

Article

Examining Differences of Invariance Alignment in the Mplus Software and the R Package Sirt

Alexander Robitzsch ^{1,2} 

¹ IPN–Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de

² Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract: Invariance alignment (IA) is a multivariate statistical technique to compare the means and standard deviations of a factor variable in a one-dimensional factor model across multiple groups. To date, the IA method is most frequently estimated using the commercial Mplus software. IA has also been implemented in the R package sirt. In this article, the performance of IA in the software packages Mplus and R are compared. It is argued and empirically shown in a simulation study and an empirical example that differences between software packages are primarily the cause of different identification constraints in IA. With a change of the identification constraint employing an argument in the IA function in sirt, Mplus and sirt resulted in comparable performance. Moreover, in line with previous work, the simulation study also highlighted that the tuning parameter $\varepsilon = 0.001$ in IA is preferable to $\varepsilon = 0.01$. Furthermore, an empirical example raises the question of whether IA, in its current implementations, behaves as expected in the case of many groups.

Keywords: invariance alignment; confirmatory factor analysis; measurement invariance; statistical software

MSC: 62H10; 62H25; 68U01



Citation: Robitzsch, A. Examining Differences of Invariance Alignment in the Mplus Software and the R Package Sirt. *Mathematics* **2024**, *12*, 770. <https://doi.org/10.3390/math12050770>

Academic Editor: Yen-Kuang Lin

Received: 17 February 2024

Revised: 24 February 2024

Accepted: 3 March 2024

Published: 5 March 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the comparison of multiple groups in confirmatory factor analysis (CFA) regarding factor variables, some identifying assumptions have to be made. It is frequently assumed that item parameters are equal across groups, denoted as measurement invariance [1]. The concept of invariance has been very prominent in psychology and the social sciences in general [2,3]. For example, in international large-scale assessment studies in education, like the Programme for International Student Assessment (PISA), the necessity of invariance is strongly emphasized [4].

In the violation of measurement invariance, the invariance alignment (IA) method [5,6] (also referred to as alignment optimization [7,8]) has been proposed to tackle such situations. The IA method tries to make item parameters as invariant as possible while allowing a few deviations from invariance. By doing so, group comparisons can be made more robust against violations of measurement invariance.

Nowadays, the IA method is frequently applied in social sciences for analyzing questionnaire data [9–12]. Unfortunately, most methodological developments of IA (but see [13–15] for exceptions) are strongly coupled to the popular but commercial (and closed-source) Mplus software [16]. Previous simulation studies for one-dimensional factor models investigated the case of continuous items [5,8,17–19], dichotomous items [20,21], and polytomous items [14,22]. IA to multidimensional factor models with continuous items has been investigated in [23,24]. Moreover, IA was studied in longitudinal models in [25–27]. The optimization function used in IA also gave rise to extending it to a general framework used in penalized structural equation models [28].

Besides the Mplus software, there exists an alternative implementation in the R package *sirt* [29]. However, several researchers have pointed out that there could be subtle differences of IA between Mplus and *sirt*. Unfortunately, there is no systematic comparison of the performance of invariance alignment implementations in Mplus and *sirt*. This article tries to shed some light on the subtleties of implementation differences of IA. It turns out that different identification constraints are likely the cause of the different results of software packages. By changing the default identification constraint in *sirt*, Mplus and *sirt* provided much more similar results. Moreover, the results from a simulation study also question the default choices of tuning parameters in the software packages.

The rest of this article is structured as follows. In Section 2, the background of IA is reviewed. Section 3 discusses the syntax code and estimation options of IA in Mplus and *sirt*. In Section 4, the two software packages are compared by means of a simulation study. An empirical example is presented in Section 5. Finally, the paper closes with a discussion in Section 6.

2. Invariance Alignment

Let the random variable X_{ig} denote item i ($i = 1, \dots, I$) in group g ($g = 1, \dots, G$). A one-dimensional factor model [30] is defined as

$$X_{ig} = v_{ig} + \lambda_{ig}F_g + \epsilon_{ig}, F_g \sim N(\mu_g, \sigma_g^2), \epsilon_{ig} \sim N(0, \omega_{ig}), \tag{1}$$

where λ_{ig} are item loadings, and v_{ig} are item intercepts. Item loadings can be assumed to be positive. If some loading is negative, the corresponding random variable X_{ig} must be multiplied by -1 . The factor variables F_g and all residual variables ϵ_{ig} are independent and univariate normally distributed. The factor variable F_g has a factor mean μ_g and a factor standard deviation σ_g .

Without additional assumptions, the parameters in (1) are not identified. An identified model is obtained by assuming a standardized latent variable F_g (i.e., with a mean of 0 and a standard deviation of 1):

$$X_{ig} = v_{ig,0} + \lambda_{ig,0}F_g + \epsilon_{ig}, F_g \sim N(0, 1), \epsilon_{ig} \sim N(0, \omega_{ig}). \tag{2}$$

The parameters in (1) and (2) are related to each other by

$$\lambda_{ig,0} = \lambda_{ig}\sigma_g \text{ and } v_{ig,0} = v_{ig} + \lambda_{ig}\mu_g = v_{ig} + \frac{\lambda_{ig,0}}{\sigma_g}\mu_g. \tag{3}$$

In many applications, the factor means μ_g and factor standard deviations σ_g should be compared across groups. To achieve this, a typical assumption in the social sciences is the property is measurement invariance [1,3]. Measurement invariance presupposes that item loadings λ_{ig} and item intercepts v_{ig} are equal across groups. That is, there exist common item loadings λ_i such that $\lambda_i = \lambda_{ig}$ for all $g = 1, \dots, G$ and common item intercepts v_i such that $v_i = v_{ig}$ for $g = 1, \dots, G$ for all items $i = 1, \dots, I$. The absence of measurement invariance is also labeled as differential item functioning (DIF; [2,31]) in the item response theory literature. If measurement invariance holds, (3) can be rewritten as

$$\lambda_{ig,0} = \lambda_i\sigma_g \text{ and } v_{ig,0} = v_i + \frac{\lambda_{ig,0}}{\sigma_g}\mu_g. \tag{4}$$

The IA method of Asparouhov and Muthén [5,6] tackles situations under sparse violations of measurement invariance. In this case, a few item loadings or item intercepts are allowed to differ across groups, while the majority of items (approximately) fulfills the invariance assumption [32]. This situation is called partial invariance in the literature [33].

The IA estimation method proceeds in two steps. In the first step, the one-dimensional factor model (2) is separately estimated by the maximum likelihood method for all groups in the first step. The estimated item parameters $\hat{\lambda}_{ig,0}$ and $\hat{v}_{ig,0}$ ($i = 1, \dots, I; g = 1, \dots, G$)

are used as the input of the IA. By rewriting (3) and inserting the estimated item loadings and item intercepts, we obtain

$$\lambda_{ig} - \lambda_{ih} = \frac{\hat{\lambda}_{ig,0}}{\sigma_g} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \text{ and } v_{ig} - v_{ih} = \hat{v}_{ig,0} - \hat{v}_{ih,0} - \frac{\hat{\lambda}_{ig,0}}{\sigma_g} \mu_g + \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \mu_h. \tag{5}$$

These relations motivate the minimization of the following linking function in IA to determine group means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_G)$ and standard deviations $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_G)$:

$$H(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G w_{i1,gh} \rho \left(\frac{\hat{\lambda}_{ig,0}}{\sigma_g} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \right) + \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G w_{i2,gh} \rho \left(\hat{v}_{ig,0} - \hat{v}_{ih,0} - \frac{\hat{\lambda}_{ig,0}}{\sigma_g} \mu_g + \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \mu_h \right), \tag{6}$$

where the weights $w_{i1,gh}$ and $w_{i2,gh}$ are known, and ρ is a nonnegative, symmetric loss function with $\rho(0) = 0$ and is monotonically increasing for nonnegative x values. Asparouhov and Muthén [5] proposed using $w_{i1,gh} = w_{i2,gh} = \sqrt{n_g n_h}$ and $\rho(x) = \sqrt{|x|}$, where n_g denotes the sample size of group g .

In the minimization of (6), additional identification constraints must be imposed. As a first alternative, the distribution parameters of the first (or any other) group can be fixed. That is, we set $\mu_1 = 0$ and $\sigma_1 = 1$. As a second alternative, one can simultaneously constrain all estimated parameters. Then, the following identification constraints can be imposed:

$$\sum_{i=1}^G \mu_g = 0 \text{ and } \prod_{g=1}^G \sigma_g = 1. \tag{7}$$

The constraints in (7) state that the arithmetic mean of the factor means equals zero, and the geometric mean of the factor standard deviation equals one.

Note that the optimization function H of IA defined in (6) can be rewritten as

$$H(\boldsymbol{\mu}, \boldsymbol{\sigma}) = H_1(\boldsymbol{\sigma}) + H_2(\boldsymbol{\mu}, \boldsymbol{\sigma}), \text{ where} \tag{8}$$

$$H_1(\boldsymbol{\sigma}) = \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G \rho \left(\frac{\hat{\lambda}_{ig,0}}{\sigma_g} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \right) \text{ and } H_2(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G \rho \left(\hat{v}_{ig,0} - \hat{v}_{ih,0} - \frac{\hat{\lambda}_{ig,0}}{\sigma_g} \mu_g + \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \mu_h \right). \tag{9}$$

However, the function H_2 can be conveniently substituted by an alternative. Note that Equation (3) can be rewritten as

$$\log \lambda_{ig,0} = \log \lambda_{ig} + \log \sigma_g \text{ and } v_{ig,0} = v_{ig} + \frac{\lambda_{ig,0}}{\sigma_g} \mu_g. \tag{10}$$

This motivates the alternative optimization function H_1^* for determining standard deviations, which employs logarithmized item loadings (see [34,35])

$$H_1^*(\boldsymbol{\sigma}^*) = \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G \rho \left(\log \hat{\lambda}_{ig,0} - \log \hat{\lambda}_{ih,0} - \sigma_g^* + \sigma_h^* \right), \tag{11}$$

where $\sigma_g^* = \log \sigma_g$ for $g = 1, \dots, G$. Due to the required identification constraints, we fix $\sigma_1^* = 0$ (i.e., $\sigma_1 = \exp(\sigma_1^*) = 1$). By minimizing H_1^* , a vector of standard deviations $\hat{\boldsymbol{\sigma}}^*$ on the logarithm metric is obtained; that is, $\hat{\boldsymbol{\sigma}}^* = (\hat{\sigma}_1^*, \dots, \hat{\sigma}_G^*)$. The vector of estimated standard deviations $\hat{\boldsymbol{\sigma}}$ can be obtained by exponentiating all entries in $\hat{\boldsymbol{\sigma}}^*$.

2.1. Numerical Optimization

As mentioned above, IA uses the loss function $\rho(x) = \sqrt{|x|} = |x|^{0.5}$ as the default in the Mplus software package [16]. However, the loss function $\rho(x) = |x|^{0.25}$ is also available in Mplus [16]. The more general L_p loss function $\rho(x) = |x|^p$ for $p > 0$ has been studied

for IA in [13,35]. It has been shown that values of the power p smaller than 0.5 can be advantageous in some situations [13].

In the practical minimization of H involved in IA, the nondifferentiable L_p loss function $\rho(x) = |x|^p$ (for $0 < p \leq 1$) is replaced by a differentiable approximation ρ_D (see [5,35])

$$\rho_D(x) = (x^2 + \varepsilon)^{p/2}, \tag{12}$$

where $\varepsilon > 0$ is a tuning parameter that controls the approximation error of ρ_D for ρ . The approximation error becomes smaller with ε values close to zero. However, the minimization of H in IA becomes more difficult when choosing too small values of ε . Practical experience led to proposals $\varepsilon = 0.01$ [5] or $\varepsilon = 0.001$ [35]. The choice $\varepsilon = 0.01$ is the default in Mplus (see [13]).

2.2. A More in-Depth Look into the Identification Constraint for Standard Deviations for Many Groups

The IA method measures the similarity between item loadings in the optimization function H_1 by

$$H_1(\sigma) = \sum_{i=1}^I \sum_{g=1}^{G-1} \sum_{h=g+1}^G \rho \left(\frac{\hat{\lambda}_{ig,0}}{\sigma_g} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \right). \tag{13}$$

As mentioned above, an identification constraint could be to fix the standard deviation of the first group to 1 or to fix the product of standard deviations to 1. Regarding the choice of the chosen identification constraint in their Mplus software, Asparouhov and Muthén [5] state that “[...] in Mplus by default the parameters are indeed reported in that metric, however, the alignment optimization is carried out using Equation (10) [i.e., the product identification constraint in (7)] to ensure full symmetry between the different groups”. To illustrate this motivation a bit, we rewrite (13) as

$$H_1(\sigma) = \sum_{i=1}^I \sum_{h=2}^G \rho \left(\frac{\hat{\lambda}_{i1,0}}{\sigma_1} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \right) + \sum_{i=1}^I \sum_{g=2}^{G-1} \sum_{h=g+1}^G \rho \left(\frac{\hat{\lambda}_{ig,0}}{\sigma_g} - \frac{\hat{\lambda}_{ih,0}}{\sigma_h} \right), \tag{14}$$

where we decomposed the terms that do involve and do not involve the first group, respectively. If the optimization would only have been carried out based on the second term in (14), the optimization value would tend to zero if standard deviations tend to infinity. Hence, fixing the standard deviation σ_1 to 1 prevents obtaining infinite estimates of σ_g for $g = 2, \dots, G$. If $\sigma_1 = 1$ is specified in the minimization of (14), it becomes clear that the first term in the sum involving the first group becomes less relevant if the number of groups increases. Hence, there is a danger that estimated standard deviations are larger if more groups are involved in the analysis. For this reason, the identification constraint $\sigma_1 = 1$ is likely not appropriate in the case of many groups. In contrast, the constraint $\prod_{g=1}^G \sigma_g = 1$ would be preferable in this case. The behavior of IA for many groups is analyzed in a simulation study in Section 4 and an empirical example in Section 5.

3. Implementation of Invariance Alignment in Mplus and Sirt

We now describe how IA can be estimated with the commercial Mplus software (Version 8.9; [16]) and the R (Version 4.3; [36]) package *sirt* [29].

Listing 1 contains command-line syntax for the specification of IA in Mplus (see [16,37]). The dataset is locally saved in `mydata.dat` (see Line 4 in Listing 1) in an appropriate working directory. The IA method should be applied for five items $I1, \dots, I5$ (see Line 6 in Listing 1). The numeric grouping variable `group` is included in the dataset. The grouping variable has to be specified as a known class variable in Mplus (see Lines 8 and 9 in Listing 1).

Listing 1. Specification of invariance alignment in Mplus software.

```

1  TITLE:
2  Invariance Alignment;
3  DATA:
4  FILE IS mydata.dat;
5  VARIABLE:
6  NAMES ARE group I1 I2 I3 I4 I5;
7  USEVARIABLES ARE group I1 I2 I3 I4 I5;
8  CLASSES = c(3);
9  KNOWNCLASS = c(group = 1 group = 2 group = 3);
10 ANALYSIS:
11 TYPE = MIXTURE;
12 ESTIMATOR = MLR;
13 ALIGNMENT = FIXED(2); ! group=2 is reference group with zero mean;
14 ! ALIGNMENT = FREE for method 'FREE';
15
16 TOLERANCE = 0.01; ! epsilon value;
17 SIMPLICITY = SQRT; ! for p=0.5;
18 ! SIMPLICITY = FOURTHRT for p=0.25;
19 MODEL:
20 %overall%
21 f1 BY I1 I2 I3 I4 I5;
22 OUTPUT:
23 alignment;

```

Mplus has only implemented the product constraint $\prod_{g=1}^G \sigma_g = 1$ for standard deviations. The method `FIXED` (i.e., Line 13 in Listing 1 that states `ALIGNMENT=FIXED`) utilizes the zero constraint of the factor of the first group; that is, $\mu_1 = 0$. The reference to the first group can be changed using the command `ALIGNMENT=FIXED(2)` (see Line 13 in Listing 1). In this case, Group 2 is used as the reference group. Alternatively, “the `FREE` alignment optimization estimates α_1 as an additional parameter” [5]. This specification seems to be overparametrized, and Mplus must have implemented some fix to prevent nonconvergence of the IA optimization problem. The Mplus manual states, “In the `FREE` setting, all factor means are estimated. `FREE` is the most general approach” [16]. This statement does not certainly provide enough details for an independent implementation of the black-box algorithms in the Mplus software. Furthermore, the `TOLERANCE` argument in Line 15 in Listing 1 specifies the tuning parameter ε that appears in the differentiable approximation (12). The default in Mplus is $\varepsilon = 0.01$. Finally, the `SIMPLICITY` argument can either choose the power $p = 0.5$ (i.e., square root `SQRT`) or $p = 0.25$ (i.e., fourth root `FOURTHRT`).

Listing 2 shows how IA can be estimated in the R package `sirt` [29,38,39]. In the first step, group-specific estimation of the one-dimensional factor models can be carried out with the function `sirt::invariance_alignment_cfa_config()` (see Line 5 in Listing 2). The group-specific estimated item loadings `lambda` and item intercepts `nu` can be extracted from the output of this function (see Lines 9 and 10 in Listing 2). Moreover, the weights $w_{g1,g2}$ in IA (see Equation (6)) are specified in Line 14 in Listing 2. The specification in this listing ensures the same chosen weights as in Mplus. The function `sirt::invariance.alignment()` performs IA based on estimated item loadings `lambda` and item intercepts `nu` (see Line 17 in Listing 2). The power p in IA can be separately chosen for item loadings (first entry in `align.pow`) and item intercepts (second entry in `align.pow`). If the power $p = 0.25$ instead of the default $p = 0.5$ should be used in the analysis, users have to specify the argument `align.pow=c(0.25,0.25)` in the `sirt::invariance.alignment()` function. The tuning parameter ε in Equation (12) can be specified with the argument `eps`.

Listing 2. Specification of invariance alignment in the R package sirt.

```

1  ## define items
2  items <- paste0('I', 1:5)
3
4  ## separate estimate of factor model in groups
5  prep <- sirt::invariance_alignment_cfa_config(dat=dat[,items],
6  group=dat$group )
7
8  # extract item loadings and item intercepts
9  lambda <- prep$lambda
10 nu <- prep$nu
11
12 #- define weights
13 Ng <- prep$N
14 wgts <- matrix(sqrt(Ng), length(Ng), ncol(nu))
15
16 ## perform invariance alignment
17 res <- sirt::invariance_alignment(lambda=lambda, nu=nu,
18 align.pow=c(.5, .5), eps=0.01, wgt=wgts, meth=3)
19
20 #- extract estimated means and standard deviations
21 res$pars

```

The IA function in the sirt package has four different estimation methods that can be requested with the argument `meth`. The default `meth = 1` uses the optimization Function (6) with the identification constraints $\mu_1 = 0$ and $\sigma_1 = 1$. The method `meth = 2` performs IA on logarithmized item loadings (see Equation (11)), also using the constraints $\mu_1 = 0$ and $\sigma_1 = 1$. The method `meth = 3` implements the product constraint $\prod_{g=1}^G \sigma_g = 1$ for standard deviations and the zero mean constraint for the first group (i.e., $\mu_1 = 0$). Hence, this method is expected to perform similarly to Mplus' FIXED alignment method. Finally, `meth = 4` also utilizes the product constraint for standard deviations but freely estimates the first group mean μ_1 . To identify the model, a penalty term $\omega W \sum_{g=1}^G \mu_g^2$ is added to the optimization function, where W is the sum of the involved weights in the IA optimization function and $\omega = 0.01$ is a small factor to achieve convergence in optimization. Likely, this method has only conceptual similarity with Mplus' FREE method, and no equivalent performance can be expected.

The estimated distributed parameters can be requested by the list entry `$pars` (see Line 21 in Listing 2).

4. Simulation Study

4.1. Method

The datasets in this simulation study were simulated from a one-dimensional factor model consisting of $I = 5$ items and $G = 3, 6, 9,$ or 12 groups. In the case of three groups, the group means were 0, 0.3, and 0.8, and the group standard deviations were 1, 1.225, and 1.095, respectively. With more than three groups, all parameters (i.e., distribution and item parameters) were replicated accordingly. For example, for six groups, the parameters were twice replicated.

All measurement error variances were set to 1 in all groups and uncorrelated with each other. The factor variable and residual variables were normally distributed. There was noninvariance in item intercepts and item loadings. All item intercepts had a value of zero except for a few cases. In the first group, the fifth item intercept was 0.5. In the second group, the first item intercept was -0.5 , while the second item had an intercept of -0.5 in the third group. All item loadings had a value of one except for a few cases. In the first group, the third item loading was 1.5. In the second group, the fifth item loading was 0.5,

while the fourth item loading was 0.5 in the third group. These parameters were duplicated with more than three groups as described above.

The sample size per group was chosen as $N = 250$, $N = 500$, $N = 1000$, $N = 2000$, or $N = \text{Inf}$ (i.e., infinite sample size). In the case of an infinite sample size, there was no sampling error, and the population parameters were the data-generating parameters. The mean vectors and the covariance matrices are sufficient statistics for the IA method. Datasets with a sample size of $N = 9999$, whose empirical means and covariances equaled the population means and covariances, respectively, were simulated in this case.

The IA method was applied in the Mplus software (Version 8.9; [16]), and the function `invariance.alignment()` in the R package `sirt` (Version 4.1-15; [29]) was applied. Both software packages utilized the power $p = 0.5$ and the tuning parameter choices $\varepsilon = 0.01$ and $\varepsilon = 0.001$. Mplus was used with the `FIXED` or the `FREE` methods, while the method `meth` in `sirt` was specified as `meth = 1`, `meth = 2`, `meth = 3`, or `meth = 4`. To compare the performance across methods, the estimates were linearly transformed such that the mean and the SD of the first group were 0 and 1, respectively.

In total, $R = 1000$ replications were conducted for each cell of the simulation study. Bias, standard deviation (SD), root mean square error (RMSE), and relative RMSE were computed to assess the performance of the different estimators for factor means and factor standard deviations. To ease the comparability between the different estimation methods, we computed a relative RMSE value, which was defined as the quotient of the RMSE for a particular method and the RMSE of a reference method. This quotient was multiplied by 100 afterward. The reference method was Mplus' `FIXED` method with $p = 0.5$ and $\varepsilon = 0.01$, which is the default in this software package. We also computed the mean absolute difference between estimates of Mplus and `sirt` to determine possible differences between software packages. Information about model specifications can be found in the material located at <https://osf.io/84ne5> (accessed on 17 February 2024).

4.2. Results

In this section, we only present results for the distribution parameters for the second group. The findings for the other groups were very similar.

Table 1 contains the bias of the estimated factor mean μ_2 for different estimation methods in Mplus and `sirt`. Overall, noticeable bias occurred for $\varepsilon = 0.01$ and $p = 0.5$. However, the bias decreased with increasing sample size but still appeared in infinite sample sizes. Moreover, note that the bias did not disappear with an increasing number of groups. Interestingly, bias was substantially reduced with the tuning parameter $\varepsilon = 0.001$, particularly for sample sizes of at least 1000. For three, six, or nine groups, the method `meth = 1` in `sirt` performed best in terms of bias. In general, the bias of both Mplus methods `FIXED` and `FREE` was similar to those obtained from the four methods implemented in `sirt`. Interestingly, `sirt`'s method `meth = 1` had issues with an increasing number of groups. For $G = 12$ and $N = 250$, there was a large bias in estimated factor means, which showed that `meth = 1` failed for a large number of groups.

Table 2 shows the relative RMSE of the estimated factor mean μ_2 in the second group. The `FREE` method in Mplus was slightly inferior to the `FIXED` method in Mplus for more than three groups. The tuning parameter $\varepsilon = 0.001$ outperformed $\varepsilon = 0.01$ in terms of relative RMSE. This observation was primarily an effect of the larger bias for $\varepsilon = 0.01$. The simulation study also highlighted that the SD for the different estimates was larger for $\varepsilon = 0.001$ than for $\varepsilon = 0.01$.

Table 3 presents the average absolute difference between the estimates of the factor mean in the second group between Mplus and `sirt`. It can be seen that Mplus' `FIXED` method was closest to the `sirt` method `meth = 3`. The differences were larger to `sirt`'s `meth = 1`, which is the default in the R package `sirt`. Furthermore, the `FREE` method of Mplus turned out to perform most similarly to `sirt`'s `meth = 4`. However, the differences between the two methods are noticeable. Hence, it can be concluded that there is no equivalent implementation of the Mplus `FREE` method in the `sirt` package.

Table 1. Simulation Study: Bias of the estimated factor mean in the second group μ_2 for different estimation methods in Mplus and sirt for $p = 0.5$ as a function of the invariance alignment parameter ε , sample size per group N , and number of groups G .

G	N	$\varepsilon = 0.01$						$\varepsilon = 0.001$					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
3	250	-0.075	-0.073	-0.062	-0.071	-0.071	-0.070	-0.048	-0.048	-0.042	-0.050	-0.050	-0.049
	500	-0.059	-0.056	-0.050	-0.057	-0.056	-0.052	-0.032	-0.028	-0.029	-0.033	-0.033	-0.027
	1000	-0.046	-0.045	-0.037	-0.043	-0.042	-0.041	-0.017	-0.016	-0.014	-0.017	-0.017	-0.016
	2500	-0.041	-0.040	-0.033	-0.038	-0.037	-0.036	-0.012	-0.011	-0.009	-0.011	-0.011	-0.010
	Inf	-0.037	-0.037	-0.030	-0.034	-0.033	-0.033	-0.006	-0.006	-0.005	-0.006	-0.006	-0.006
6	250	-0.073	-0.069	-0.052	-0.070	-0.070	-0.066	-0.046	-0.045	-0.033	-0.046	-0.047	-0.045
	500	-0.065	-0.064	-0.051	-0.063	-0.063	-0.060	-0.037	-0.036	-0.029	-0.036	-0.037	-0.036
	1000	-0.048	-0.048	-0.035	-0.046	-0.045	-0.044	-0.020	-0.020	-0.015	-0.020	-0.020	-0.019
	2500	-0.040	-0.040	-0.029	-0.038	-0.037	-0.037	-0.012	-0.011	-0.008	-0.011	-0.011	-0.011
	Inf	-0.036	-0.037	-0.027	-0.034	-0.033	-0.033	-0.006	-0.006	-0.005	-0.006	-0.006	-0.006
9	250	-0.075	-0.070	-0.048	-0.075	-0.075	-0.069	-0.047	-0.043	-0.028	-0.050	-0.050	-0.046
	500	-0.063	-0.061	-0.043	-0.062	-0.061	-0.058	-0.037	-0.033	-0.023	-0.035	-0.036	-0.033
	1000	-0.050	-0.049	-0.033	-0.048	-0.048	-0.046	-0.022	-0.020	-0.014	-0.022	-0.022	-0.020
	2500	-0.041	-0.042	-0.027	-0.040	-0.039	-0.039	-0.014	-0.013	-0.009	-0.014	-0.013	-0.013
	Inf	-0.035	-0.037	-0.023	-0.034	-0.033	-0.033	-0.006	-0.006	-0.004	-0.006	-0.006	-0.006
12	250	-0.086	-0.081	0.320	-0.086	-0.086	-0.078	-0.059	-0.053	3.014	-0.060	-0.060	-0.055
	500	-0.054	-0.053	-0.024	-0.054	-0.053	-0.050	-0.027	-0.024	-0.009	-0.027	-0.028	-0.024
	1000	-0.048	-0.048	-0.025	-0.048	-0.047	-0.045	-0.021	-0.019	-0.010	-0.021	-0.021	-0.019
	2500	-0.041	-0.043	-0.023	-0.041	-0.040	-0.040	-0.014	-0.014	-0.008	-0.014	-0.014	-0.014
	Inf	-0.034	-0.037	-0.019	-0.034	-0.033	-0.033	-0.006	-0.006	-0.003	-0.006	-0.006	-0.006

Note: Inf = infinite sample size (i.e., using population parameters); Absolute biases larger than 0.03 are printed in bold font. After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

Table 2. Simulation Study: Relative root mean square error of the estimated factor mean in the second group μ_2 for different estimation methods in Mplus and sirt for $p = 0.5$ as a function of the invariance alignment parameter ε , sample size per group N , and number of groups G .

G	N	$\varepsilon = 0.01$						$\varepsilon = 0.001$					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
3	250	100 ‡	100.3	98.4	98.9	98.9	98.9	94.8	97.2	96.6	96.1	96.2	97.6
	500	100 ‡	98.3	97.2	99.1	99.0	95.4	92.0	89.7	93.7	94.3	94.4	89.3
	1000	100 ‡	99.4	93.7	96.9	96.4	94.9	84.3	84.5	84.0	83.9	83.9	82.3
	2500	100 ‡	99.1	90.1	95.8	94.8	93.2	71.4	70.4	71.0	71.5	71.5	69.4
6	250	100 ‡	101.0	99.2	100.6	100.5	99.7	95.9	95.9	99.3	98.5	98.3	96.2
	500	100 ‡	100.8	95.2	99.8	99.5	98.6	86.3	87.0	86.7	87.9	88.0	86.8
	1000	100 ‡	101.6	91.3	98.1	97.6	96.9	82.4	82.6	81.1	82.2	82.2	81.0
	2500	100 ‡	101.5	87.5	97.5	96.5	96.1	70.0	70.0	70.1	71.0	71.0	69.8
9	250	100 ‡	100.3	97.8	100.8	100.7	99.5	91.8	91.5	95.6	94.2	94.0	96.2
	500	100 ‡	102.8	94.7	100.1	100.0	100.4	89.1	90.9	89.9	90.6	90.6	90.7
	1000	100 ‡	101.4	89.3	99.1	98.5	97.9	80.4	80.9	79.9	81.3	81.3	80.6
	2500	100 ‡	102.7	85.6	98.7	97.7	97.9	72.0	73.1	71.8	73.0	72.9	72.9
12	250	100 ‡	102.2	734.9	100.6	100.4	100.3	91.9	94.5	5980	93.6	93.5	94.8
	500	100 ‡	105.0	96.3	101.4	101.2	103.0	88.9	93.1	92.6	91.8	91.7	93.7
	1000	100 ‡	105.0	89.1	100.7	100.1	101.7	82.5	86.3	83.2	84.3	84.2	86.2
	2500	100 ‡	104.9	80.6	99.9	98.9	100.1	69.8	72.4	69.0	71.2	71.1	72.4

Note: ‡ = The reference method for the computation of the relative RMSE was “Mplus, FIXED” with $p = 0.5$ and $\varepsilon = 0.01$. Absolute RMSE values smaller than 100 are printed in bold font. After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

Table 3. Simulation Study: Mean absolute difference between different Mplus and sirt estimation methods of the estimated factor mean in the second group μ_2 for $p = 0.5$ and $\varepsilon = 0.01$ as a function of sample size per group N and number of groups G .

G	N	Mplus, FIXED and sirt, meth				Mplus, FIXED and sirt, meth			
		1	2	3	4	1	2	3	4
3	250	0.0174	0.0111	0.0113	0.0152	0.0200	0.0160	0.0161	0.0109
	500	0.0123	0.0085	0.0087	0.0101	0.0135	0.0109	0.0110	0.0078
	1000	0.0100	0.0064	0.0067	0.0067	0.0101	0.0072	0.0074	0.0061
	2500	0.0083	0.0043	0.0049	0.0050	0.0081	0.0044	0.0049	0.0044
6	250	0.0234	0.0045	0.0025	0.0280	0.0411	0.0435	0.0415	0.0110
	500	0.0159	0.0059	0.0046	0.0715	0.0847	0.0903	0.0877	0.0145
	1000	0.0126	0.0050	0.0031	0.0651	0.0791	0.0831	0.0808	0.0134
	2500	0.0109	0.0046	0.0026	0.0490	0.0625	0.0657	0.0636	0.0122
9	250	0.0293	0.0045	0.0025	0.0365	0.0502	0.0529	0.0509	0.0119
	500	0.0218	0.0045	0.0025	0.0276	0.0411	0.0435	0.0415	0.0114
	1000	0.0168	0.0061	0.0041	0.0728	0.0881	0.0939	0.0914	0.0153
	2500	0.0140	0.0051	0.0029	0.0616	0.0766	0.0806	0.0784	0.0140
12	250	0.4073	0.0093	0.0093	0.0214	0.4028	0.0171	0.0169	0.0106
	500	0.0302	0.0070	0.0071	0.0135	0.0298	0.0099	0.0098	0.0075
	1000	0.0221	0.0053	0.0054	0.0097	0.0228	0.0070	0.0070	0.0055
	2500	0.0182	0.0030	0.0032	0.0053	0.0198	0.0044	0.0046	0.0040

Note: After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

Table 4 shows the bias for the factor SD of the second group for $p = 0.5$. As for the factor mean, the tuning parameter $\varepsilon = 0.001$ had superior performance compared to $\varepsilon = 0.01$. For the SD, the Mplus methods FIXED and FREE as well as sirt’s meth = 3 and meth = 4 coincide. Overall, the sirt method meth = 1 was preferable for three or six groups, while its performance deteriorated for a larger number of groups. It should be emphasized that the bias did not even disappear in infinite sample sizes for $\varepsilon = 0.01$.

Table 5 presents the relative RMSE for the factor SD in the second group. The specifications with $\varepsilon = 0.001$ were generally preferable over $\varepsilon = 0.01$ in terms of RMSE. The Mplus and sirt methods performed very similarly. Obviously, the bias issues of sirt’s meth = 1 for many groups (i.e., 9 or 12 groups) also translated into substantially increased RMSE values.

Table 6 displays the mean absolute difference for the estimate of the factor SD in the second group between Mplus and sirt. The Mplus method FIXED had a similar performance to the sirt meth = 3, while Mplus’ FREE method has comparable performance with sirt’s meth = 4.

To conclude, this simulation study demonstrated that the performance of IA estimates in Mplus can be similar to sirt if an appropriate estimation method meth in sirt is chosen. The default sirt method meth = 1 resulted in larger differences to Mplus. However, sirt’s meth = 1 can be preferred over Mplus and the other sirt methods for three or six groups but cannot be recommended for many groups (i.e., at least nine groups). Overall, the tuning parameter $\varepsilon = 0.001$ should be preferred over $\varepsilon = 0.01$ in terms of bias and RMSE.

Table 4. Simulation Study: Bias of the estimated factor standard deviation in the second group σ_2 for different estimation methods in Mplus and sirt for $p = 0.5$ as a function of the invariance alignment parameter ε , sample size per group N , and number of groups G .

G	N	$\varepsilon = 0.01$						$\varepsilon = 0.001$					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
3	250	-0.064	-0.064	-0.020	-0.067	-0.066	-0.066	-0.037	-0.037	-0.007	-0.041	-0.042	-0.042
	500	-0.058	-0.058	-0.028	-0.062	-0.059	-0.059	-0.032	-0.032	-0.015	-0.035	-0.036	-0.036
	1000	-0.046	-0.046	-0.023	-0.050	-0.047	-0.047	-0.020	-0.020	-0.010	-0.022	-0.022	-0.022
	2500	-0.041	-0.041	-0.022	-0.045	-0.041	-0.041	-0.015	-0.015	-0.008	-0.015	-0.015	-0.015
	Inf	-0.033	-0.033	-0.018	-0.038	-0.033	-0.033	-0.006	-0.006	-0.003	-0.006	-0.006	-0.006
6	250	-0.071	-0.071	0.014	-0.074	-0.073	-0.073	-0.043	-0.043	0.020	-0.044	-0.045	-0.045
	500	-0.053	-0.053	0.003	-0.057	-0.054	-0.054	-0.026	-0.026	0.008	-0.026	-0.027	-0.027
	1000	-0.049	-0.049	-0.006	-0.053	-0.050	-0.050	-0.023	-0.023	-0.002	-0.024	-0.024	-0.024
	2500	-0.037	-0.037	-0.002	-0.041	-0.038	-0.038	-0.011	-0.011	0.002	-0.011	-0.011	-0.011
	Inf	-0.033	-0.033	-0.005	-0.038	-0.033	-0.033	-0.006	-0.006	-0.001	-0.006	-0.006	-0.006
9	250	-0.067	-0.067	0.069	-0.071	-0.070	-0.070	-0.040	-0.040	0.061	-0.042	-0.042	-0.042
	500	-0.049	-0.049	0.042	-0.053	-0.051	-0.051	-0.021	-0.021	0.034	-0.023	-0.023	-0.023
	1000	-0.042	-0.042	0.026	-0.046	-0.042	-0.042	-0.015	-0.015	0.018	-0.016	-0.016	-0.016
	2500	-0.038	-0.038	0.016	-0.042	-0.038	-0.038	-0.011	-0.011	0.008	-0.012	-0.012	-0.012
	Inf	-0.033	-0.033	0.010	-0.038	-0.033	-0.033	-0.006	-0.006	0.002	-0.006	-0.006	-0.006
12	250	-0.062	-0.062	2.041	-0.067	-0.065	-0.065	-0.036	-0.036	14.39	-0.037	-0.038	-0.038
	500	-0.047	-0.047	0.090	-0.051	-0.049	-0.049	-0.020	-0.020	0.059	-0.021	-0.021	-0.021
	1000	-0.043	-0.043	0.055	-0.048	-0.044	-0.044	-0.017	-0.017	0.030	-0.018	-0.018	-0.018
	2500	-0.039	-0.039	0.038	-0.043	-0.039	-0.039	-0.012	-0.012	0.014	-0.013	-0.013	-0.013
	Inf	-0.033	-0.033	0.028	-0.038	-0.033	-0.033	-0.006	-0.006	0.004	-0.006	-0.006	-0.006

Note: Inf = infinite sample size (i.e., using population parameters); Absolute biases larger than 0.03 are printed in bold font. After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

Table 5. Simulation Study: Relative root mean square error of the estimated factor standard deviation in the second group σ_2 for different estimation methods in Mplus and sirt for $p = 0.5$ as a function of the invariance alignment parameter ε , sample size per group N , and number of groups G .

G	N	$\varepsilon = 0.01$						$\varepsilon = 0.001$					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
3	250	100 ‡	100.0	87.5	102.8	102.2	102.3	97.5	97.4	92.6	99.1	99.1	99.1
	500	100 ‡	100.0	82.8	104.1	102.4	102.4	90.5	90.5	83.1	93.6	93.8	93.8
	1000	100 ‡	100.0	83.2	104.8	101.5	101.5	83.9	83.9	80.5	85.1	85.4	85.4
	2500	100 ‡	100.0	75.3	107.2	101.3	101.3	70.3	70.3	66.4	71.4	71.3	71.3
6	250	100 ‡	100.0	88.9	102.9	101.6	101.6	93.5	93.5	92.8	94.8	95.1	95.1
	500	100 ‡	100.0	80.8	102.7	100.7	100.7	88.2	88.2	84.3	88.1	88.2	88.2
	1000	100 ‡	100.0	73.5	104.6	101.5	101.5	80.4	80.4	75.0	81.2	81.1	81.1
	2500	100 ‡	100.0	64.8	107.8	101.8	101.8	68.4	68.4	65.7	69.7	69.5	69.5
9	250	100 ‡	100.0	111.6	102.8	101.4	101.4	93.5	93.5	113.3	95.1	94.5	94.5
	500	100 ‡	100.0	104.4	103.4	101.4	101.4	87.1	87.1	101.3	88.4	88.1	88.1
	1000	100 ‡	100.0	89.2	104.7	101.0	101.0	80.0	80.0	83.7	80.7	80.6	80.6
	2500	100 ‡	100.0	74.5	107.2	100.6	100.6	68.4	68.4	68.0	69.4	68.5	68.5
12	250	100 ‡	100.0	4125	102.9	101.5	101.5	94.2	94.2	30256	95.8	95.3	95.3
	500	100 ‡	100.0	151.4	103.3	101.1	101.2	87.5	87.5	122.6	88.0	88.1	88.1
	1000	100 ‡	100.0	121.0	106.0	101.3	101.3	79.1	79.1	93.2	79.9	79.4	79.4
	2500	100 ‡	100.0	103.1	107.1	101.2	101.3	67.0	67.0	70.7	67.3	67.1	67.1

Note: ‡ = The reference method for the computation of the relative RMSE was “Mplus, FIXED” with $p = 0.5$ and $\varepsilon = 0.01$. Absolute RMSE values smaller than 100 are printed in bold font. After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

Table 6. Simulation Study: Mean absolute difference between different Mplus and sirt estimation methods of the estimated factor standard deviation in the second group σ_2 for $p = 0.5$ and $\varepsilon = 0.01$ as a function of sample size per group N and number of groups G .

G	N	Mplus, FIXED and sirt, meth				Mplus, FIXED and sirt, meth			
		1	2	3	4	1	2	3	4
3	250	0.0438	0.0084	0.0076	0.0076	0.0438	0.0084	0.0076	0.0076
	500	0.0300	0.0064	0.0051	0.0051	0.0300	0.0064	0.0051	0.0051
	1000	0.0222	0.0054	0.0037	0.0037	0.0222	0.0054	0.0037	0.0037
	2500	0.0186	0.0046	0.0024	0.0024	0.0186	0.0046	0.0024	0.0024
6	250	0.0850	0.0045	0.0025	0.0280	0.0411	0.0435	0.0415	0.0110
	500	0.0565	0.0059	0.0046	0.0715	0.0847	0.0903	0.0877	0.0145
	1000	0.0431	0.0050	0.0031	0.0651	0.0791	0.0831	0.0808	0.0134
	2500	0.0345	0.0046	0.0026	0.0490	0.0625	0.0657	0.0636	0.0122
9	250	0.1362	0.0045	0.0025	0.0365	0.0502	0.0529	0.0509	0.0119
	500	0.0907	0.0045	0.0025	0.0276	0.0411	0.0435	0.0415	0.0114
	1000	0.0680	0.0061	0.0041	0.0728	0.0881	0.0939	0.0914	0.0153
	2500	0.0537	0.0051	0.0029	0.0616	0.0766	0.0806	0.0784	0.0140
12	250	2.1033	0.0092	0.0069	0.0069	2.1033	0.0092	0.0069	0.0069
	500	0.1374	0.0065	0.0045	0.0045	0.1374	0.0065	0.0045	0.0045
	1000	0.0984	0.0057	0.0035	0.0035	0.0984	0.0057	0.0035	0.0035
	2500	0.0766	0.0046	0.0023	0.0023	0.0766	0.0046	0.0023	0.0023

Note: After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

5. Empirical Example: Asparouhov and Muthén (2014) Dataset

This empirical example uses a dataset that was previously also analyzed in [5,40,41]. The dataset came from the European social survey (ESS) conducted in the year 2005 (ESS 2005), which included subjects from 26 countries. The factor variable of tradition and conformity was assessed by four items presented in portrait format, where the scale of the items is such that a high value represents a low level of tradition conformity. The wording of the four items were as follows (see [5]): “It is important for him to be humble and modest. He tries not to draw attention to himself” (item TR9); “Tradition is important to him. He tries to follow the customs handed down by his religion or family” (item TR20); “He believes that people should do what they’re told. He thinks people should follow rules at all times, even when no one is watching” (item C07); and “It is important for him to always behave properly. He wants to avoid doing anything people would say is wrong” (item C016). The dataset for this empirical example (and used in [5]) was downloaded from <https://www.statmodel.com/Alignment.shtml> (accessed on 17 February 2024).

5.1. Original Data

We analyzed the original ESS dataset but included subjects with no missing values on the four items. The dataset used in this article can be found at <https://osf.io/84ne5> (accessed on 17 February 2024). In the 26 countries, the sample sizes ranged between 1031 and 2963 persons with a mean of 1869.5 and an SD of 454.7. The IA method was applied with the specifications $p = 0.5$ and $\varepsilon = 0.01$ in Mplus and sirt. The same six estimation methods (i.e., FIXED and FREE in Mplus as well as meth = 1, meth = 2, meth = 3, and meth = 4 in sirt) were applied to the dataset.

Table 7 shows the estimated factor means and SDs for the 26 countries and the six estimation methods. It can be seen that sirt’s default meth = 1 provides implausible estimates in this example with many groups. However, the sirt methods meth = 2, meth = 3, and meth = 4 performed comparably to Mplus’ FIXED and FREE methods. It turned out that Mplus’ FIXED method was relatively close to sirt’s meth = 3 in terms of absolute differences in estimated factor means ($M = 0.010$, $SD = 0.013$, $Min = 0.000$, $Max = 0.070$). In addition, estimated factor means were also similar between the Mplus FIXED method and the sirt meth = 2 method (absolute differences: $M = 0.012$, $SD = 0.014$, $Min = 0.000$, $Max = 0.068$).

Moreover, Mplus' FREE method also performed similarly to sirt's meth = 4 for estimated factor means (absolute differences: $M = 0.010$, $SD = 0.016$, $Min = 0.000$, $Max = 0.086$). There was also a close resemblance for estimated factor standard deviations between the Mplus FIXED and sirt meth = 3 methods (absolute differences: $M = 0.007$, $SD = 0.006$, $Min = 0.000$, $Max = 0.020$). However, the differences between the estimation methods FIXED and FREE in Mplus (or meth = 3 and meth = 4 in sirt) are noteworthy.

Table 7. Empirical Example, Original Data: Factor means and factor standard deviations estimates of invariance alignment for 26 countries estimated estimated with Mplus and sirt.

Country	N	Mean						Standard Deviation					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
1	1525	0	0	0	0	0	0	1	1	1	1	1	1
2	1695	0.079	0.026	0.470	0.074	0.074	0.036	0.994	0.994	6.270	0.989	0.993	0.992
3	2320	-0.432	-0.474	-2.958	-0.438	-0.443	-0.474	1.109	1.109	7.396	1.096	1.107	1.107
4	1468	0.263	0.205	1.671	0.260	0.262	0.220	1.086	1.086	6.973	1.085	1.095	1.095
5	1031	-0.579	-0.635	-3.841	-0.588	-0.589	-0.626	0.987	0.987	6.456	0.988	0.990	0.989
6	2296	0.134	0.085	0.786	0.121	0.121	0.085	1.081	1.081	7.002	1.078	1.076	1.076
7	2963	0.316	0.268	2.007	0.305	0.309	0.274	1.135	1.135	7.388	1.123	1.137	1.137
8	1550	0.168	0.105	1.030	0.159	0.164	0.121	1.098	1.098	6.989	1.076	1.112	1.112
9	1793	0.152	0.107	0.930	0.143	0.143	0.108	0.989	0.989	6.447	0.992	0.989	0.989
10	1857	-0.245	-0.293	-1.605	-0.248	-0.251	-0.287	0.992	0.992	6.336	0.979	0.989	0.989
11	1630	0.353	0.311	2.213	0.335	0.336	0.302	1.170	1.170	7.677	1.161	1.164	1.164
12	1703	0.311	0.255	1.942	0.305	0.309	0.266	1.210	1.210	7.515	1.179	1.197	1.197
13	2356	0.106	0.060	0.636	0.099	0.101	0.065	1.145	1.145	7.286	1.128	1.157	1.157
14	2622	-0.424	-0.463	-2.783	-0.425	-0.423	-0.451	1.083	1.083	7.060	1.078	1.072	1.072
15	1562	-0.149	-0.185	-1.052	-0.158	-0.157	-0.185	1.138	1.138	7.475	1.127	1.118	1.118
16	1450	-0.232	-0.278	-1.554	-0.231	-0.231	-0.266	1.105	1.105	7.319	1.091	1.089	1.089
17	2361	0.083	0.031	0.495	0.077	0.076	0.039	1.374	1.374	8.890	1.377	1.369	1.369
18	2166	-0.303	-0.360	-2.090	-0.320	-0.323	-0.361	1.100	1.100	7.145	1.095	1.105	1.105
19	1770	0.334	0.287	2.097	0.327	0.325	0.289	1.065	1.065	6.833	1.066	1.059	1.059
20	1685	-0.288	-0.325	-2.009	-0.305	-0.302	-0.332	1.031	1.031	6.748	1.024	1.016	1.016
21	2120	0.283	0.207	2.376	0.351	0.353	0.293	0.971	0.971	6.501	0.960	0.964	0.964
22	2471	-0.080	-0.130	-0.582	-0.088	-0.088	-0.123	1.088	1.088	7.183	1.082	1.082	1.082
23	1439	0.878	0.822	5.334	0.835	0.877	0.833	1.136	1.136	6.950	1.088	1.143	1.140
24	1358	-0.397	-0.454	-2.669	-0.402	-0.405	-0.444	0.917	0.917	6.033	0.909	0.915	0.914
25	1783	-0.330	-0.377	-2.274	-0.326	-0.333	-0.368	0.997	0.997	6.669	0.955	0.977	0.977
26	1632	0.159	0.110	0.909	0.142	0.140	0.103	1.241	1.241	8.001	1.245	1.234	1.234

Note: N = sample size per country. Table entries with absolute differences smaller than 0.01 between the methods "Mplus, FIXED" and "sirt, meth = 3" are displayed in a gray-colored background. Table entries with absolute differences smaller than 0.01 between the methods "Mplus, FREE" and "sirt, meth = 4" are displayed in a yellow-colored background. After a linear transformation of the obtained parameter estimates, the first country had a factor mean of 0 and a factor standard deviation of 1.

5.2. Pseudo-Datasets

In this section, the original ESS dataset is used to create pseudo-datasets that should provide more insights about the different behavior of the estimation methods implemented in Mplus and sirt. The first five countries from the original datasets with sample sizes 1525, 1695, 2320, 1468, and 1031 subjects are used in the creation of the datasets. It is investigated whether the size of the estimates depends on the number of groups. To enable clean but idealized settings, we varied the number of included groups by replicating the original dataset accordingly. For example, with $G = 10$ groups, the first five groups were the original five countries, while groups six to ten are also the five countries but labeled as unique groups in the IA estimation. Usually, one would expect that the results of the first five groups should not change if the same dataset appears as duplications in the pseudo-dataset.

Table 8 presents estimated factor means and SDs for the third and the sixth group in the pseudo-datasets involving $G = 5, 10, 15, 20, 25,$ or 30 groups. Note that the sixth group

coincided with the first group in the pseudo-datasets and the first country in the original dataset. The distribution parameter estimates were transformed such that the mean and the SD of the first group were 0 and 1, respectively.

The factor mean estimates changed as a function of a number of groups for the Mplus FIXED and all sirt methods. Only for the Mplus FREE method were the estimates invariant with respect to the number of groups. In particular, large differences in the estimates were observed when comparing results in a model with 25 and 30 groups. Because the first group had a (transformed) mean of 0, it would also be expected that Group 6 would have factor mean estimates of 0. However, this was not the case for the estimation method, except for Mplus' FREE and sirt's meth = 4 methods. Overall, this pattern is surprising because it implies that the choice of the reference group (i.e., the first group in our case) and the number of groups strongly affect the estimates of factor means. For the SD, only sirt's meth = 1 had estimates that depended on the number of groups.

Table 8. Empirical Example, Pseudo-Datasets: Factor mean and factor standard deviations estimates of invariance alignment for Groups 3 and 6 estimated with Mplus and sirt.

Group	G	Mean						Standard Deviation					
		Mplus		sirt, meth=				Mplus		sirt, meth=			
		FIXED	FREE	1	2	3	4	FIXED	FREE	1	2	3	4
3	5	−0.371	−0.510	−0.422	−0.393	−0.408	−0.470	1.045	1.045	1.115	1.041	1.078	1.078
	10	−0.358	−0.510	−0.416	−0.375	−0.389	−0.462	1.045	1.045	1.153	1.041	1.078	1.079
	15	−0.345	−0.510	−0.412	−0.357	−0.370	−0.455	1.045	1.045	1.201	1.041	1.078	1.079
	20	−0.330	−0.510	−0.413	−0.337	−0.349	−0.449	1.045	1.045	1.277	1.040	1.078	1.079
	25	−0.310	−0.510	−0.441	−0.314	−0.325	−0.445	1.045	1.045	1.462	1.040	1.078	1.079
	30	−0.113	−0.510	−0.620	−0.286	−0.296	−0.441	1.045	1.045	2.258	1.041	1.078	1.080
6	10	0.027	0.000	0.042	0.040	0.040	0.000	1.000	1.000	1.063	1.000	1.000	1.000
	15	0.042	0.000	0.067	0.060	0.060	0.000	1.000	1.000	1.105	1.000	1.000	1.000
	20	0.059	0.000	0.096	0.082	0.082	0.000	1.000	1.000	1.171	1.000	1.000	1.000
	25	0.082	0.000	0.142	0.106	0.106	0.000	1.000	1.000	1.334	1.000	1.000	1.000
	30	0.290	0.000	0.277	0.137	0.137	0.000	1.000	1.000	2.026	1.000	1.000	1.000

Note: G = number of groups. After a linear transformation of the obtained parameter estimates, the first group had a factor mean of 0 and a factor standard deviation of 1.

6. Discussion

In this article, we compared the performance of IA estimates of the Mplus software and the R package sirt. There are two alternative identification constraints for estimating standard deviations ψ_g . Mplus uses the product constraint $\prod_{g=1}^G \psi_g = 1$, which is used in the sirt methods meth = 3 and meth = 4. However, one can alternatively fix the standard deviation of the first group to 1. This is the default in the R package sirt (i.e., meth = 1). The differences between Mplus and the IA function in the sirt package can primarily be traced back to the different identification constraints for standard deviations. The difference between Mplus and sirt can be made smaller by choosing meth = 3, which mimics the identification constraint used in Mplus. Notably, the latter method is preferred for a large(r) number of groups (say, more than eight), while the default of meth = 1 might be preferable for at most six groups. The simulation study and the empirical example demonstrated that the default meth = 1 in the sirt package does not provide trustworthy results, and users are strongly recommended switching to meth = 2 or meth = 3.

Overall, it turned out in the simulation study that the tuning parameter $\epsilon = 0.001$ generally outperforms the default Mplus choice $\epsilon = 0.01$. A previous study indicated that the choice of ϵ is more critical than the choice between the power $p = 0.5$ or $p = 0.25$ [15]. Minor reductions regarding bias can be obtained with the power $p = 0.25$ instead of $p = 0.5$. However, for reasonably large sample sizes (e.g., more than 500 subjects per group), an L_0 loss function [42] can even outperform the L_p loss function for $p = 0.5$ or $p = 0.25$ [15].

Regardless of the use of a particular estimation method in Mplus or sirt, we wonder whether the optimization function of IA is suitable in the case of many groups. The pairwise differences between model parameters in the optimization might lead to less stable estimates than a linear model specification that does not involve pairwise differences. There is some evidence that Haberman linking with the $L_{0.5}$ IA loss function could be superior in the estimation of many groups (say, more than 20 groups) in IA (see [35]). Further research is needed to explore possible adaptations of the IA method in the case of many groups.

In this article, we only examined estimation differences between Mplus and sirt for normally distributed data. It can be expected that estimation differences due to different identification constraints would similarly be present for ordinal data [6] because it uses item loadings and item thresholds from item response theory models instead of item loadings and item intercepts from a one-dimensional factor model based on the multivariate normal distribution as the model input.

The IA method can provide consistent estimation of factor means and standard deviations if there is a sparse pattern of parameters that are noninvariant across groups. It is debatable whether such a sparse pattern of noninvariant effects can be theoretically assumed in empirical datasets [43,44]. However, if researchers believe in such a sparsity assumption, IA can be deemed an effective data-driven method.

The simulation study conducted in this article assumed a sparse structure of noninvariant parameters. It could be that the differences between Mplus and the IA function in the sirt package were larger under different data-generating models. Future research could further investigate the software differences for more data-generating models and could also involve scenarios of a large number of groups.

As a cautionary remark, we would like to add that enough implementation details must appear in publications for commercial black-box software like Mplus to enable independent judgment, evaluation, and reimplementing of existing methods. We believe that non-documented or sparsely documented modeling approaches in commercial software, like the IA method in Mplus, should not be used in substantive and methodological publications because it fundamentally contradicts the principles of open science.

Funding: This research received no external funding.

Data Availability Statement: Datasets and R code are available at <https://osf.io/84ne5> (accessed on 17 February 2024).

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFA	confirmatory factor analysis
DGM	data-generating model
IA	invariance alignment
RMSE	root mean square error
SD	standard deviation

References

1. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **1993**, *58*, 525–543. [CrossRef]
2. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* **1989**, *13*, 127–143. [CrossRef]
3. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
4. van de Vijver, F.J.R. *Invariance Analyses in Large-Scale Studies*; OECD: Paris, France, 2019. [CrossRef]
5. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* **2014**, *21*, 495–508. [CrossRef]
6. Muthén, B.; Asparouhov, T. IRT studies of many groups: The alignment method. *Front. Psychol.* **2014**, *5*, 978. [CrossRef] [PubMed]
7. Ciecuch, J.; Davidov, E.; Schmidt, P. Alignment optimization. Estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In *Cross-Cultural Analysis: Methods and Applications*; Davidov, E., Schmidt, P., Billiet, J., Eds.; Routledge: London, UK, 2018; pp. 571–592. [CrossRef]

8. Pokropek, A.; Davidov, E.; Schmidt, P. A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Struct. Equ. Model. Multidiscip. J.* **2019**, *26*, 724–744. [[CrossRef](#)]
9. Leitgöb, H.; Seddig, D.; Asparouhov, T.; Behr, D.; Davidov, E.; De Roover, K.; Jak, S.; Meitinger, K.; Menold, N.; Muthén, B.; et al. Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Soc. Sci. Res.* **2023**, *110*, 102805. [[CrossRef](#)] [[PubMed](#)]
10. Luong, R.; Flake, J.K. Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychol. Methods* **2023**, *28*, 905–924. [[CrossRef](#)]
11. Sideridis, G.; Alghamdi, M.H. Bullying in middle school: Evidence for a multidimensional structure and measurement invariance across gender. *Children* **2023**, *10*, 873. [[CrossRef](#)]
12. Tsaousis, I.; Jaffari, F.M. Identifying bias in social and health research: Measurement invariance and latent mean differences using the alignment approach. *Mathematics* **2023**, *11*, 4007. [[CrossRef](#)]
13. Pokropek, A.; Lüdtke, O.; Robitzsch, A. An extension of the invariance alignment method for scale linking. *Psychol. Test Assess. Model.* **2020**, *62*, 303–334. Available online: <https://bit.ly/2UEp9GH> (accessed on 17 February 2024).
14. Mansolf, M.; Vreeker, A.; Reise, S.P.; Freimer, N.B.; Glahn, D.C.; Gur, R.E.; Moore, T.M.; Pato, C.N.; Pato, M.T.; Palotie, A.; et al. Extensions of multiple-group item response theory alignment: Application to psychiatric phenotypes in an international genomics consortium. *Educ. Psychol. Meas.* **2020**, *80*, 870–909. [[CrossRef](#)] [[PubMed](#)]
15. Robitzsch, A. Implementation aspects in invariance alignment. *Stats* **2023**, *6*, 1160–1178. [[CrossRef](#)]
16. Muthén, L.; Muthén, B. *Mplus User's Guide*; Version 8.9, 1998–2023; Muthén & Muthén: Los Angeles, CA, USA, 2023.
17. Kim, E.S.; Cao, C.; Wang, Y.; Nguyen, D.T. Measurement invariance testing with many groups: A comparison of five approaches. *Struct. Equ. Model. Multidiscip. J.* **2017**, *24*, 524–544. [[CrossRef](#)]
18. Lai, M.H.C.; Liu, Y.; Tse, W.W.Y. Adjusting for partial invariance in latent parameter estimation: Comparing forward specification search and approximate invariance methods. *Behav. Res. Methods* **2022**, *54*, 414–434. [[CrossRef](#)] [[PubMed](#)]
19. Muthén, B.; Asparouhov, T. Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociol. Methods Res.* **2018**, *47*, 637–664. [[CrossRef](#)]
20. DeMars, C.E. Alignment as an alternative to anchor purification in DIF analyses. *Struct. Equ. Model. Multidiscip. J.* **2020**, *27*, 56–72. [[CrossRef](#)]
21. Finch, W.H. Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Appl. Meas. Educ.* **2016**, *29*, 30–45. [[CrossRef](#)]
22. Flake, J.K.; McCoach, D.B. An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Struct. Equ. Model. Multidiscip. J.* **2018**, *25*, 56–70. [[CrossRef](#)]
23. Byrne, B.M.; van de Vijver, F.J.R. The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema* **2017**, *29*, 539–551. [[CrossRef](#)]
24. Marsh, H.W.; Guo, J.; Parker, P.D.; Nagengast, B.; Asparouhov, T.; Muthén, B.; Dicke, T. What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychol. Methods* **2018**, *23*, 524–545. [[CrossRef](#)]
25. Kim, E.; Cao, C.; Liu, S.; Wang, Y.; Dedrick, R. Testing measurement invariance over time with intensive longitudinal data and identifying a source of non-invariance. *Struct. Equ. Model. Multidiscip. J.* **2023**, *30*, 393–411. [[CrossRef](#)]
26. Lai, M.H.C. Adjusting for measurement noninvariance with alignment in growth modeling. *Multivar. Behav. Res.* **2023**, *58*, 30–47. [[CrossRef](#)] [[PubMed](#)]
27. Winter, S.D.; Depaoli, S. An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size. *Int. J. Behav. Dev.* **2020**, *44*, 371–382. [[CrossRef](#)]
28. Asparouhov, T.; Muthén, B. Penalized Structural Equation Models. *Struct. Equ. Model. A Multidiscip. J.* **2023**. [[CrossRef](#)]
29. Robitzsch, A. *Sirt: Supplementary Item Response Theory Models*. 2024. R Package Version 4.1-15. Available online: <https://CRAN.R-project.org/package=sirt> (accessed on 6 February 2024).
30. Bartholomew, D.J.; Knott, M.; Moustaki, I. *Latent Variable Models and Factor Analysis: A Unified Approach*; Wiley: New York, NY, USA, 2011. [[CrossRef](#)]
31. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [[CrossRef](#)]
32. van de Schoot, R.; Kluytmans, A.; Tummers, L.; Lugtig, P.; Hox, J.; Muthén, B. Facing off with scylla and charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* **2013**, *4*, 770. [[CrossRef](#)] [[PubMed](#)]
33. Byrne, B.M.; Shavelson, R.J.; Muthén, B. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* **1989**, *105*, 456–466. [[CrossRef](#)]
34. Haberman, S.J. *Linking Parameter Estimates Derived from an Item Response Model through Separate Calibrations*; Research Report No. RR-09-40; Educational Testing Service: Princeton, NJ, USA, 2009. [[CrossRef](#)]
35. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. [[CrossRef](#)]
36. R Core Team. *R: A Language and Environment for Statistical Computing*. 2023. Available online: <https://www.R-project.org/> (accessed on 15 March 2023).

37. Rudnev, M. Alignment Method for Measurement Invariance: Tutorial. Internet Blog Entry. 2019. Available online: <http://tinyurl.com/mry3vw99> (accessed on 17 February 2024).
38. Fischer, R.; Karl, J.A. A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Front. Psychol.* **2019**, *10*, 1507. [[CrossRef](#)]
39. Han, H. Using measurement alignment in research on adolescence involving multiple groups: A brief tutorial with R J. *Res. Adolesc.* **2023**, *34*, 235–242. [[CrossRef](#)]
40. Knoppen, D.; Saris, W. Do we have to combine values in the Schwartz' human values scale? A comment on the Davidov studies. *Surv. Res. Methods* **2009**, *3*, 91–103. [[CrossRef](#)]
41. Beierlein, C.; Davidov, E.; Schmidt, P.; Schwartz, S.H.; Rammstedt, B. Testing the discriminant validity of Schwartz' portrait value questionnaire items—A replication and extension of Knoppen and Saris (2009). *Surv. Res. Methods* **2012**, *6*, 25–36. [[CrossRef](#)]
42. O'Neill, M.; Burke, K. Variable selection using a smooth information criterion for distributional regression models. *Surv. Res. Methods* **2023**, *33*, 71. [[CrossRef](#)]
43. Robitzsch, A. Model-robust estimation of multiple-group structural equation models. *Algorithms* **2023**, *16*, 210. [[CrossRef](#)]
44. Robitzsch, A.; Lüdtke, O. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Struct. Equ. Model. Multidiscip. J.* **2023**, *30*, 859–870. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.