Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 305-334

An extension of the invariance alignment method for scale linking

Artur Pokropek[1], *Oliver Lüdtke*[2] *& Alexander Robitzsch*[3]

Abstract

We examine the extension of the invariance alignment (IA) method originally proposed by Asparouhov and Muthén (2014). The generalized form of a loss function for the IA is discussed, and different forms of the loss function are evaluated using Monte Carlo studies and an empirical example using European Social Survey Data. We compare results obtained by the Mplus software (Muthén & Muthén, 1998-2017) with the R package sirt (Robitzsch, 2019). It is shown that different forms of loss functions that are implemented in the sirt package differ in their performance according to the recovery of group means. This suggests that the performance of IA heavily depends on the form of the loss functions, type of the data (mostly sample size), and type of invariance that could be encountered. The results show that the loss function proposed by Asparouhov and Muthén (2014) might not be optimal in all situations.

Keywords: invariance alignment, CFA, multiple-group model, linking, simulation

[1] *Correspondence concerning this article should be addressed to:* Artur Pokropek, PhD, Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland; email: apokropek@ifispan.waw.pl

[2] IPN − Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany

[3] Centre for International Student Assessment, Munich, Germany

Introduction

In recent years, the concept of approximate measurement invariance (AMI) gained considerable attention from statisticians and aroused high expectations among the applied researchers working with data from many populations (see, e.g., Cieciuch, Davidov, & Schmidt, 2018; Munck 2018; Byrne & van de Vijver, 2017). The AMI postulates that the estimation of reliable and comparable parameters for the groups in multiple group models is possible despite the fact that there exist small "natural" differences between item parameters from different groups or a few completely non-invariant item parameters.

One of the proposed methods to handle AMI was invariance alignment (IA) introduced for confirmatory factor models (CFA) by Asparouhov and Muthén (2014) and subsequently for item response theory (IRT) models (Muthén & Asparouhov, 2014). This procedure replaces the requirement of cross-country equality constraints of parameters by the alignment procedure that minimizes the amount of non-invariance using a particular simplicity function for the identification of the multiple group model. The IA procedure estimates a solution that minimizes overall differences between group-specific parameters using a loss function that is optimized at a few large non-invariant item parameters and many approximately invariant item parameters.

The IA method is available both in the commercial Mplus statistical software (Muthén & Muthén, 1998-2017) and open-source software environment of R (R Core Team, 2019) through the sirt package (Robitzsch, 2019). The advantages of the approach lay in the flexibility of handling various data constellations and very large datasets and the fact that it is computationally not very demanding (compared to alternative AMI methods like Bayesian structural equation modeling or multilevel modeling).

The main aim of this article is to discuss the generalized form of the loss function for IA proposed by Robitzsch (2019) and evaluate different forms of the generalized loss function under different types of non-invariance situations using a Monte Carlo study and an empirical example.

Invariance alignment

Asparouhov and Muthén (2014) describe the IA approach as a procedure that aligns item parameters from group-specific configural CFA or IRT models (Muthén & Asparouhov, 2014) into a most optimal invariance pattern that allows estimating group-specific factor means and variances without requiring exact measurement invariance. In this article, for simplicity, we will focus on CFA models, although our results might be easily translated into the IRT framework (see Muthén & Asparouhov, 2014).

The starting point is a multiple group CFA model:

$$y_{ipg} = \nu_{ig} + \lambda_{ig}\eta_{pg} + \epsilon_{ipg} \;,\; \eta_{pg} \sim \mathrm{N}\left(\alpha_g, \psi_g^2\right) \;, \tag{1}$$

where $i = 1,\ldots,I$ denotes the item index, p the person index and $g = 1,\ldots,G$ the group index, y_{ipg} is a response to the item, ν_{ig} and λ_{ig} are the item parameters, factor intercept, and loading respectively, ϵ_{ipg} is a normally distributed residual variable with $\epsilon_{ipg} \sim \mathrm{N}\left(0,\theta_{ig}^2\right)$, η_{pg} is the latent factor variable and it is assumed to be normally distributed in each group. Note that this multiple group CFA model is not identified because not all item loadings and item intercepts can be simultaneously estimated along with group means $\boldsymbol{\alpha} = \left(\alpha_1,\ldots,\alpha_G\right)$ and standard deviations $\boldsymbol{\psi} = \left(\psi_1,\ldots,\psi_G\right)$. The IA procedure solves the identification issue by determining $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ in such a way that the amount of measurement non-invariance is minimized. This is achieved by utilizing an appropriate alignment function, which optimally aligns group-specific item parameters.

The alignment procedure consists of two steps. In the first step, configural measurement models are estimated for each group. Those models might be CFA models for continuous indicators or IRT models for categorical indicators. The configural model for each group is identified by setting the mean to zero and the standard deviations to one while all item parameters are estimated freely in each group. This results in group-wise estimated item loadings $\hat{\lambda}_{ig,0}$ and item intercepts $\hat{\nu}_{ig,0}$. For each group, the CFA model defined in Equation 1 can be equivalently written as

$$y_{ipg} = \underbrace{\nu_{ig} + \lambda_{ig}\alpha_g}_{=\nu_{ig,0}} + \underbrace{\lambda_{ig}\psi_g}_{=\lambda_{ig,0}}\eta_{pg}^{*} + \epsilon_{ipg} \quad , \quad \eta_{pg}^{*} \sim \mathrm{N}(0,1) \; , \tag{2}$$

In Equation 2, it can be seen that the estimated item parameters $(\lambda_{ig,0},\nu_{ig,0})$ are functions of the group-specific parameters (λ_{ig},ν_{ig}), group means α_g and standard deviations ψ_g.

In the second step, the alignment algorithm determines aligned means $\boldsymbol{\alpha}$ and standard deviations $\boldsymbol{\psi}$ by minimizing an alignment optimization function F. More specifically, the alignment function F minimizes deviations $\lambda_{ig} - \lambda_{ih} = \dfrac{\lambda_{ig,0}}{\psi_g} - \dfrac{\lambda_{ih,0}}{\psi_h}$ and deviations $\nu_{ig} - \nu_{ih} = \nu_{ig,0} - \dfrac{\lambda_{ig,0}}{\psi_g}\cdot\alpha_g - \nu_{ih,0} + \dfrac{\lambda_{ih,0}}{\psi_h}\cdot\alpha_h$, and is defined as[4,5]:

[4] It should be noted that $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ can be estimated by only using estimated intercepts $\hat{\boldsymbol{\nu}}$, i.e., by only considering the second term f_ν in Equation 3. However, in a preliminary simulation it turned out that this approach is less efficient than the approach using both terms, f_λ and f_ν.

[5] The optimization function in Equation 3 can be rewritten as two optimization problems that involve only the f_λ and f_ν terms, respectively. To this end, reparameterized parameters $(\tilde{\boldsymbol{\alpha}},\boldsymbol{\psi})$ are defined, where $\tilde{\alpha}_g = \alpha_g/\psi_g$. The optimization of $F(\tilde{\boldsymbol{\alpha}},\boldsymbol{\psi})$ can then be independently conducted for $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\psi}$.

$$F(\boldsymbol{\alpha},\boldsymbol{\psi}) = \sum_i \sum_{g<h} w_{igh} f_\lambda \left(\frac{\hat{\lambda}_{ig,0}}{\psi_g} - \frac{\hat{\lambda}_{ih,0}}{\psi_h} \right) + \sum_i \sum_{g<h} w_{igh} f_\nu \left(\hat{\nu}_{ig,0} - \frac{\hat{\lambda}_{ig,0}}{\psi_g}\alpha_g - \hat{\nu}_{ih,0} + \frac{\hat{\lambda}_{ih,0}}{\psi_h}\alpha_h \right) \quad , \tag{3}$$

where f_λ and f_ν are loss functions for item slopes and intercepts, respectively. The weights w_{igh} are often chosen as $w_{igh} = \sqrt{N_g N_h}$ (Asparouhov & Muthén, 2014; but see also Mansolf et al., 2020, for using different weights) to take uncertainty in item parameter estimates into account. However, it is also plausible to choose equal weights so that all groups contribute equally in the alignment optimization. For identification reasons, the product of standard deviations of all groups is set to one (i.e., $\prod_{g=1}^{G} \psi_g = 1$) and the mean α_g of the first group is set to zero (or the average of group means is set to zero, i.e., $\sum_{g=1}^{G} \alpha_g = 0$). Therefore, IA penalizes differences in item intercepts and item slopes between groups and, hence, minimizes the extent of measurement non-invariance. Given estimates of group means α_g and standard deviations ψ_g, aligned item ters $\hat{\lambda}_{ig}$ and $\hat{\nu}_{ig}$ can be calculated as:

$$\hat{\lambda}_{ig} = \frac{\hat{\lambda}_{ig,0}}{\psi_g} \quad \text{and} \quad \hat{\nu}_{ig} = \hat{\nu}_{ig,0} - \frac{\hat{\lambda}_{ig,0}}{\psi_g}\alpha_g$$

Muthén and Asparouhov (2014) proposed the same loss function for slopes and intercepts:

$$f_\lambda(x) = f_\nu(x) = \sqrt{|x|} = |x|^{1/2}$$

The loss functions can be generalized to be different for slopes and intercepts

$$f_\lambda(x) = |x|^{p_\lambda} \quad \text{and} \quad f_\nu(x) = |x|^{p_\nu} \, , \tag{4}$$

where powers p_λ and p_ν are nonnegative and define the shape of the loss function. This general form of the loss function was introduced in the R package sirt (Robitzsch, 2019). The default loss function used by Muthén and Asparouhov (2014) and implemented in the Mplus software (Muthén & Muthén, 1998-2017) is obtained with $p_\lambda = p_\nu = 0.5$, but optionally allows for the power value $p_\lambda = p_\nu = 0.25$. Note that the power values p_λ and p_ν govern the amount of admitted non-invariant item parameters in the alignment optimization. For low values of p like $p = 0.1$ or $p = 0.5$, the alignment function is optimized at a few large non-invariant item parameters and many approximately invariant parameters.

Hence, it mimics a situation of estimating a CFA model assuming partial invariant item parameters. In the limiting case of $p = 0$[6], the number of non-invariant item parameters is minimized (see Oelker, Pößnecker, & Tutz, 2015). A power of $p = 2$ corresponds to a least-squares estimation in which all parameter deviations equally contribute to the estimation of aligned means and standard deviations. Therefore, it will be more suitable for IA without large biases[7] and should provide results similar to the Bayesian structural equation modeling (BSEM) approach for invariance modeling (Muthén & Asparouhov, 2013).

Relationship of invariance alignment to IRT linking approaches

It is worth to mention that the ideas behind IA are not a completely novel contribution to the literature. The idea that a linear transformation of the ability scale could be used to link a number of test administrations (groups) using differences between item parameters from a measurement model (i.e., an item response model) has been established in the literature of linking or equating (Kolen & Brennan, 2014). Several methods were developed to estimate transformation parameters: the mean-sigma method (Marco, 1977), the mean-mean method (Loyd & Hoover, 1980), the Haebara (1980), and the Stocking-Lord (Stocking & Lord, 1983) approach. These approaches use the same principles of aligning item parameters for scale transformation, although they were originally designed for two groups only. Generalizations to multiple groups for linking scales were proposed (Arai & Mayekawa, 2011; Battauz, 2017; Haberman, 2009), which closely relate linking methods to IA.

Haberman linking (HL; Haberman, 2009) is a linking approach based on item loadings and item difficulties from multiple groups, which are obtained from separate calibrations. Based on these item parameters, HL is conducted in two steps. In the first step, item loadings are aligned, and in the second step, item difficulties are aligned. In each step, linear regression with estimated item parameters as the dependent variable is used to calculate joint item parameters referring to all groups (loadings or intercepts) and distribution parameters (means and standard deviations). The linear regression is estimated by ordinary least squares. As highlighted by researcher Matthias von Davier, Haberman linking is very similar to the IA approach (mentioned in Avvisati, Le Donné, & Paccagnella, 2019). The IA approach appears to be closest to IA in the case of $p = 2$, which corresponds to least squares estimation. However, both approaches make different assumptions about the residual effects (i.e., the differences in group-specific item parameters). For linking item loadings, IA with $p = 2$ minimizes discrepancies of the form

[6] By defining $0^0 = 0$.

[7] The definition of what is a large bias would always be somehow arbitrary and depends on the context. In this study, we are defining an intercept which is greater or equal to 0.3 as large bias.

$\left(\frac{\hat{\lambda}_{ig,0}}{\psi_g}-\frac{\hat{\lambda}_{ih,0}}{\psi_h}\right)^2$ for determining standard deviations $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_G)$, while HL[8] minimizes $\left(\log\frac{\hat{\lambda}_{ig,0}}{\psi_g}-\log\frac{\hat{\lambda}_{ih,0}}{\psi_h}\right)^2$. For linking item intercepts, IA with $p = 2$ minimizes discrepancies of the form $\left(\hat{\nu}_{ig,0}-\frac{\hat{\lambda}_{ig,0}}{\psi_g}\alpha_g-\hat{\nu}_{ih,0}+\frac{\hat{\lambda}_{ih,0}}{\psi_h}\alpha_h\right)^2$ for determining means $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_G)$, while HL[9] minimizes $\left(\frac{\hat{\psi}_g}{\hat{\lambda}_{ig,0}}\hat{\nu}_{ig,0}-\alpha_g-\frac{\hat{\psi}_h}{\hat{\lambda}_{ih,0}}\hat{\nu}_{ih,0}+\alpha_h\right)^2$ using estimated standard deviations $\hat{\boldsymbol{\psi}}=(\hat{\psi}_1,\ldots,\hat{\psi}_G)$ from the first step. Due to these different optimization functions, slightly different behavior in the performance of HL and IA approaches can be expected, but such an investigation is left to future research.

Estimation

The loss functions $f_\lambda(x)=f_\nu(x)=|x|^p$ are not differentiable, but are replaced in the estimation by differentiable approximating functions $f_D(x)=\left(x^2+\varepsilon\right)^{p/2}$ with a small $\varepsilon > 0$ (e.g., $\varepsilon = .01$)[10]. Because the function f_D is differentiable, quasi-Newton minimization approaches can be used that are implemented in standard optimizers in R (R Core Team, 2019). In our experience, in the case of small ε values, the optimization of the alignment function is very sensitive to starting values (see also Asparouhov & Muthén, 2014). Therefore, the implementation in the sirt (Robitzsch, 2019) package specifies a sequence of decreasing values of ε in the optimization, each ε using the previous solution as initial values (see Battauz, 2019, for a similar approach).

[8] Haberman (2009) decomposed logarithmized item loadings into common item loadings and group standard deviations in a linear model. By considering differences of logarithms of item loadings, common item loadings can be removed from estimation. This approach is used subsequently to illustrate the relationship of IA and HL.

[9] For linking item difficulties in the HL approach, the group-specific item difficulties $-\hat{\nu}_{ig}/\hat{\lambda}_{ig}$ are decomposed into common item difficulties and group means in a linear regression. Again, by considering differences of item difficulties, common item difficulties can be removed from estimation.

[10] An anonymous reviewer pointed out that Mplus uses $\varepsilon = .01$.

An illustrative example with different values of the power *p*

In order to illustrate the choice of different values of the power p in the alignment optimization, we consider a fictional example (see Table 1) consisting of six items and two groups. We only focus on the alignment of intercepts to show the general idea. It is assumed that the two groups do not differ in their mean. Hence, invariance alignment should recover this true mean difference of zero. The first four items do not differ in item intercepts between the two groups. However, items I5 and I6 have non-invariant item parameters where these items favor the second group.

Table 1:
Estimated Item Intercepts in Two Groups

Item	Group 1	Group 2	Diff.
I1	−1.5	−1.5	0
I2	−0.5	−0.5	0
I3	0.5	0.5	0
I4	1.5	1.5	0
I5	0.0	1.0	1
I6	1.0	2.0	1

Note: Diff. = difference between item intercepts

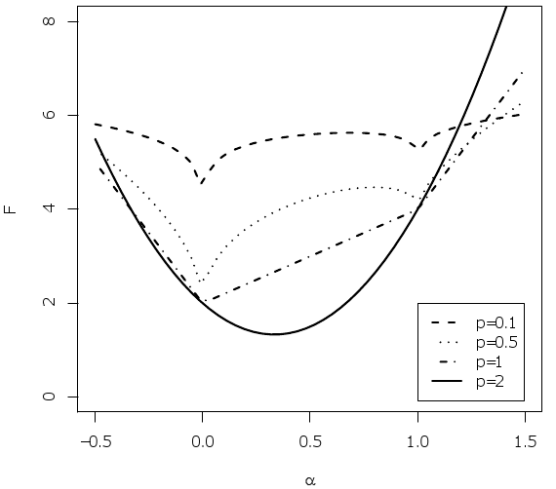

Figure 1:
Alignment Optimization Function as a Function of the Group Mean Difference α for Different Powers for the Illustrative Dataset

It is interesting to compare the performance of invariance alignment for different powers ($p = 0.1$, 0.5, 1, and 2) in this dataset. Figure 1 shows for the different powers p the alignment optimization function F as a function of the group mean difference α. With $p = 2$, a biased mean difference of $\alpha = .33$ is obtained because all differences between item intercepts contribute equally to the estimation of the group means difference. However, in the case of $p = 1$, the sum of absolute deviations is minimized, which results in the median difference of $\alpha = 0$ and an unbiased estimate of the group mean difference. The alignment estimate for $p = 0.1$ and $p = 0.5$ is also unbiased. However, as displayed in Figure 1, this toy dataset also shows that for powers smaller than 1, one additional local minimum at $\alpha = 1$ occurs. Thus, for small power values (i.e., $p < 1$), estimation issues might more frequently occur. However, the probability of estimating the global minimum instead of a local minimum can be substantially increased by using multiple starting values in the optimization as implemented in Mplus.

Monte Carlo simulation study

Previous research

Up-to-date, only a few studies investigated the performance of IA. Asparouhov and Muthén (2014) provided a small scale simulation using 5-item scale varying the sample size (100 or 1,000 per group), the number of groups (2, 3, 15 or 60), and the extent of non-invariance (0%, 10%, 20% of non-invariant items). AMI situations were not tested, but under the partial invariance conditions, the recovery of latent group means was very good even with 20% of non-invariant items. These results were confirmed by Flake and McCoach (2017), who used a two-factor model with 7 items per factor. Additionally, Marsh et al. (2018) showed that latent group means were more accurately estimated with alignment than with the scalar CFA-MI, and partial invariance models. Muthén and Asparouhov (2018) recommended alignment in a variety of situations, even with a small number of items. They concluded that the alignment procedure could be used with a small number of groups (even with two groups), but rather large sample sizes should be used. Moreover, according to their recommendations, IA is suitable only for partial invariance patterns, i.e., the majority of the parameters are invariant and a minority of the parameters non-invariant.

Pokropek, Davidov, and Schmidt (2019) tested different measurement invariance situations (including AMI) using 3-, 4- and 5-items scales, 24 groups, and 1000 observations. In general, they confirmed the excellent performance of IA under partial non-invariance. However, the performance of the IA under the AMI situation was not substantially superior to a simple scalar CFA model that just ignores the non-invariance problem. Similarly, the combination of AMI and partial invariance leads to conditions in which precise recovery of latent group parameters is very problematic.

Some additional studies investigated the detection of non-invariant items by using the IA in Mplus (DeMars, 2020; Finch, 2016; Kim et al., 2017) and showed both acceptable type I error rates and power rates. However, the present article focuses on the estimation

of group mean differences under IA and will not address the detection of non-invariant items.

Overall, previous simulation research was restricted to a limited number of conditions. Mostly, short scales were tested, and partial non-invariance was of primary interest. Moreover, all of the previous studies were examining only the loss function originally proposed by Muthén and Asparouhov (2014). The present simulation study provides a more comprehensive evaluation of the performance of the IA approach using different forms of the loss function.

Design of simulation

We focus on three basic and most common situations in social sciences: a small (Study 1), a medium (Study 2), and a large (Study 3) number of groups. Study 1 focuses on two groups and sample sizes from 100 to 1000 persons per group. This scenario would mimic small-scale psychological studies, field trial studies, but also a situation where measurement invariance is tested in a two-group case (like gender) in a large national representative sample. Monte Carlo simulation studies that investigate the performance of IA with two groups were conducted by Asparouhov and Muthén (2014), DeMars (2020), and Finch (2016). In this article, however, we extend the settings and test IA with different specifications. Study 2 focuses on 4 groups and sample sizes from 100 to 2000 per group. This situation might reflect a large-scale national study in which 4 groups are

Table 2:
Design of simulation study

Study	Questionnaire (5 items)	Test (20 items)
Study 1 (2 groups; sample size 100 to 1000 per group)	**2 N-I items in 1 group:** 1) Size of N-I (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)	**8 N-I items in 1 group:** 1) Size of N-I (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)
Study 2 (4 groups; sample size per group: 100, 200, 500, 1000, 2000)	**2 N-I items in each of 2 groups:** 1) Bias of N-I items (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)	**8 N-I items in each of 2 groups:** 1) Size of N-I (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)
Study 3 (25 groups; sample size 1000 per group)	**2 N-I items in each of 0, 5, 15 of groups** 1) Size of N-I (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)	**8 N-I items in each of 0, 5, 15 of groups** 1) Size of N-I (0, 0.3, 0.6, 0.9); 2) Approximate (0.001, 0.005, 0.010, 0.05) * N-I items (0, 0.6)

Note: N-I Non-Invariant

compared, a field trial of a larger international study or an evaluation that wants to compare a limited number of countries. Study 3 mimics large scale international surveys like the European Social Survey (ESS), the World Values Survey (WVS), or large-scale assessments like the Programme for the International Assessment of Adult Competencies (PIAAC) or the Programme for International Student Assessment (PISA) with 25 groups and sample sizes larger than 1000.

In each situation, we investigate two lengths of the scales: 5 items and 20 items. The scale with 5 items reflects a typical questionnaire scale (e.g., Math anxiety, Cultural possessions at home index, Enjoyment of science index in PISA; Advanced Reading at home, Advanced Reading at work in PIAAC or Schwartz's Human Values that are usually measured by 4 to 6 items). The scale with 20 items reflects a psychological or cognitive test.

An overview of the simulation designs is presented in Table 2. In all scenarios, we limited the number of non-invariant items to 40%. In Study 1 and in Study 2, half of the groups were affected by non-invariant items, while in Study 3, the number of countries that were affected by non-invariant items varied depending on the scenario.

Data generating procedures and analysis

In all conditions, data were generated using the same procedure. The true means for the groups were assigned as a sequence of equally spaced values between −1 and 1. The SDs for the groups were generated in a similar manner but with a minimum value of 0.8 and a maximum of 1.2. The means and SDs were matched randomly for each group.

For simplicity, we consider a situation in which all slopes (λ) were set to 1. The intercept parameters (τ) were generated using the same rule as it was used for the group means and were assigned as a sequence of equally spaced values between −1 and 1. The values of the item parameters were initially set to be equal across groups. Then, depending on the scenario, different invariance situations were generated. For the typical non-invariant N-I situation, biases were assigned to successive even groups starting from the second group. The direction of the bias was generated in such a way that it has the same sign in the group, while on average, it tends to be balanced between groups. For instance, in the 4-group scenario with 5 items, the first group has no biased items, the second has only positive biases, the third has no biased items, and the fourth group has only negative biases.

Approximate non-invariance was implemented by adding random values generated from a standard normal distribution with mean zero and variance defined by the level of approximate invariance of differences between item parameters. In other words, for achieving approximate non-invariance of item parameters that would be characterized by the expected variance of the differences between parameters of 0.05, random biases for all item parameters were sampled from a standard normal distribution with mean zero and variance of 0.025.

In each scenario, 400 replications were generated. Once the data were generated, six models were estimated for each generated dataset: the scalar CFA invariance model in

which all item parameters were fixed to be the same among groups, and five CFA invariance alignment models, two models with a power of $p = 0.5$ that were specified in Mplus and sirt, and three alignment models with $p = 0.1$, $p = 1$, and $p = 2$ that were all specified in sirt.

One should notice that in some settings, the estimating models are not fully corresponding with the true generating models. This choice was made on purpose because we believe that it is useful to assess how models perform in different settings and to check the robustness of the findings for different data scenarios. This strategy was also applied in numerous simulation studies for checking the behavior of different models that do not match the true population model (see Nylund-Gibson & Masyn, 2016; Muthén & Kaplan, 1985; Muthén & Asparouhov, 2008; Rhemtulla, Brosseau-Liard, & Savalei, 2012, for some examples).

Performance measures

For each model, we investigated the average absolute bias and accuracy of the estimated latent group means. The average absolute bias reports the average absolute differences between an estimated mean by a particular method and a true group mean used to generate the data:

$$Bias\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{G}\sum_{g=1}^{G}\left|\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{gr} - \theta_g\right| \ ,$$

where R is the total number of replications, G is the number of groups, and $\hat{\theta}_{gr}$ is the group mean estimate of group g in the r^{th} replication.

Accuracy is a combination of bias and variability that quantifies the overall performance of an estimator. The more biased and the less precise an estimator is, the worse is its accuracy. In this study, we use the average root mean square error (RMSE) for assessing accuracy:

$$RMSE\left(\hat{\boldsymbol{\theta}}\right) = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{gr} - \theta_g\right)^2\right)}$$

Results

Study 1

Questionnaire case (5 items)

In Figure 2, absolute bias (left panel), and accuracy measured by RMSE (right) are presented. This condition reflects the situation of a small study with two groups, each with 100 respondents, and a short 5-item scale. In this scenario, two items are non-invariant in each group, and the figure displays results with different sizes of DIF. Both absolute bias

and RMSE increase substantially with a higher bias of the non-invariant items. However, the most crucial observation is that all alignment models perform worse than the scalar model that simply ignores the problem of non-invariance. The quality of mean recovery is related to the number of estimated parameters (the scalar model is estimating with fewer parameters), and this might cause the better performance of the scalar model in the presented situations.

Figure 3 shows that conditions with a small sample size are problematic, resulting in a less accurate recovery of group means. It is evident that a sample size of 100 is too small to produce reliable estimates in the alignment approach. There is a high increase of accu-

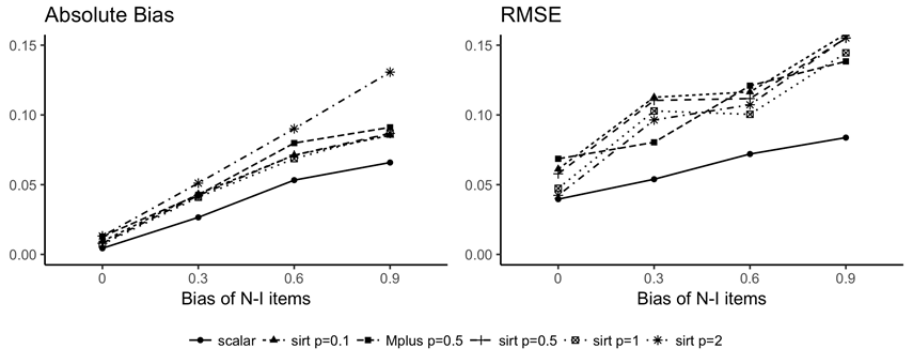

Figure 2:
Results of a simulation study for small scale study reflecting questionnaire scale (5 items; 100 individuals per group; 2 groups). Condition: size of DIF (two non-invariant items for one group)

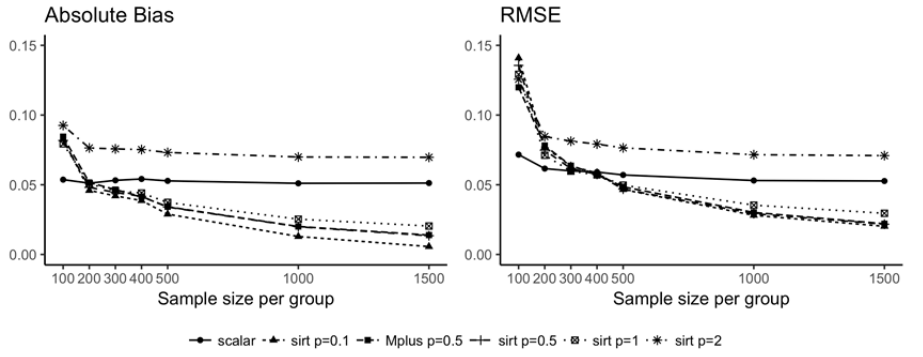

Figure 3:
Results of a simulation study for small scale study reflecting questionnaire scale (5 items; 2 groups, two non-invariant items in one group with bias=0.6). Condition: size of the group.

racy in the IA approach when the sample size is increased to 200, but only with the sample size of 500, both RMSE and absolute bias of alignment procedure (with powers 0.1, 0.5 and 1) are substantially lower than the RMSE and absolute bias of the scalar model. In the scenario with a majority of invariant items and a minority of non-invariant items, a power of $p = 0.1$ is most effective, produces smaller absolute bias and RMSE, although the results of models with the power of $p = 0.5$ are very close to it. A power of $p = 1$ gives an only slightly worse recovery of the group means. However, IA that uses a power of $p = 2$ in the loss function gives much worse results even compared to the scalar model.

In Table 3, we present results of a simulation where approximate invariance or approximate non-invariance (AN-I) was applied to all items. The following factors were varied: three sample sizes of one group, 100, 500, and 1000, as well as four sizes of AN-I. 0.001, 0.005, 0.010, and 0.050. Similar to the results presented in Figure 2 and Figure 3, results in Table 3 indicate that the sample size of 100 is too small for the IA approach to beat the results of the scalar model. With higher sample sizes, the results of the IA approach

Table 3:
Bias and RMSE for a scenario with 5 items, two groups. Different levels of approximate non-invariance (AN-I). No N-I items with large biases. Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
100	scalar model	**0.002**	**0.004**	**0.005**	**0.008**	**0.041**	**0.041**	**0.039**	**0.048**
	sirt $p = 0.1$	0.022	0.020	0.020	0.025	0.072	0.071	0.067	0.078
	Mplus $p = 0.5$	0.008	0.011	0.009	0.017	0.062	0.069	0.063	0.074
	sirt $p = 0.5$	0.022	0.019	0.019	0.022	0.069	0.068	0.063	0.074
	sirt $p = 1$	0.020	0.017	0.016	0.019	0.056	0.055	0.052	0.062
	sirt $p = 2$	0.017	0.015	0.013	0.015	0.046	0.048	0.042	0.051
500	scalar model	**0.001**	**0.000**	**0.002**	0.007	**0.019**	**0.020**	**0.021**	**0.030**
	sirt $p = 0.1$	0.004	0.003	0.002	0.004	0.025	0.026	0.027	0.041
	Mplus $p = 0.5$	0.003	0.003	0.002	**0.002**	0.023	0.025	0.025	0.038
	sirt $p = 0.5$	0.004	0.003	0.002	0.003	0.024	0.025	0.025	0.038
	sirt $p = 1$	0.004	0.004	0.002	0.002	0.021	0.023	0.022	0.033
	sirt $p = 2$	0.004	0.004	0.002	0.002	0.020	0.021	0.021	0.030
1000	scalar model	**0.000**	**0.001**	**0.002**	0.005	**0.013**	**0.015**	**0.017**	**0.029**
	sirt $p = 0.1$	0.000	0.002	0.000	0.002	0.016	0.020	0.022	0.038
	Mplus $p = 0.5$	0.000	0.002	0.000	0.002	0.015	0.018	0.021	0.036
	sirt $p = 0.5$	0.000	0.002	0.000	**0.002**	0.015	0.018	0.020	0.036
	sirt $p = 1$	0.001	0.002	0.001	0.002	0.014	0.017	0.019	0.033
	sirt $p = 2$	0.001	0.001	0.001	0.002	0.013	0.015	0.017	0.030

significantly improved but are still not as good as the results of the scalar model. The IA approach that comes closest to the scalar model in terms of accuracy is the alignment procedure with a power of 2.

Table 4 shows results where AN-I is crossed with large biases in two items. Not surprisingly, for a sample size of 100, the scalar model outperformed all IA approaches. For higher sample sizes, and size of AN-I of at most 0.01, the lowest absolute bias and RMSE were produced by the IA approach with $p = 0.1$. Both Mplus and sirt provided similar results for $p = 0.5$ with a slight advantage of sirt. Interestingly, for a high level on AN-I (i.e., 0.05), the results of the scalar model remain most accurate.

Table 4:
Bias and RMSE for a scenario with 5 items, two groups. Two N-I items in one group with bias 0.6 Different level of approximate non-invariance (AN-I). Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
100	scalar model	**0.054**	**0.048**	**0.049**	**0.042**	**0.071**	**0.067**	**0.069**	**0.069**
	sirt $p = 0.1$	0.095	0.091	0.098	0.101	0.132	0.131	0.143	0.166
	Mplus $p = 0.5$	0.087	0.083	0.089	0.085	0.120	0.119	0.130	0.133
	sirt $p = 0.5$	0.093	0.089	0.098	0.099	0.128	0.123	0.141	0.159
	sirt $p = 1$	0.089	0.086	0.093	0.095	0.112	0.111	0.128	0.148
	sirt $p = 2$	0.090	0.086	0.093	0.092	0.104	0.100	0.122	0.126
500	scalar model	0.053	0.050	0.052	**0.048**	0.057	0.054	**0.058**	**0.061**
	sirt $p = 0.1$	**0.030**	**0.033**	**0.042**	0.057	**0.049**	0.052	0.064	0.086
	Mplus $p = 0.5$	0.035	0.036	0.045	0.059	0.051	0.052	0.063	0.084
	sirt $p = 0.5$	0.034	0.036	0.045	0.059	0.050	**0.051**	0.061	0.083
	sirt $p = 1$	0.043	0.043	0.051	0.062	0.052	0.053	0.062	0.081
	sirt $p = 2$	0.074	0.071	0.074	0.076	0.077	0.074	0.078	0.085
1000	scalar model	0.052	0.052	0.051	**0.046**	0.055	0.055	0.055	**0.058**
	sirt $p = 0.1$	**0.016**	**0.023**	**0.029**	0.054	**0.031**	**0.039**	**0.046**	0.084
	Mplus $p = 0.5$	0.022	0.029	0.034	0.056	0.033	0.040	0.047	0.081
	sirt $p = 0.5$	0.022	0.029	0.033	0.055	0.032	0.040	0.046	0.078
	sirt $p = 1$	0.032	0.038	0.042	0.058	0.039	0.045	0.050	0.077
	sirt $p = 2$	0.071	0.071	0.071	0.074	0.073	0.074	0.074	0.082

Long scale (20 items)

Let us now turn to a situation with a longer scale that might reflect more refined psychological scales or cognitive tests in the case of a small sample size of 100. In Figure 4, we present results where the size of the N-I bias is manipulated. With such a sample size, results of the IA approach were very similar to the scalar model, and only with a high bias of 0.9, IA with $p = 0.1$ and $p = 0.5$ produced slightly lower absolute bias and RMSE than the scalar model.

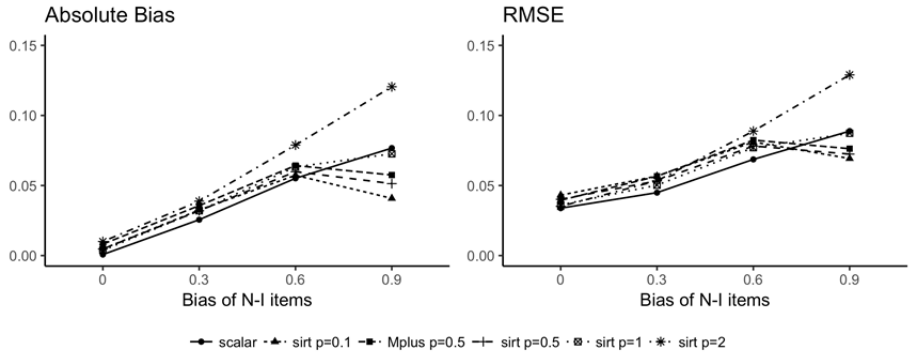

Figure 4:
Results of a simulation study for small scale study reflecting test scale (20 items; 100 individuals per group; 2 groups). Condition: size of DIF (eight non-invariant items for one group).

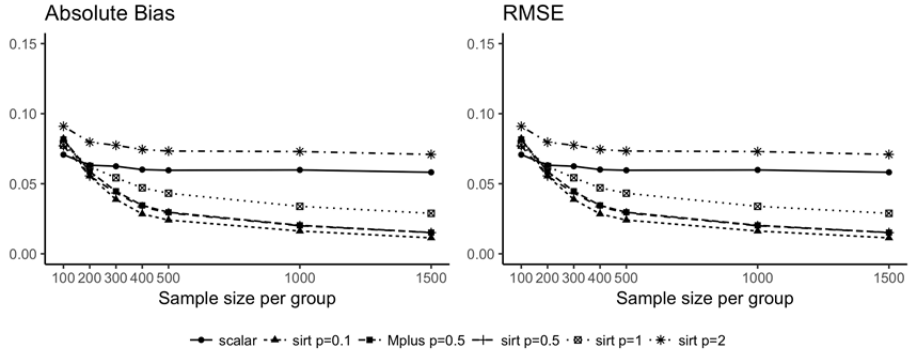

Figure 5:
Results of a simulation study for small scale study reflecting questionnaire scale (20 items; 2 groups, eight non-invariant items in one group with bias=0.6). Condition: size of the group.

In Figure 5, we varied the sample sizes of the groups and fixed the size of N-I bias to 0.6. Clearly, IA (with powers .1, .5, and 1) needs at least 200 to 300 observations per group to show superior performance to the scalar model that ignores non-invariance. Alignment with a power $p = 0.1$ performed best in terms of absolute bias and RMSE, but the procedure with the power of 0.5 provided very similar results. Alignment with a power of 1 gives slightly worse results than $p = 0.1$ and $p = 0.5$ but was still superior to the scalar model when the sample size per group exceeded 200.

The approximate non-invariance for the 20-items scenario with varying sample sizes is depicted in Table 5. In contrast to the 5-items scale, the IA approach for the 20-item scale could mitigate even a large size of approximate N-I using simply a scalar model. In fact, the differences between the recovery of group means for the scalar model and different settings of alignment were not substantial, especially with sample sizes 500 and 1000.

Table 5:
Bias and RMSE for a scenario with 5 items, two groups. Different level of approximate non-invariance (AN-I). No N-I items with large biases. Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
100	scalar model	**0.001**	**0.001**	**0.001**	**0.005**	**0.035**	**0.035**	**0.035**	**0.037**
	sirt $p = 0.1$	0.010	0.010	0.013	0.011	0.042	0.044	0.045	0.048
	Mplus $p = 0.5$	0.008	0.008	0.011	0.008	0.039	0.041	0.041	0.043
	sirt $p = 0.5$	0.010	0.010	0.013	0.011	0.039	0.041	0.042	0.043
	sirt $p = 1$	0.010	0.010	0.013	0.010	0.037	0.038	0.039	0.039
	sirt $p = 2$	0.010	0.010	0.013	0.009	0.036	0.036	0.037	0.038
500	scalar model	**0.001**	**0.001**	**0.001**	**0.003**	**0.015**	**0.015**	**0.016**	**0.018**
	sirt $p = 0.1$	0.002	0.003	0.002	0.005	0.016	0.018	0.019	0.025
	Mplus $p = 0.5$	0.001	0.003	0.001	0.004	0.015	0.017	0.018	0.023
	sirt $p = 0.5$	0.002	0.003	0.002	0.005	0.015	0.017	0.018	0.023
	sirt $p = 1$	0.002	0.003	0.002	0.004	0.015	0.016	0.017	0.020
	sirt $p = 2$	0.002	0.003	0.002	0.003	0.015	0.016	0.016	0.019
1000	scalar model	**0.000**	**0.000**	0.002	**0.002**	**0.011**	**0.012**	**0.012**	**0.016**
	sirt $p = 0.1$	0.001	0.002	0.000	0.003	0.012	0.013	0.013	0.020
	Mplus $p = 0.5$	0.001	0.001	0.000	0.002	0.011	0.013	0.013	0.018
	sirt $p = 0.5$	0.001	0.002	0.000	0.002	0.011	0.013	0.013	0.018
	sirt $p = 1$	0.001	0.002	0.000	0.002	0.011	0.012	0.012	0.017
	sirt $p = 2$	0.001	0.002	**0.000**	0.003	0.011	0.012	0.012	0.016

When AN-I is crossed with large biases in eight items, results mimic the conditions of those presented in Table 4. For a sample size of 100, the scalar model outperforms all IA approaches. For higher sample sizes and sizes of AN-I up to 0.01, the lowest absolute bias and RMSE were produced by alignment procedures with $p = 0.1$. For a high level of AN-I (i.e., 0.05), the results of the scalar model remained most accurate. We are not presenting the detailed table for this condition, but the results are available under request.

Study 2

Questionnaire case (5 items)

Table 6 shows the performance of the investigated methods with different sizes of non-invariance and different sample sizes. Not surprisingly, the scalar model works best in situations when there are no non-invariant items (i.e., the scalar model conforms to the data generating model). It is important to emphasize that in a situation of full invariance, IA produced results that were close to the scalar model, at least for large sample sizes (very close for a sample size of $N \geq 500$). When there are some non-invariant items, the best recovery of group means is achieved by IA with a power of $p = 0.5$, or, in most situations, it was even outperformed with the power of $p = 0.1$.

Results of the recovery of group means in situations where different levels of AN-I are applied appear to be very similar to results displayed in Table 5. It turned out that in the presence of AN-I, the scalar model outperformed the IA approach. The advantage of the scalar model over the IA approach was higher for larger sample sizes. For smaller sample sizes, the advantages are clearly visible, while for sample sizes of at least 500, the advantage of the scalar model was less noticeable. The detailed results are available under request.

In Table 7, we show results for the condition in which, in addition to AN-I, there are two N-I items with large biases (0.6) in two groups. In this situation, except for a small sample size of $N = 100$, the IA approach resulted in better recovery of group means. In most situations, IA with a power of $p = 0.1$ performed best. For sample sizes of at least 500 and an amount of AN-I variance of 0.050, IA with a power of $p = 0.5$ performed best with a slight advantage of Mplus in favor of sirt.

Long scale (20 items)

Table 8 reports simulation results for a 20-item scale for three levels of N-I. The pattern of findings is evident and very similar to earlier analyses. Overall, the scalar model performed best for data without N-I items, while in most other situations, IA with a power of $p = 0.1$ performed best.

We also analyzed conditions with different levels of AN-I. We are not presenting the detailed results here (which are available upon request) because they can be straightforwardly described in a few words. The scalar model slightly outperformed the alignment

Table 6:
Bias and RMSE for a scenario with 5 items, four groups. Different levels of non-invariance (AN-I). Two N-I items with large biases in two groups. Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of N-I Item Bias:				Size of N-I Item Bias:			
		0.0	0.3	0.6	0.9	0.0	0.3	0.6	0.9
100	scalar model	**0.001**	**0.060**	0.115	0.163	**0.072**	**0.104**	**0.153**	0.204
	sirt $p = 0.1$	0.018	0.061	**0.067**	**0.066**	0.095	0.137	0.160	0.184
	Mplus $p = 0.5$	0.016	0.063	0.076	0.071	0.088	0.126	0.153	**0.167**
	sirt $p = 0.5$	0.017	0.063	0.077	0.075	0.088	0.127	0.153	0.172
	sirt $p = 1$	0.018	0.067	0.098	0.112	0.079	0.117	0.150	0.173
	sirt $p = 2$	0.021	0.071	0.129	0.187	0.076	0.115	0.170	0.229
200	scalar model	**0.001**	0.060	0.115	0.162	**0.053**	**0.089**	0.141	0.187
	sirt $p = 0.1$	0.007	**0.041**	**0.032**	**0.019**	0.065	0.098	0.100	**0.081**
	Mplus $p = 0.5$	0.008	0.046	0.039	0.031	0.059	0.092	**0.093**	0.084
	sirt $p = 0.5$	0.008	0.046	0.040	0.031	0.060	0.092	0.094	0.084
	sirt $p = 1$	0.009	0.055	0.076	0.080	0.056	0.090	0.111	0.117
	sirt $p = 2$	0.011	0.066	0.128	0.187	0.054	0.096	0.155	0.213
500	scalar model	**0.000**	0.059	0.115	0.162	**0.032**	0.076	0.129	0.176
	sirt $p = 0.1$	0.003	**0.021**	**0.010**	**0.009**	0.038	0.059	**0.048**	**0.045**
	Mplus $p = 0.5$	0.003	0.028	0.020	0.017	0.035	**0.057**	0.050	0.046
	sirt $p = 0.5$	0.003	0.027	0.019	0.016	0.036	0.058	0.049	0.046
	sirt $p = 1$	0.003	0.044	0.052	0.053	0.034	0.065	0.074	0.074
	sirt $p = 2$	0.004	0.063	0.125	0.185	0.033	0.080	0.140	0.200
1000	scalar model	**0.001**	0.060	0.114	0.162	**0.023**	0.071	0.123	0.170
	sirt $p = 0.1$	0.001	**0.013**	**0.005**	**0.007**	0.026	**0.036**	**0.031**	**0.030**
	Mplus $p = 0.5$	0.001	0.021	0.014	0.011	0.024	0.040	0.034	0.032
	sirt $p = 0.5$	0.001	0.019	0.012	0.011	0.024	0.039	0.033	0.032
	sirt $p = 1$	0.001	0.038	0.041	0.041	0.024	0.052	0.055	0.056
	sirt $p = 2$	0.002	0.063	0.125	0.185	0.023	0.074	0.134	0.195
2000	scalar model	**0.000**	0.060	0.114	0.162	**0.016**	0.066	0.119	0.167
	sirt $p = 0.1$	0.001	**0.009**	**0.004**	**0.004**	0.017	**0.024**	**0.020**	**0.020**
	Mplus $p = 0.5$	0.001	0.017	0.011	0.009	0.016	0.029	0.024	0.022
	sirt $p = 0.5$	0.001	0.015	0.009	0.008	0.016	0.028	0.023	0.022
	sirt $p = 1$	0.001	0.034	0.035	0.035	0.016	0.043	0.044	0.045
	sirt $p = 2$	0.001	0.063	0.124	0.185	0.016	0.069	0.130	0.191

Table 7:
Bias and RMSE for a scenario with 5 items, four groups. Different levels of approximate non-invariance (AN-I) combined with two N-I items with large biases (0.6) in two groups. Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
100	scalar model	0.115	0.117	0.115	0.119	0.155	0.158	**0.160**	**0.168**
	sirt $p = 0.1$	**0.072**	**0.075**	**0.077**	**0.100**	0.160	0.164	0.168	0.204
	Mplus $p = 0.5$	0.078	0.082	0.091	0.107	0.150	0.154	0.166	0.195
	sirt $p = 0.5$	0.079	0.084	0.092	0.108	0.151	0.155	0.169	0.197
	sirt $p = 1$	0.102	0.103	0.108	0.118	**0.150**	**0.152**	0.161	0.181
	sirt $p = 2$	0.134	0.131	0.134	0.135	0.172	0.173	0.178	0.182
200	scalar model	0.112	0.114	0.116	0.115	0.137	0.140	0.144	0.155
	sirt $p = 0.1$	**0.028**	**0.034**	**0.038**	**0.076**	**0.066**	**0.068**	**0.073**	**0.103**
	Mplus $p = 0.5$	0.042	0.042	0.049	0.078	0.096	0.100	0.112	0.160
	sirt $p = 0.5$	0.041	0.042	0.048	0.079	0.096	0.100	0.111	0.159
	sirt $p = 1$	0.073	0.073	0.081	0.096	0.110	0.113	0.123	0.151
	sirt $p = 2$	0.126	0.125	0.127	0.126	0.152	0.153	0.158	0.166
500	scalar model	0.114	0.116	0.115	0.118	0.127	0.132	0.133	0.150
	sirt $p = 0.1$	**0.011**	**0.016**	**0.019**	**0.056**	**0.049**	**0.059**	**0.067**	0.139
	Mplus $p = 0.5$	0.020	0.027	0.031	0.067	0.051	0.061	0.070	**0.134**
	sirt $p = 0.5$	0.019	0.026	0.031	0.068	0.051	0.061	0.072	0.136
	sirt $p = 1$	0.054	0.059	0.064	0.090	0.076	0.083	0.092	0.136
	sirt $p = 2$	0.125	0.127	0.126	0.126	0.139	0.143	0.144	0.158
1000	scalar model	0.114	0.115	0.114	0.117	0.122	0.126	0.127	0.147
	sirt $p = 0.1$	**0.008**	**0.011**	**0.016**	**0.056**	**0.033**	**0.043**	**0.055**	0.131
	Mplus $p = 0.5$	0.014	0.020	0.023	0.062	0.035	0.045	0.056	**0.125**
	sirt $p = 0.5$	0.013	0.018	0.022	0.062	0.035	0.045	0.056	0.126
	sirt $p = 1$	0.042	0.049	0.054	0.085	0.057	0.068	0.077	0.129
	sirt $p = 2$	0.125	0.125	0.124	0.124	0.134	0.137	0.138	0.154
2000	scalar model	0.115	0.114	0.117	0.116	0.120	0.123	0.128	0.142
	sirt $p = 0.1$	**0.005**	**0.007**	**0.012**	**0.053**	0.033	0.043	0.055	0.131
	Mplus $p = 0.5$	0.013	0.016	0.024	0.056	0.027	0.038	**0.050**	**0.121**
	sirt $p = 0.5$	0.012	0.015	0.023	0.057	**0.026**	**0.037**	0.050	0.123
	sirt $p = 1$	0.038	0.042	0.053	0.080	0.048	0.058	0.072	0.122
	sirt $p = 2$	0.126	0.125	0.126	0.122	0.131	0.134	0.138	0.149

Table 8:
Bias and RMSE for a scenario with 20 items, four groups. Different levels of approximate non-invariance (AN-I). Eight N-I items with large biases in two groups. Smallest average absolute bias and RMSE for each condition were bolded.

		Average Absolute Bias				**RMSE**			
Sample size	Model	Size of N-I Item Bias:				Size of N-I Item Bias:			
		0.0	0.3	0.6	0.9	0.0	0.3	0.6	0.9
100	scalar model	**0.003**	0.060	0.119	0.168	**0.063**	**0.098**	0.152	0.201
	sirt $p = 0.1$	0.013	**0.055**	**0.042**	**0.026**	0.069	0.104	**0.099**	**0.081**
	Mplus $p = 0.5$	0.009	0.059	0.061	0.049	0.066	0.102	0.106	0.092
	sirt $p = 0.5$	0.006	0.058	0.057	0.045	0.067	0.102	0.105	0.092
	sirt $p = 1$	0.007	0.063	0.094	0.103	0.065	0.103	0.132	0.140
	sirt $p = 2$	0.014	0.069	0.129	0.190	0.065	0.107	0.164	0.223
200	scalar model	**0.001**	0.060	0.117	0.168	**0.048**	0.084	0.139	0.187
	sirt $p = 0.1$	0.013	0.055	0.042	**0.026**	0.049	**0.071**	**0.058**	**0.058**
	Mplus $p = 0.5$	0.006	0.044	0.035	0.029	0.049	0.075	0.067	0.061
	sirt $p = 0.5$	0.004	**0.042**	**0.034**	0.027	0.049	0.074	0.066	0.060
	sirt $p = 1$	0.005	0.054	0.073	0.076	0.048	0.081	0.098	0.100
	sirt $p = 2$	0.008	0.064	0.126	0.186	0.048	0.090	0.149	0.207
500	scalar model	**0.001**	0.060	0.116	0.166	**0.029**	0.074	0.128	0.177
	sirt $p = 0.1$	0.003	**0.018**	**0.010**	**0.007**	0.030	**0.039**	**0.035**	**0.034**
	Mplus $p = 0.5$	0.002	0.029	0.018	0.015	0.029	0.048	0.039	0.036
	sirt $p = 0.5$	0.002	0.027	0.017	0.014	0.029	0.046	0.038	0.035
	sirt $p = 1$	0.002	0.045	0.050	0.051	0.029	0.060	0.066	0.066
	sirt $p = 2$	0.003	0.063	0.124	0.184	0.029	0.077	0.137	0.196
1000	scalar model	**0.000**	0.060	0.115	0.167	**0.021**	0.068	0.123	0.174
	sirt $p = 0.1$	0.001	**0.012**	**0.007**	**0.004**	0.022	**0.026**	**0.024**	**0.021**
	Mplus $p = 0.5$	0.001	0.021	0.013	0.012	0.021	0.033	0.027	0.027
	sirt $p = 0.5$	0.001	0.019	0.012	0.011	0.021	0.032	0.026	0.026
	sirt $p = 1$	0.001	0.038	0.039	0.042	0.021	0.048	0.050	0.052
	sirt $p = 2$	0.002	0.062	0.123	0.185	0.021	0.071	0.131	0.193
2000	scalar model	**0.000**	0.061	0.117	0.168	0.015	0.066	0.121	0.172
	sirt $p = 0.1$	0.001	**0.008**	**0.004**	**0.003**	**0.014**	**0.019**	**0.016**	**0.016**
	Mplus $p = 0.5$	0.001	0.017	0.011	0.009	0.015	0.025	0.021	0.019
	sirt $p = 0.5$	0.000	0.015	0.010	0.008	0.015	0.024	0.020	0.018
	sirt $p = 1$	0.000	0.033	0.035	0.035	0.015	0.040	0.041	0.042
	sirt $p = 2$	0.001	0.063	0.125	0.185	0.015	0.068	0.129	0.190

procedure, especially for small sample sizes. While the biases were approximately zero for all powers in the IA approach, the RMSE was lowest for $p = 2$ and close to the RMSE in the scalar model.

The final results for Study 2 refer to the situation where AN-I biases were combined with large biases of 0.6 in eight items from two out of four groups. In those settings, IA with $p = 0.1$ clearly produced the highest accuracy of group means producing lowest biases and highest RMSE for all sample sizes and conditions except for sample size 100 and the level of AN-I of 0.05 where RMSE is slightly lower for $p = 0.05$. In all other situations, the $p = 0.05$ gives the second-best results. The worst recovery was obtained with $p = 2$ where recovery of the means is even slightly worse than for the scalar model. We are not providing detailed results, but the results could be obtained upon request.

Study 3

Questionnaire case (5 item)

In Study 3, we consider a situation typical for large scale-scale comparative studies, that is a large number of groups (i.e., 25) and large sample sizes (i.e., 1000). First, we analyze a 5-item scale. In Table 9, the results of a simulation study with no N-I item or two N-I

Table 9:
Average absolute bias and RMSE for a scenario with 5 items, 25 groups, 1000 observations per group,and 2 N-I items with large biases (in 5 and 15 groups). Smallest average absolute bias and RMSE for each condition were bolded.

N affected groups	Model	**Average Absolute Bias**				**RMSE**			
		Size of N-I Item Bias:				Size of N-I Item Bias:			
		0.0	0.3	0.6	0.9	0.0	0.3	0.6	0.9
5	scalar model	**0.002**	0.027	0.052	0.073	**0.030**	0.048	0.070	0.089
	sirt $p = 0.1$	0.003	**0.015**	**0.006**	**0.005**	0.031	**0.040**	**0.033**	**0.032**
	Mplus $p = 0.5$	0.003	0.019	0.025	0.022	0.031	0.041	0.049	0.051
	sirt $p = 0.5$	0.003	0.019	0.011	0.009	0.031	0.041	0.037	0.034
	sirt $p = 1$	0.003	0.023	0.027	0.027	0.031	0.044	0.047	0.047
	sirt $p = 2$	0.003	0.029	0.057	0.084	0.031	0.050	0.074	0.099
15	scalar model	**0.002**	0.074	0.152	0.254	**0.030**	0.086	0.160	0.260
	sirt $p = 0.1$	0.002	**0.012**	**0.007**	**0.006**	0.030	**0.038**	**0.035**	**0.034**
	Mplus $p = 0.5$	0.003	0.024	0.016	0.014	0.031	0.044	0.038	0.037
	sirt $p = 0.5$	0.003	0.021	0.014	0.012	0.031	0.042	0.037	0.036
	sirt $p = 1$	0.003	0.043	0.045	0.046	0.031	0.058	0.060	0.061
	sirt $p = 2$	0.003	0.076	0.150	0.223	0.031	0.087	0.157	0.228

items in 5 and 15 affected groups (out of 25 groups) are presented. Additionally, different sizes of N-I biases were considered: 0.3, 0.6, and 0.9. Results clearly show that in a situation where all items were invariant, the scalar model performed best. The IA approach, however, did not significantly differ in performance. This means that applying IA to the data without N-I items did not affect results in a significant way. On the other hand, IA optimization performed much better than the scalar model when N-I items are present in the data. The best mean recovery in terms of average absolute bias and RMSE was achieved with a power $p = 0.1$.

In Table 10, the results of a simulation study for different numbers of affected groups of N-I items with a large bias of 0.6 and different values of AN-I variance are presented. The general pattern is apparent. For AMI scenarios without N-I items (i.e., no affected groups), the scalar model performed best. When some groups are affected by N-I items,

Table 10:
Average absolute bias and RMSE for a scenario with 5 items, 25 groups, 1000 observations per group, and different levels of AN-I and two N-I items with large biases (0.6 in 0, 5 and 15 groups). Smallest average absolute bias and RMSE for each condition were bolded.

N affected groups	Model	**Average Absolute Bias**				**RMSE**			
		Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
0	scalar model	**0.002**	**0.001**	**0.001**	**0.003**	**0.031**	**0.036**	**0.043**	**0.072**
	sirt $p = 0.1$	0.003	0.003	0.003	0.005	0.032	0.038	0.046	0.079
	Mplus $p = 0.5$	0.003	0.003	0.003	0.005	0.032	0.037	0.044	0.075
	sirt $p = 0.5$	0.003	0.003	0.003	0.005	0.032	0.037	0.044	0.075
	sirt $p = 1$	0.003	0.003	0.003	0.004	0.032	0.037	0.043	0.072
	sirt $p = 2$	0.003	0.003	0.003	0.004	0.032	0.036	0.043	0.072
5	scalar model	0.052	0.052	0.052	0.054	0.072	0.076	0.081	0.110
	sirt $p = 0.1$	**0.007**	**0.009**	0.020	**0.030**	**0.036**	**0.043**	**0.052**	0.110
	Mplus $p = 0.5$	0.022	0.020	0.019	0.038	0.048	0.051	0.056	**0.104**
	sirt $p = 0.5$	0.012	0.014	**0.017**	0.037	0.039	0.046	0.054	0.105
	sirt $p = 1$	0.027	0.030	0.033	0.047	0.049	0.056	0.064	0.106
	sirt $p = 2$	0.057	0.057	0.057	0.058	0.076	0.080	0.084	0.111
15	scalar model	0.152	0.153	0.154	0.158	0.162	0.164	0.168	0.187
	sirt $p = 0.1$	**0.008**	**0.010**	**0.014**	**0.037**	**0.038**	**0.045**	**0.053**	**0.111**
	Mplus $p = 0.5$	0.018	0.022	0.027	0.060	0.041	0.049	0.058	0.112
	sirt $p = 0.5$	0.015	0.020	0.025	0.058	0.040	0.048	0.057	0.112
	sirt $p = 1$	0.048	0.054	0.062	0.098	0.063	0.073	0.083	0.135
	sirt $p = 2$	0.150	0.150	0.150	0.151	0.158	0.160	0.163	0.178

the recovery of the group means was better for the IA methods than the scalar method. For the settings used in this simulation, the power of $p = 0.1$ resulted in the lowest average absolute bias and RMSE.

Long scale (20 items)

The results of the simulation where zero or eight N-I items were simulated in 5 and 15 out of 25 groups and different sizes of N-I biases were considered: 0.3, 0.6, and 0.9. The IA approach with $p = 0.1$ gave the best recovery of the group means. Differently from scenarios for a 5-item scale, the IA approach provided better results even for conditions in which all items were invariant. With increasing values of p, mean recovery is getting worse. The IA with $p = 2$ gives similar results to the scalar model. The detailed results are available upon request.

Table 11:
Average absolute bias and RMSE for a scenario with 20 items, 25 groups, 1000 observations per group, and different levels of AN-I and eight N-I items with large biases (0.6 in 0, 5 and 15 groups). Smallest average absolute bias and RMSE for each condition were bolded.

N affected groups	Model	**Average Absolute Bias**				**RMSE**			
		Size of AN-I (dif. variance):				Size of AN-I (dif. variance):			
		0.001	0.005	0.010	0.050	0.001	0.005	0.010	0.050
0	scalar model	**0.002**	**0.002**	0.003	0.003	**0.028**	**0.029**	**0.031**	**0.043**
	sirt $p = 0.1$	0.002	0.002	**0.002**	**0.003**	0.028	0.030	0.032	0.045
	Mplus $p = 0.5$	0.007	0.003	0.003	0.003	0.074	0.030	0.032	0.044
	sirt $p = 0.5$	0.003	0.003	0.003	0.003	0.029	0.030	0.032	0.045
	sirt $p = 1$	0.004	0.003	0.003	0.004	0.029	0.030	0.032	0.044
	sirt $p = 2$	0.004	0.003	0.003	0.004	0.029	0.030	0.032	0.043
5	scalar model	0.053	0.053	0.053	0.055	0.070	0.071	0.072	0.085
	sirt $p = 0.1$	**0.007**	**0.008**	**0.010**	**0.031**	**0.030**	**0.033**	**0.037**	**0.069**
	Mplus $p = 0.5$	0.013	0.015	0.019	0.043	0.034	0.038	0.042	0.075
	sirt $p = 0.5$	0.011	0.014	0.018	0.041	0.033	0.037	0.041	0.074
	sirt $p = 1$	0.026	0.030	0.034	0.051	0.045	0.050	0.054	0.081
	sirt $p = 2$	0.057	0.057	0.057	0.057	0.073	0.074	0.075	0.086
15	scalar model	0.150	0.151	0.151	0.157	0.158	0.159	0.160	0.170
	sirt $p = 0.1$	**0.008**	**0.011**	**0.014**	**0.042**	**0.008**	**0.011**	**0.014**	**0.042**
	Mplus $p = 0.5$	0.018	0.024	0.029	0.068	0.036	0.041	0.047	0.090
	sirt $p = 0.5$	0.016	0.022	0.026	0.066	0.035	0.039	0.045	0.088
	sirt $p = 1$	0.049	0.058	0.065	0.108	0.061	0.070	0.078	0.124
	sirt $p = 2$	0.150	0.150	0.150	0.150	0.157	0.157	0.158	0.164

In Table 11, the results of a simulation study for different levels of AMI scenarios (number of affected groups) and where AMI is combined with N-I items with large biases (0.6) are presented. Similar to previous simulation conditions, when only AMI is present. in general, the scalar model performed best. For conditions where AMI occurred in combination with large biases (i.e., there exist N-I items), the IA approach with $p = 0.1$ provided the best mean recovery in terms of bias and RMSE.

Real data example

As an illustrative example of the IA approach, we use data on a measure of depression from Round 6 of the European Social Survey (ESS) from 2012 (ESS, 2014). This data set was also used by Kuha and Moustaki (2015) to illustrate the estimation of latent group mean differences in the case of non-invariant items. We used the same six depression items, each with four response categories (see Kuha & Moustaki, 2015). The ESS dataset contains probability samples from adult populations of 29 countries (listed in Table 12). In our analysis, sample sizes ranged from 752 to 2,958 respondents per country ($M = 1886$, $SD = 520$), and the total sample size was 54,673. In the IA approach, sampling weights were used in the estimation, and the sampling weights were normalized within a country to correspond to a target population of 5,000 respondents. Country means and standard deviations were subsequently transformed such that the total population comprising all 29 countries (all countries have equal contributions) has a mean of zero and a standard deviation of one.

In Table 12, we present the results of the estimated country means under different powers of the IA approach. Again, Mplus and sirt were used for estimating the IA approach. It can be seen that the IA results of Mplus and sirt (IA with $p = 0.5$) closely match and maximally differ by .024. The average absolute deviation between both software packages was 0.009, which can be seen as close enough for practical applications. However, the range of means for a country produced by different powers of IA (i.e., $p = 0.1$, 0.5, 1, and 2) varied considerably ($M = 0.051$, $SD = 0.035$, $Max = 0.121$), although the ranking among countries remained relatively stable.

Table 17 shows the correlations of the estimated country means for the IA approach with different power values. The Mplus and sirt ($p = 0.5$) estimates were quite close ($r = .999$). In addition, the powers $p = 1$ and $p = 0.5$ ($r = .998$), as well as $p = 0.5$ and $p = 0.1$ ($r = .998$), turned out to be very similar. When comparing the differences between the different country mean estimates (see Table 12) and the correlations of the different estimates at the country level (see Table 13), it can be seen that even with a very large correlation at the country level (e.g., $r = .97$), absolute differences of country mean estimates obtained from different methods can be non-negligible.

Table 12:
Estimated Country Means for Depression Scale for ESS Data (Round 6)

Country	**sirt** $p = 2$	**sirt** $p = 1$	**sirt** $p = 0.5$	**Mplus** $p = 0.5$	**sirt** $p = 0.1$
NOR	−0.481	−0.479	−0.478	−0.477	−0.476
DEN	−0.326	−0.378	−0.415	−0.405	−0.433
FIN	−0.375	−0.386	−0.384	−0.376	−0.357
SWE	−0.346	−0.333	−0.330	−0.318	−0.329
IRE	−0.309	−0.297	−0.292	−0.283	−0.290
ISL	−0.257	−0.266	−0.287	−0.285	−0.309
NLD	−0.226	−0.263	−0.283	−0.278	−0.293
SVN	−0.272	−0.272	−0.277	−0.273	−0.281
CHE	−0.225	−0.226	−0.235	−0.246	−0.246
GER	−0.070	−0.115	−0.155	−0.151	−0.215
UK	−0.081	−0.117	−0.151	−0.127	−0.170
BEL	−0.030	−0.080	−0.092	−0.068	−0.095
CYP	−0.146	−0.092	−0.058	−0.057	−0.023
ISR	−0.055	−0.043	−0.036	−0.044	−0.032
POR	0.082	0.005	−0.009	−0.023	−0.014
POL	−0.036	−0.011	−0.004	−0.001	−0.006
FRA	0.000	0.036	0.039	0.046	0.037
SPA	0.043	0.074	0.095	0.085	0.110
EST	0.134	0.126	0.124	0.121	0.124
BGR	0.075	0.138	0.171	0.167	0.184
SVK	0.250	0.222	0.212	0.209	0.208
ITA	0.179	0.211	0.237	0.240	0.252
LIT	0.143	0.210	0.244	0.229	0.262
CZE	0.300	0.312	0.333	0.322	0.357
RUS	0.374	0.386	0.388	0.378	0.389
KOS	0.499	0.427	0.402	0.378	0.389
HUN	0.559	0.556	0.540	0.554	0.531
UKR	0.600	0.656	0.703	0.682	0.725

Note: p = Used power in alignment optimization function. Mplus = IA estimation with ML in Mplus.

Table 13:
Correlations of Estimated Country Means for Depression Scale for ESS Data (Round 6) of Different Linking Approaches

	$p = 2$	$p = 1$	$p = 0.5$	$p = 0.1$
$p = 1$	.991			
$p = 0.5$	.981	**.998**		
$p = 0.1$	.972	.993	**.998**	
Mplus	.981	**.998**	**.999**	**.997**

Note: p = used power in the IA approach in sirt. Mplus = IA estimation in Mplus (using $p = 0.5$). Correlations larger than .995 are written in bold.

Discussion

In this article, we discussed the generalized form of the loss function for IA proposed by Robitzsch (2019) and evaluated different forms of the loss function under different types of non-invariance situations using a Monte Carlo study and an empirical example. We compared two software implementations: Mplus and R, and we conclude that no significant differences exist in terms of parameter recovery (for $p = 0.5$). Our results suggest that the performance of IA heavily depends on the form of the loss function, type of the data (mostly sample size), and type of invariance that could be encountered. In the case of small sample sizes (200 and smaller), alignment models performed worse than the scalar model that simply ignores the problem of non-invariance. In a typical partial invariance situation or when partial invariance is combined with approximate invariance, the alignment model with a power of $p = 0.1$ is the favourable model, closely followed by alignment with $p = 0.5$. The reason behind this is that those small deviations receive larger values in the optimization function for $p = 0.1$ than for $p = 0.5$. In the limit of $p \to 0$, the number of invariant parameters is maximized. When data is generated under partial invariance or partial invariance combined with approximate invariance, we can expect that $p = 0.1$ is superior to $p = 0.5$.

For approximate invariance without large non-invariance biases, the scalar model and alignment with $p = 2$ provided the best recovery of the means. However, it needs to be emphasized that those results are limited to the specific situations and conditions used in this study. In our opinion, we chose the most relevant conditions and most plausible types of item non-invariance to provide a broad general picture of IA performance in estimating group means. More studies examining different conditions and recovering other parameters of the models (e.g., item parameters) are needed to understand the performance and usefulness of IA better. The most urgent issue is to establish proper procedures for choosing the most suitable powers of the loss function for the actual data. It should be noted that for models without equality constraints on item parameters, classical CFA or IRT fit measures are not useful for choosing the proper loss function for IA, and new approaches need to be developed. One direction could be to use a cross-validation approach where data are split into an estimation set and cross-validation set.

The first one is used for estimation, the second one for checking the predictive power of the model. However, this approach could become problematic, especially when sample sizes are small. Some solutions exist, for example, the leave-one-out cross-validation (LOO-CV; Gelfand, Dey & Chang 1992), where a series of single data points are used to test the model's predictive power.

Another direction for the development of IA that, among others, could allow establishing proper model fit is regularization. The invariance alignment approach bears strong similarities to regularization techniques, which are often used in variable selection problems and machine learning (for an overview, see Hastie, Tibshirani, & Wainwright, 2015). In a regularization approach to invariance (Bauer, Belzak, & Cole, 2020; Huang, 2018; Liang & Jacobucci, 2019; Lindstrøm & Dahl, 2020), group-specific item parameters are decomposed into a common item parameter and a group-specific deviation. For example, for item loadings it is assumed that $\lambda_{ig} = \lambda_i + e_{ig}$, which results in an overidentified model. However, in regularization, penalty terms regarding the non-identifiable group-specific deviations e_{ig} are added to the log-likelihood function in the estimation, which ensures empirical identifiability of model parameters and imposes assumptions about the distribution of parameters of non-invariance. IA with $p = 1$ is similar to using the least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) penalty function, while $p = 2$ corresponds to the ridge penalty (Hoerl and Kennard, 1970). IA with $p = 0.5$ is similar to using a bridge penalty in regularized estimation (Hastie et al., 2015). Regularization, as well as IA, show optimal performance in data constellations that result in many invariant parameters for which the differences in parameters among groups are nearly zero, and a few non-invariant item parameters for which the differences in parameters among groups are large. However, regularized estimation typically sets many item parameter deviations to zero, while IA estimates them close to zero. This property has the potential that regularized estimates could produce more stable estimates of group means than the alignment approach. In future studies, it could be interesting to examine whether this assumption holds true.

Finally, it is worth mentioning that alignment shows conceptual similarities with the anchor point selection (APS) procedure proposed by Strobl et al. (2018; see also Bechger & Maris, 2015, for an alternative approach). Although IA is different from APS from a computational point of view, it would be interesting to compare the methods and their performance.

Acknowledgements

The work of the first author 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).

References

Arai, S., & Mayekawa, S. I. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika, 38*(1), 1−16. https://doi.org/10.2333/bhmk.38.1

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

Avvisati, F., Le Donné, N., Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences, 1*:8. https://doi.org/10.1186/s42409-019-0010-z

Battauz, M. (2017). Multiple equating of separate IRT calibrations. *Psychometrika*, *82*(3), 610–636. https://doi.org/10.1007/s11336-016-9517-x

Battauz, M. (2019). Regularized estimation of the nominal response model. *Multivariate Behavioral Research*. Advance online publication. doi: 10.1080/00273171.2019.1681252

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bechger, T. M., Maris, G. A. (2015). Statistical test for differential item pair functioning. *Psychometrika, 80*, 317–340. https://doi.org/10.1007/s11336-014-9408-y

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

DeMars, C. (2020). Alignment as an alternative to anchor purification in DIF analyses. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 56–72. https://doi.org/10.1080/10705511.2019.1617151

European Social Survey. (2014). *ESS-6 2012 Documentation report. Edition 2.1*. European Social Survey Data Archive, Norwegian Social Science Data Services, Bergen.

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education, 29*(1), 30–45. https://doi.org/10.1080/08957347.2015.1102916

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: A Multidisciplinary Journal*, *25*, 56−70. https://doi.org/10.1080/10705511.2017.1374187

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In: J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics*, 4th Edition (pp. 147–167). Oxford: Oxford University Press.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations*. (ETS Research Report RR-09-40). Princeton, NJ: ETS.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. Boca Raton: CRC Press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. DOI: 10.1080/00401706.1970.10488634

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 71*(3), 499–522. https://doi.org/10.1111/bmsp.12130

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

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York: Springer.

Kuha, J., & Moustaki, I. (2015). Nonequivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis. *Psychological Methods, 20*(4), 523–536. https://doi.org/10.1037/met0000031

Liang, X., Jacobucci, R. (2019): Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive Lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*. Advance online publication. doi: 10.1080/10705511.2019.1693273

Lindstrøm, J. C., & Dahl, F. A. (2020). Model selection with lasso in multi-group structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 33–43. https://doi.org/10.1080/10705511.2019.1638262

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Mansolf, M., Vreeker, A., Reise, S. P., Freimer, N. B., Glahn, D. C., Gur, R. E., …, Bilder, R. M. (2020). Extensions of multiple-group item response theory alignment: Application to psychiatric phenotypes in an international genomics consortium. *Educational and Psychological Measurement*. Advance online publication. doi: 10.1177/0013164419897307

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139−160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

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

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

Muthén, B., & Asparouhov, T. (2013). *BSEM measurement invariance analysis*. Mplus web notes. https://www.statmodel.com/examples/webnotes/webnote17.pdf

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

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

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.2044-8317.1992.tb00975.x

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. *Eighth edition*. Los Angeles, CA: Muthén & Muthén.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782–797. https://doi.org/10.1080/10705511.2016.1221313

Oelker, M.-R., Pößnecker, W., & Tutz, G. (2015) Selection and fusion of categorical predictors with L_0-type penalties. *Statistical Modelling 15*(5), 389–410. https://doi.org/10.1177/1471082X14553366

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

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

Robitzsch, A. (2019). *sirt: Supplementary item response theory models*. R package version 3.4-64. https://CRAN.R-project.org/package=sirt

Strobl, C., Kopf, J., Hartmann, R., & Zeileis, A. (2018). *Anchor point selection: An approach for anchoring without anchor items*. Working Papers in Economics and Statistics No. 2018-03. University of Innsbruck.

Stocking, M. L., Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210. https://doi.org/10.1177/014662168300700208

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x