007** | **0.908** | 0.202 | 0.098 | 0.064 | 0.320 | **0.020** | **0.909** | **0.302** | **0.012** | 0.762 |
| 100% | 3-items | 1 | **0.302** | **0.008** | **0.919** | **0.308** | **0.026** | 0.764 | **0.302** | **0.008** | **0.921** | 0.274 | **0.029** | 0.697 | **0.309** | **0.014** | 0.891 | **0.306** | **0.023** | 0.786 |
| | | 2 | **0.302** | **0.007** | **0.924** | 0.322 | **0.040** | 0.638 | NA | NA | NA | 0.261 | **0.042** | 0.571 | NA | NA | NA | **0.316** | **0.037** | 0.629 |
| | 4-items | 1 | **0.303** | **0.007** | **0.916** | **0.307** | **0.021** | 0.827 | **0.303** | **0.008** | **0.915** | 0.266 | **0.035** | 0.647 | **0.304** | **0.012** | **0.919** | **0.306** | **0.015** | 0.888 |
| | | 2 | **0.302** | **0.008** | **0.917** | **0.309** | **0.030** | 0.705 | **0.302** | **0.008** | **0.916** | 0.237 | 0.063 | 0.354 | 0.316 | **0.018** | **0.905** | **0.307** | **0.029** | 0.711 |
| | | 3 | **0.302** | **0.007** | **0.923** | 0.318 | **0.038** | 0.639 | NA | NA | NA | 0.242 | **0.059** | 0.388 | NA | NA | NA | **0.311** | **0.035** | 0.638 |
| | 5-items | 1 | **0.302** | **0.007** | **0.918** | **0.306** | **0.017** | 0.853 | **0.303** | **0.007** | **0.922** | 0.263 | **0.037** | 0.644 | **0.305** | **0.010** | **0.937** | **0.307** | **0.012** | **0.916** |
| | | 2 | **0.302** | **0.007** | **0.917** | **0.309** | **0.024** | 0.768 | **0.303** | **0.008** | **0.909** | 0.231 | 0.069 | 0.265 | **0.314** | **0.015** | **0.918** | **0.308** | **0.019** | 0.817 |
| | | 3 | **0.303** | **0.007** | **0.910** | **0.312** | **0.029** | 0.715 | **0.301** | **0.008** | **0.915** | 0.203 | 0.097 | 0.092 | 0.326 | **0.027** | 0.884 | **0.308** | **0.026** | 0.717 |
| | | 4 | **0.305** | **0.005** | **0.908** | 0.318 | **0.035** | 0.636 | NA | NA | NA | 0.219 | 0.081 | 0.167 | NA | NA | NA | **0.311** | **0.030** | 0.665 |

*Note*. mean – means of estimated parameters (0.285 < in bold < 0.315); RMSE – root mean square error (< 0.06 in bold); CI95 – 95% CI coverage (> 0.9 in bold); NA - model not identified. Six columns report these statistics. The first is a reference which is based on data without noninvariant items and an MG-CFA model. The other columns 2-6 impose different model specifications on data with noninvariant items according to the different conditions listed above. For abbreviations of the Models 1-6, see the note to Table 1.

Table 5

*Recovery of Group Means for AMI Situation*

| Scale | AMI | 1. MG-CFA | | | 2. MG-BSEM | | | 3. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 |
| 3-items | 0.001 | **0.991** | 0.078 | 0.642 | **0.992** | 0.077 | 0.803 | **0.991** | 0.076 | 0.673 |
| | 0.005 | **0.983** | 0.090 | 0.583 | **0.984** | 0.116 | 0.794 | **0.983** | 0.085 | 0.625 |
| | 0.010 | 0.974 | 0.103 | 0.538 | 0.977 | 0.169 | 0.859 | 0.972 | 0.099 | 0.558 |
| | 0.050 | 0.909 | 0.195 | 0.334 | NA | NA | NA | 0.896 | 0.180 | 0.384 |
| 4-items | 0.001 | **0.989** | **0.059** | 0.813 | **0.990** | 0.063 | 0.863 | **0.989** | **0.058** | 0.829 |
| | 0.005 | **0.984** | 0.072 | 0.702 | **0.985** | 0.085 | 0.889 | **0.984** | 0.070 | 0.716 |
| | 0.010 | 0.977 | 0.086 | 0.628 | 0.977 | 0.109 | 0.893 | 0.976 | 0.084 | 0.624 |
| | 0.050 | 0.927 | 0.156 | 0.382 | 0.927 | 0.299 | **0.955** | 0.919 | 0.149 | 0.385 |
| 5-items | 0.001 | **0.993** | **0.044** | **0.921** | **0.993** | **0.044** | **0.955** | **0.991** | **0.045** | **0.923** |
| | 0.005 | **0.988** | **0.058** | 0.796 | **0.989** | 0.064 | **0.946** | **0.988** | **0.057** | 0.807 |
| | 0.010 | **0.983** | 0.071 | 0.711 | **0.984** | 0.086 | **0.935** | **0.983** | 0.070 | 0.699 |
| | 0.050 | 0.939 | 0.140 | 0.404 | 0.940 | 0.204 | **0.939** | 0.935 | 0.134 | 0.398 |

*Note*. Cor – correlation between the true and estimated means (> 0.98 in bold); RMSE – root mean square error (< 0.06 in bold); CI95 – 95% CI coverage (> 0.9 in bold) AMI: approximate measurement invariance. NA- no convergence. For abbreviations of the Models 1-3, see the note to Table 1.

Table 6

*Recovery of Path Coefficients for AMI Situation*

| Scale | AMI | 1. MG-CFA | | | 2. MG-BSEM | | | 3. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 |
| 3-items | 0.001 | **0.303** | **0.036** | **0.914** | **0.294** | **0.035** | **0.912** | **0.307** | **0.033** | **0.922** |
| | 0.005 | **0.303** | **0.038** | 0.891 | **0.287** | **0.038** | **0.900** | **0.306** | **0.036** | 0.892 |
| | 0.010 | **0.305** | **0.042** | 0.855 | 0.282 | **0.042** | 0.875 | **0.307** | **0.040** | 0.850 |
| | 0.050 | **0.312** | 0.069 | 0.661 | 0.266 | 0.061 | 0.821 | **0.306** | **0.064** | 0.650 |
| 4-items | 0.001 | **0.303** | **0.034** | **0.910** | **0.292** | **0.034** | **0.909** | **0.306** | **0.032** | **0.920** |
| | 0.005 | **0.305** | **0.037** | 0.886 | **0.287** | **0.037** | 0.886 | **0.309** | **0.035** | 0.896 |
| | 0.010 | **0.305** | **0.040** | 0.861 | 0.284 | **0.039** | 0.878 | **0.308** | **0.037** | 0.874 |
| | 0.050 | 0.315 | 0.061 | 0.703 | 0.279 | **0.050** | 0.878 | **0.311** | **0.056** | 0.707 |
| 5-items | 0.001 | **0.303** | **0.033** | **0.913** | **0.294** | **0.033** | **0.907** | **0.308** | **0.032** | **0.919** |
| | 0.005 | **0.303** | **0.036** | 0.886 | **0.291** | **0.035** | **0.902** | **0.307** | **0.034** | **0.903** |
| | 0.010 | **0.305** | **0.038** | 0.867 | **0.290** | **0.036** | **0.907** | **0.309** | **0.036** | 0.883 |
| | 0.050 | **0.311** | **0.055** | 0.716 | **0.289** | **0.045** | **0.919** | **0.308** | **0.051** | 0.722 |

*Note*. mean – means of estimated parameters (0.285 < in bold < 0.315); RMSE – root mean square error (< 0.06 in bold); CI95 – 95% CI coverage (> 0.9 in bold); AMI – approximate measurement invariance. For abbreviations of the Models 1-3, see the note to Table 1.

Table 7

*Recovery of Group Means for Partial and Approximate MI Situation, 5-Item Scale*

| Groups affected | AMI | Noninvariant | 1. MG-CFA | | | 2. PMG-CFA | | | 3. MG-BSEM | | | 4. PMG-BSEM | | | 5. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 | Cor | RMSE | CI95 |
| 25% | 0.005 | 1 | **0.984** | 0.064 | 0.759 | **0.987** | 0.059 | 0.790 | **0.985** | 0.088 | 0.843 | **0.988** | 0.064 | **0.933** | **0.986** | 0.062 | 0.762 |
| | | 2 | 0.979 | 0.070 | 0.742 | **0.987** | 0.060 | 0.798 | 0.980 | 0.113 | 0.765 | **0.988** | 0.062 | **0.945** | 0.980 | 0.071 | 0.723 |
| | | 3 | 0.973 | 0.078 | 0.713 | **0.984** | 0.064 | 0.777 | 0.976 | 0.119 | 0.731 | **0.985** | 0.068 | **0.940** | 0.967 | 0.086 | 0.663 |
| | | 4 | 0.969 | 0.082 | 0.709 | 0.980 | 0.069 | 0.767 | 0.974 | 0.166 | 0.627 | 0.980 | 0.072 | **0.934** | 0.952 | 0.101 | 0.630 |
| | 0.010 | 1 | 0.979 | 0.074 | 0.695 | **0.982** | 0.070 | 0.718 | 0.980 | 0.104 | 0.879 | **0.983** | 0.082 | **0.943** | 0.979 | 0.075 | 0.673 |
| | | 2 | 0.973 | 0.080 | 0.669 | 0.979 | 0.073 | 0.703 | 0.975 | 0.120 | 0.844 | **0.981** | 0.080 | **0.944** | 0.972 | 0.081 | 0.648 |
| | | 3 | 0.968 | 0.086 | 0.645 | 0.977 | 0.076 | 0.690 | 0.972 | 0.144 | 0.781 | 0.979 | 0.082 | **0.944** | 0.963 | 0.093 | 0.597 |
| | | 4 | 0.964 | 0.091 | 0.641 | 0.968 | 0.087 | 0.667 | 0.969 | 0.165 | 0.727 | 0.969 | 0.092 | **0.947** | 0.949 | 0.107 | 0.564 |
| 50% | 0.005 | 1 | 0.979 | 0.071 | 0.718 | **0.987** | 0.061 | 0.784 | 0.980 | 0.115 | 0.748 | **0.988** | 0.062 | **0.937** | **0.983** | 0.070 | 0.708 |
| | | 2 | 0.971 | 0.080 | 0.675 | **0.985** | 0.063 | 0.780 | 0.973 | 0.161 | 0.639 | **0.986** | 0.063 | **0.938** | 0.973 | 0.083 | 0.638 |
| | | 3 | 0.961 | 0.091 | 0.638 | **0.982** | 0.066 | 0.769 | 0.965 | 0.162 | 0.622 | **0.983** | 0.080 | **0.932** | 0.954 | 0.103 | 0.559 |
| | | 4 | 0.949 | 0.103 | 0.603 | 0.971 | 0.079 | 0.726 | 0.957 | 0.261 | 0.480 | 0.972 | 0.078 | **0.943** | 0.928 | 0.127 | 0.477 |
| | 0.010 | 1 | 0.975 | 0.081 | 0.649 | **0.981** | 0.073 | 0.695 | 0.977 | 0.121 | 0.827 | **0.983** | 0.081 | **0.938** | 0.976 | 0.081 | 0.626 |
| | | 2 | 0.967 | 0.088 | 0.631 | 0.978 | 0.074 | 0.696 | 0.969 | 0.161 | 0.739 | 0.979 | 0.078 | **0.952** | 0.965 | 0.094 | 0.565 |
| | | 3 | 0.957 | 0.099 | 0.580 | 0.972 | 0.082 | 0.667 | 0.962 | 0.167 | 0.705 | 0.974 | 0.100 | **0.907** | 0.948 | 0.114 | 0.482 |
| | | 4 | 0.948 | 0.110 | 0.545 | 0.954 | 0.102 | 0.614 | 0.955 | 0.249 | 0.601 | 0.954 | 0.104 | **0.945** | 0.927 | 0.132 | 0.429 |
| 75% | 0.005 | 1 | 0.976 | 0.074 | 0.684 | **0.987** | 0.060 | 0.789 | 0.977 | 0.137 | 0.689 | **0.988** | 0.063 | **0.932** | **0.981** | 0.073 | 0.690 |
| | | 2 | 0.963 | 0.090 | 0.605 | **0.984** | 0.064 | 0.769 | 0.965 | 0.213 | 0.547 | **0.985** | 0.064 | **0.940** | 0.964 | 0.097 | 0.545 |
| | | 3 | 0.949 | 0.104 | 0.555 | 0.978 | 0.071 | 0.744 | 0.954 | 0.213 | 0.526 | 0.980 | 0.081 | **0.930** | 0.943 | 0.117 | 0.459 |
| | | 4 | 0.930 | 0.121 | 0.499 | 0.960 | 0.090 | 0.688 | 0.940 | 0.365 | 0.409 | 0.961 | 0.086 | **0.940** | 0.913 | 0.140 | 0.381 |
| | 0.010 | 1 | 0.972 | 0.084 | 0.631 | 0.980 | 0.074 | 0.699 | 0.973 | 0.143 | 0.778 | **0.981** | 0.081 | **0.939** | 0.973 | 0.088 | 0.592 |
| | | 2 | 0.960 | 0.097 | 0.565 | 0.976 | 0.079 | 0.669 | 0.963 | 0.200 | 0.669 | 0.978 | 0.080 | **0.951** | 0.958 | 0.105 | 0.503 |
| | | 3 | 0.945 | 0.112 | 0.509 | 0.966 | 0.090 | 0.623 | 0.951 | 0.199 | 0.651 | 0.968 | 0.085 | **0.944** | 0.938 | 0.122 | 0.429 |
| | | 4 | 0.933 | 0.122 | 0.481 | 0.938 | 0.114 | 0.569 | 0.942 | 0.329 | 0.518 | 0.939 | 0.111 | **0.945** | 0.917 | 0.144 | 0.368 |
| 100% | 0.005 | 1 | 0.973 | 0.094 | 0.556 | **0.986** | 0.063 | 0.777 | 0.974 | 0.162 | 0.616 | **0.987** | 0.064 | **0.944** | 0.979 | 0.085 | 0.602 |
| | | 2 | 0.957 | 0.123 | 0.447 | **0.982** | 0.070 | 0.740 | 0.959 | 0.261 | 0.482 | **0.983** | 0.070 | **0.948** | 0.957 | 0.119 | 0.435 |
| | | 3 | 0.938 | 0.149 | 0.373 | 0.972 | 0.087 | 0.678 | 0.944 | 0.272 | 0.431 | 0.974 | 0.105 | 0.890 | 0.934 | 0.144 | 0.358 |
| | 0.010 | 1 | 0.968 | 0.100 | 0.534 | 0.979 | 0.077 | 0.677 | 0.970 | 0.163 | 0.733 | **0.981** | 0.087 | **0.940** | 0.970 | 0.099 | 0.532 |
| | | 2 | 0.952 | 0.129 | 0.419 | 0.973 | 0.090 | 0.618 | 0.956 | 0.248 | 0.582 | 0.974 | 0.092 | **0.944** | 0.950 | 0.128 | 0.401 |
| | | 3 | 0.936 | 0.152 | 0.365 | 0.958 | 0.109 | 0.565 | 0.943 | 0.247 | 0.563 | 0.959 | 0.135 | 0.887 | 0.933 | 0.145 | 0.364 |

*Note.* Cor – correlation between the true and estimated means (> 0.98 in bold); RMSE – root mean square error (< 0.06 in bold); CI95 – 95% CI coverage (> 0.9 in bold) AMI – approximate measurement invariance. For abbreviations of the Models 1-5, see the note to Table 1.

Table 8

*Recovery of Path Coefficients for Partial and Approximate MI Situation, 5-Item Scale*

| Groups affected | AMI | Noninvariant | 1. MG-CFA | | | 2. PMG-CFA | | | 3. MG-BSEM | | | 4. PMG-BSEM | | | 5. AMG-CFA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 | mean | RMSE | CI95 |
| 25% | 0.005 | 1 | **0.304** | **0.010** | 0.884 | **0.304** | **0.010** | 0.893 | **0.297** | **0.010** | **0.936** | **0.307** | **0.011** | **0.945** | **0.308** | **0.013** | **0.902** |
| | | 2 | **0.304** | **0.011** | 0.871 | **0.304** | **0.010** | 0.884 | **0.289** | **0.013** | **0.907** | **0.309** | **0.012** | **0.938** | **0.307** | **0.012** | 0.888 |
| | | 3 | **0.303** | **0.011** | 0.861 | **0.303** | **0.010** | 0.888 | 0.282 | **0.019** | 0.865 | **0.311** | **0.014** | **0.940** | **0.306** | **0.013** | 0.870 |
| | | 4 | **0.303** | **0.011** | 0.847 | **0.304** | **0.011** | 0.876 | 0.275 | **0.025** | 0.810 | **0.313** | **0.015** | **0.936** | **0.306** | **0.013** | 0.835 |
| | 0.010 | 1 | **0.305** | **0.014** | 0.856 | **0.305** | **0.014** | 0.864 | 0.278 | **0.023** | 0.839 | **0.293** | **0.013** | **0.911** | **0.307** | **0.015** | 0.870 |
| | | 2 | **0.305** | **0.013** | 0.847 | **0.305** | **0.013** | 0.858 | 0.267 | **0.033** | 0.740 | **0.296** | **0.012** | **0.918** | **0.307** | **0.014** | 0.873 |
| | | 3 | **0.303** | **0.013** | 0.841 | **0.304** | **0.013** | 0.857 | 0.256 | **0.044** | 0.614 | **0.299** | **0.012** | **0.922** | **0.306** | **0.015** | 0.847 |
| | | 4 | **0.303** | **0.013** | 0.832 | **0.304** | **0.013** | 0.850 | 0.247 | **0.053** | 0.507 | **0.304** | **0.012** | **0.931** | **0.306** | **0.015** | 0.817 |
| 50% | 0.005 | 1 | **0.304** | **0.011** | 0.883 | **0.304** | **0.010** | 0.891 | **0.289** | **0.014** | **0.908** | **0.308** | **0.012** | **0.941** | **0.306** | **0.012** | 0.899 |
| | | 2 | **0.302** | **0.011** | 0.859 | **0.302** | **0.011** | 0.887 | 0.275 | **0.025** | 0.809 | **0.311** | **0.014** | **0.940** | **0.304** | **0.012** | 0.878 |
| | | 3 | **0.303** | **0.011** | 0.834 | **0.303** | **0.010** | 0.885 | 0.263 | **0.037** | 0.681 | 0.317 | **0.017** | **0.936** | **0.304** | **0.012** | 0.847 |
| | | 4 | **0.303** | **0.011** | 0.812 | **0.304** | **0.011** | 0.866 | 0.250 | **0.050** | 0.499 | 0.323 | **0.023** | **0.917** | **0.304** | **0.013** | 0.780 |
| | 0.010 | 1 | **0.304** | **0.013** | 0.849 | **0.304** | **0.013** | 0.862 | 0.266 | **0.034** | 0.731 | **0.295** | **0.013** | **0.910** | **0.307** | **0.015** | 0.866 |
| | | 2 | **0.305** | **0.014** | 0.836 | **0.306** | **0.014** | 0.858 | 0.247 | **0.053** | 0.501 | **0.303** | **0.012** | **0.930** | **0.308** | **0.015** | 0.851 |
| | | 3 | **0.303** | **0.014** | 0.813 | **0.304** | **0.013** | 0.850 | 0.264 | **0.036** | 0.695 | 0.323 | **0.024** | **0.917** | **0.305** | **0.015** | 0.826 |
| | | 4 | **0.305** | **0.015** | 0.791 | **0.307** | **0.015** | 0.822 | 0.217 | **0.083** | 0.190 | 0.322 | **0.023** | **0.907** | **0.306** | **0.016** | 0.765 |
| 75% | 0.005 | 1 | **0.303** | **0.010** | 0.868 | **0.304** | **0.010** | 0.891 | 0.282 | **0.019** | 0.864 | **0.311** | **0.013** | **0.938** | **0.306** | **0.012** | 0.891 |
| | | 2 | **0.302** | **0.010** | 0.837 | **0.302** | **0.010** | 0.880 | 0.262 | **0.038** | 0.666 | 0.316 | **0.017** | **0.931** | **0.303** | **0.012** | 0.856 |
| | | 3 | **0.302** | **0.010** | 0.806 | **0.302** | **0.010** | 0.880 | 0.245 | **0.055** | 0.442 | 0.322 | **0.022** | **0.914** | **0.302** | **0.011** | 0.813 |
| | | 4 | **0.304** | **0.013** | 0.767 | **0.306** | **0.012** | 0.852 | 0.232 | **0.068** | 0.279 | 0.333 | **0.033** | 0.890 | **0.305** | **0.015** | 0.739 |
| | 0.010 | 1 | **0.304** | **0.013** | 0.851 | **0.304** | **0.013** | 0.868 | 0.256 | **0.044** | 0.630 | **0.299** | **0.011** | **0.927** | **0.306** | **0.014** | 0.870 |
| | | 2 | **0.305** | **0.014** | 0.814 | **0.305** | **0.014** | 0.855 | 0.231 | **0.069** | 0.305 | **0.310** | **0.014** | **0.924** | **0.306** | **0.015** | 0.830 |
| | | 3 | **0.305** | **0.014** | 0.790 | **0.306** | **0.013** | 0.844 | 0.211 | **0.089** | 0.149 | 0.323 | **0.024** | **0.902** | **0.306** | **0.014** | 0.792 |
| | | 4 | **0.305** | **0.015** | 0.757 | **0.308** | **0.015** | 0.794 | 0.194 | 0.106 | 0.088 | 0.348 | **0.048** | 0.786 | **0.302** | **0.016** | 0.737 |
| 100% | 0.005 | 1 | **0.306** | **0.018** | 0.823 | **0.303** | **0.011** | 0.883 | 0.275 | **0.026** | 0.791 | **0.313** | **0.016** | **0.934** | **0.307** | **0.016** | 0.868 |
| | | 2 | **0.309** | **0.022** | 0.762 | **0.303** | **0.012** | 0.879 | 0.250 | **0.050** | 0.501 | 0.327 | **0.028** | 0.892 | **0.307** | **0.021** | 0.786 |
| | | 3 | **0.314** | **0.032** | 0.686 | **0.305** | **0.015** | 0.858 | 0.232 | **0.068** | 0.303 | 0.355 | **0.055** | 0.736 | **0.310** | **0.029** | 0.696 |
| | 0.010 | 1 | **0.307** | **0.020** | 0.809 | **0.304** | **0.015** | 0.849 | 0.248 | **0.052** | 0.516 | **0.302** | **0.012** | **0.926** | **0.307** | **0.018** | 0.838 |
| | | 2 | **0.310** | **0.026** | 0.736 | **0.304** | **0.017** | 0.829 | 0.219 | **0.081** | 0.209 | 0.322 | **0.024** | 0.899 | **0.309** | **0.026** | 0.745 |
| | | 3 | 0.315 | **0.032** | 0.692 | **0.304** | **0.019** | 0.807 | 0.196 | 0.104 | 0.099 | 0.340 | **0.052** | 0.710 | **0.311** | **0.029** | 0.708 |

*Note:* mean – means of estimated parameters (0.285 < in bold < 0.315); RMSE – root mean square error (< 0.01 in bold); CI95 – 95% CI coverage (> 0.9 in bold) AMI – approximate measurement invariance. For abbreviations of the Models 1-6, see the note to Table1

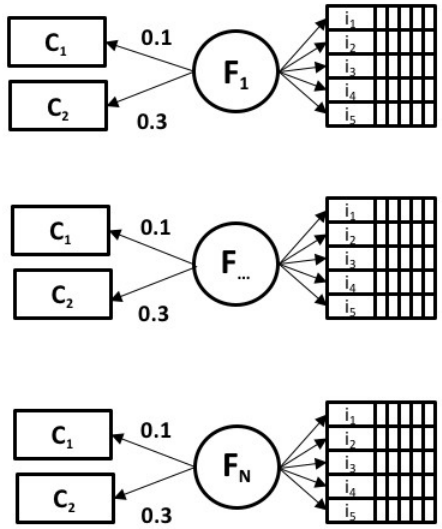


*Figure* 1. Example of a multigroup SEM model used in the simulations. The rectangles on the left represent two observed variables, and rectangles on the right represent categorical items. Circles are latent factors for *n*th group.

# Supplemental material

Assessing Measurement Invariance using Traditional and Newer Approaches: A Monte Carlo Simulation Study

Artur Pokropek
*Institute of Philosophy and Sociology, Polish Academy of Sciences*
Peter Schmidt
*Department of Political Science, Justus Liebig University Giessen*
Eldad Davidov
*Institute of Sociology and Social Psychology, University of Cologne, and Department of Sociology, University of Zurich*

*Figure* S1. Generating the data for the simulation study. Example for a 5-item scale, with large biases for some items (PMI bias) reflecting PMI situation and small “wiggle room” biases reflecting AI situation.