Comparing Robust Haberman Linking and Invariance Alignment

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: Linking methods are widely used in the social sciences to compare group differences regarding the mean and the standard deviation of a factor variable. This article examines a comparison between robust Haberman linking (HL) and invariance alignment (IA) for factor models with dichotomous and continuous items, utilizing the $L_{0.5}$ and $L_0$ loss functions. A simulation study demonstrates that HL outperforms IA when item intercepts are used for linking, rather than the original HL approach, which relies on item difficulties. The results regarding the choice of loss function were mixed: $L_0$ showed superior performance in the simulation study with continuous items, while $L_{0.5}$ performed better in the study with dichotomous items.

Keywords: linking; Haberman linking; invariance alignment; factor model; differential item functioning; item response model; $L_p$ loss function; $L_0$ loss function

1. Introduction

When comparing multiple groups in confirmatory factor analysis (CFA; [1]) or item response theory (IRT; [2,3]) models with respect to a factor variable, certain identification assumptions are required. A common approach assumes that item parameters remain consistent across groups, a property referred to as measurement invariance [4,5]. This principle is frequently discussed in the social sciences [5,6]. In the IRT literature, the lack of invariance is called differential item functioning (DIF; [7–11]).

To address violations of measurement invariance, the invariance alignment (IA) method [12–15], also known as alignment optimization [16,17], has been proposed for comparing groups in unidimensional factor models. The IA method aims to maximize the number of item parameters that are (approximately) invariant while permitting limited deviations. This approach enhances the robustness of group comparisons in the presence of measurement invariance violations. The IA method is frequently applied in social sciences for analyzing questionnaire data [18–25].

The Haberman linking (HL; [26]) method, as a particular linking method [27–29], has been developed for group comparisons in IRT models. While it is widely applied in educational testing, it remains largely unfamiliar to practitioners in other fields (but see [30–34] for exceptions). This trend is also reflected in the citation counts of the original sources of IA and HL. The original IA article of Asparouhov and Muthén (2014, Structural Equation Modeling [12]) has been cited 849 times according to Google Scholar, 440 times according to Web of Science, and 392 times according to CrossRef (as of 22 November 2024). In contrast, the original HL article of Haberman (2009, ETS Research Report [26]) has only 80 citations according to Google Scholar and 32 citations according to CrossRef (as of 22 November 2024).
Notably, HL appears to share significant similarities with the invariance alignment (IA) method. As noted in [35], researcher “Matthias von Davier, in particular, also highlighted how the alignment optimization [i.e., invariance alignment] method is very similar to the simultaneous test-linking approach proposed by Haberman (2009)” [35].

This article examines a comparison between HL and the IA method. Specifically, a robust version of HL, termed robust HL, is contrasted with IA in the presence of fixed differential item functioning (DIF) effects, which follows a sparsity structure; that is, only a subset (i.e., the minority) of item parameters exhibit DIF. This article builds on previous work of the author in a precursor article [36], published in 2020 in the Stats journal. To facilitate the comparison between the HL and IA linking methods, the robust loss functions $L_{0.5}$ and $L_0$ are employed in both HL and IA. A recently proposed differentiable approximation of the $L_0$ loss function is applied for HL and IA for the first time. As in [36], HL was modified to enable linking based on item intercepts instead of item difficulties.

The findings in this article, from two simulation studies, show that the modified HL specification with linking based on item intercepts closely resembles IA and, in several contexts, outperforms it. Moreover, in contrast to previous work in [36], the simulation studies in this article involved only three groups, rather than a larger number of groups, and manipulated the DIF effect size and the number of items to more thoroughly investigate the generalizability of the findings from the comparison of HL and IA. Previous work pointed out that the performance of different loss functions strongly depends on the size of the DIF effect sizes, where more robust (i.e., less biased) results are typically obtained with larger DIF effect sizes. In addition, more robust results can be expected with a larger number of items.

The structure of the article is as follows. Section 2 reviews unidimensional factor models for dichotomous and continuous items. The robust $L_{0.5}$ and $L_0$ functions are discussed in Section 3. Descriptions of IA and HL with robust loss functions are provided in Sections 4 and 5, respectively. Section 6 presents results from a simulation study with dichotomous items, while Section 7 discusses findings from a simulation study using continuous items. Finally, Section 8 concludes the article with a discussion.

2. Unidimensional Factor Model

This section discusses unidimensional factor models for dichotomous items in Section 2.1 and for continuous items in Section 2.2. Let $X_g = (X_{1g}, \ldots, X_{Ig})$ denote the vector of $I > 1$ items in group $g = 1, \ldots, G$. The vector of items $X$ is related to a normally distributed factor variable $\theta_g$, commonly referred to as the trait or ability variable in IRT. It is assumed that $\theta_g$ is normally distributed with mean $\mu_g$ and standard deviation (SD) $\sigma_g$ in group $g$. For identification reasons, we assume that $\mu_1 = 0$ and $\sigma_1 = 1$ holds in the first group. In the unidimensional factor models, it is assumed that items $X_{ig}$ are conditionally independent given the latent variable $\theta_g$. This implies that

$$P(X_{1g}, \ldots, X_{Ig} | \theta_g) = \prod_{i=1}^{I} P(X_{ig} | \theta_g).$$

(1)

Equation (1) represents the local independence assumption in IRT for dichotomous items, and it implies uncorrelated residual variables in the factor model for continuous items.

2.1. Dichotomous Items

We now describe the unidimensional factor model for dichotomous items $X_g \in \{0, 1\}^I$. The function $p_{ig}(\theta) = P(X_{ig} = 1 | \theta_g)$ for $i = 1, \ldots, I$ in group $g$ is referred to as the
item response function (IRF). The IRF of the two-parameter logistic (2PL) model [37] is defined as
\[
P_{ig}(\theta) = \Psi \left( a_{ig}(\theta - b_{ig}) \right) \text{ with } \theta \sim N(\mu_g, \sigma^2_g),
\] (2)
where \(a_{ig}\) and \(b_{ig}\) denote the item discrimination and the item difficulty \(b_{ig}\), respectively, and \(\Psi(x) = (1 + \exp(-x))^{-1}\) is the logistic distribution function.

The 2PL model in (2) can be equivalently expressed with item intercepts \(v_{ig}\) instead of item difficulties \(b_{ig}\). Then, the 2PL model is parametrized as
\[
P_{ig}(\theta) = \Psi \left( a_{ig}\theta + v_{ig} \right) \text{ with } \theta \sim N(\mu_g, \sigma^2_g),
\] (3)
where \(v_{ig} = -a_{ig}b_{ig}\), or \(b_{ig} = -a_{ig}/v_{ig}\).

2.2. Continuous Items

The unidimensional factor model for continuous items in CFA is defined as
\[
X_{ig} = v_{ig} + a_{ig}\theta + \xi_{ig}\text{ with } \theta \sim N(\mu_g, \sigma^2_g)\text{ and } \xi_{ig} \sim N(0, \omega_{ig}),
\] (5)
where \(v_{ig}\) is the item intercept and \(a_{ig}\) is the item discrimination, respectively. The residual variables \(\xi_{ig}\) are uncorrelated with the factor variable \(\theta\).

The parameters of the unidimensional factor model (5) can be estimated with maximum likelihood [1,41]. When estimating the unidimensional model (5) separately in group \(g\), the identification constraints \(\mu_g = 0\) and \(\sigma_g = 1\) must again be applied. Under this specification, the identified item parameters \(\hat{a}_{ig}\) and \(\hat{b}_{ig}\) are expressed as
\[
\hat{a}_{ig} = a_{ig}\sigma_g \text{ and } \hat{b}_{ig} = \frac{b_{ig} - \mu_g}{\sigma_g}.
\] (4)

3. \(L_{0.5}\) and \(L_0\) Loss Functions

In this paper, DIF effects are treated as outliers that should be excluded from group comparisons [42–48]. Group differences are calculated using robust location measures, which are estimated via robust loss functions [49–51].

A flexible class of such a robust loss function is the \(L_p\) (with \(p > 0\)) loss function, refs. [36,52–55] defined as
\[
\rho(x) = |x|^p\text{ for } p > 0.
\] (7)
The case \(p = 2\) (i.e., the \(L_2\) loss function) corresponds to the widely known square loss function, while \(p = 1\) (i.e., the \(L_1\) loss function) corresponds to median regression [56,57]. However, the loss function \(\rho\) is not differentiable at \(x = 0\) for \(p \leq 1\). To address this, a differentiable approximation of \(\rho\) is given by
\[
\tilde{\rho}(x) \simeq \left(|x|^2 + \epsilon\right)^{p/2},
\] (8)
where $\varepsilon > 0$ is a tuning parameter that controls the approximation error of $\tilde{\rho}$ relative to $\rho$ [12]. In practice, values such as $\varepsilon = 0.01$ (see [12] for $p = 0.5$ and $p = 0.25$) or $\varepsilon = 0.001$ [36,58] have been effectively applied in various contexts.

The left panel in Figure 1 depicts the exact $L_{0.5}$ loss function as a black solid line and its approximation as a red dashed line. The two functions are nearly identical except near the nondifferentiable point $x = 0$.

As an alternative to the $L_p$ loss function, the $L_0$ loss function [59,60] is defined as

$$\rho(x) = 1(x \neq 0),$$  \hspace{1cm} (9)

where $1$ denotes the indicator function. This loss function takes the value $0$ for $x = 0$ and $1$ for $x \neq 0$, effectively counting the number of deviations from $0$. The $L_0$ loss function is particularly well suited for scenarios involving partial invariance, where only a small number of DIF effects deviate from zero. An approximation of this loss function is given by (see [61])

$$\tilde{\rho}(x) \approx \frac{x^2}{x^2 + \varepsilon},$$  \hspace{1cm} (10)

where $\varepsilon > 0$ is a tuning parameter. The choice $\varepsilon = 0.01$ has been shown to perform well across various settings [62,63].

In the previous work of the author [36], the $L_0$ loss function was approximated using the $L_p$ loss function with a very small value of $p = 0.02$. However, in our experience based on simulation studies involving structural equation models [62], the $L_0$ approximation (10) outperforms the differentiable approximation of the $L_{0.02}$ loss function in (8) in terms of the precision of parameter estimates. Consequently, this article focuses solely on the $L_0$ loss function.

The right panel in Figure 1 depicts the exact $L_0$ loss function as a black solid line and its approximation as a red dashed line. Note that the $L_0$ function is even discontinuous at the point $x = 0$, and there are moderate deviations of the approximation $\tilde{\rho}$ from the exact $L_0$ loss function $\rho$.
The $L_p$ loss function with $p < 1$, such as $p = 0.5$, is preferred over $p = 1$ in situations involving asymmetric error distributions (e.g., outliers). For instance, all outlying DIF effects in one group might be either positive or negative, creating a scenario of unbalanced DIF. In such cases, the $L_1$ loss function has been found to perform inadequately.

4. Invariance Alignment

In this section, we review the estimation of the IA method proposed by Asparouhov and Muthén [12,14]. The IA method has been discussed for continuous items [12], dichotomous items [14], and polytomous items [64].

The originally proposed IA method estimates group means $\mu$ and group SDs $\sigma$ in a single step. However, it has been shown that the optimization can be split into two successive steps: first, estimating $\sigma$, and then estimating $\mu$, as in HL [36]. The IA method relies on estimated item discriminations $\hat{a}_{ig}$ (or the log-transformed item discriminations $\log \hat{a}_{ig}$ and item intercepts $\hat{\nu}_{ig}$ ($i = 1, \ldots, I; g = 1, \ldots, G$).

First, group SDs can be obtained by minimizing the optimization function

$$H_1(\sigma) = \sum_{i=1}^{I} \sum_{g \not= h} \rho \left( \frac{\hat{a}_{ig} - \hat{a}_{ih}}{\sigma_g - \sigma_h} \right),$$

where $\rho$ represents the $L_p$ or $L_0$ loss function. Note that the original IA method also involves weights for pairs of groups $g$ and $h$, which accounts for different sample sizes in the optimization function. We omit the inclusion of weights here, as these weights become irrelevant when all sample sizes per group are equal.

For identification purposes, one can either set $\sigma_1 = 1$ or apply the constraint $\prod_{g=1}^{G} \sigma_g = 1$; the latter constraint is implemented in the popular Mplus software [65].

As an alternative to the optimization function $H_1$ in (11), log-transformed group SDs $s = (s_1, \ldots, s_G)$ with $s_g = \log(\sigma_g)$ could alternatively be estimated based on log-transformed item discriminations $\log \hat{a}_{ig}$ by minimizing

$$\tilde{H}_1(s) = \sum_{i=1}^{I} \sum_{g \not= h} \rho (\log \hat{a}_{ig} - \log \hat{a}_{ih} - s_g + s_h),$$

where group SDs can be estimated as $\tilde{\sigma}_g = \exp(\tilde{s}_g)$ based on estimated values $\tilde{s}_g$ for $g = 1, \ldots, G$.

In the second step, group means $\mu$ are estimated based on previously estimated group SDs $\tilde{\sigma}$ by minimizing

$$H_2(\mu) = \sum_{i=1}^{I} \sum_{g \not= h} \rho \left( \hat{\nu}_{ig} - \hat{\nu}_{ih} - \frac{\hat{a}_{ig}}{\sigma_g} \mu_g + \frac{\hat{a}_{ih}}{\sigma_h} \mu_h \right).$$

Again, the originally proposed IA method uses the $L_{0.5}$ loss function, which employs the power $p = 0.5$.

In the Mplus implementation of the $L_p$ loss function with $p = 0.5$ and $p = 0.25$, the default tuning parameter in the differentiable approximation of the loss function is $\varepsilon = 0.01$ [36,63,65]. However, selecting $\varepsilon = 0.001$ may be beneficial in many contexts [63,66]. The choice $\varepsilon = 0.01$ in the $L_0$ loss function was found to yield the best performance [66].

5. Haberman Linking

The originally proposed HL method [26] estimates group means $\mu$ and group SDs $\sigma$ based on log-transformed item discriminations $\hat{a}_{ig}$ and item difficulties $\hat{b}_{ig}$.
In the first step, the log-transformed group standard deviations $s_g = \log \sigma_g$ are estimated ($g = 1, \ldots, G$). In the second step, group means $\mu_g$ are estimated. For identification purposes, the constraints $s_1 = 0$ (implying $\sigma_1 = 1$) and $\mu_1 = 0$ are imposed.

Log-transformed SDs $s$ and log-transformed joint item discriminations $\alpha = (\alpha_1, \ldots, \alpha_I)$, where $a_i = \exp(\alpha_i) \text{ for } i = 1, \ldots, I$, are obtained by minimizing the following optimization function:

$$H_1(s, \alpha) = \sum_{i=1}^{I} \sum_{g=1}^{G} \rho(\log \hat{\alpha}_{ig} - a_i - s_g),$$

where $\rho$ is the $L_p$ or $L_0$ loss function. The original HL method uses the $L_2$ loss function [26]. However, because distribution parameters need to be consistently estimated without being influenced by outlying DIF effects, this article employs differentiable approximations of robust loss functions, specifically $L_0.5$ and $L_0$.

Subsequently, group means $\mu$ and joint item difficulties $b = (b_1, \ldots, b_I)$ are determined by minimizing

$$H_2(\mu, b) = \sum_{i=1}^{I} \sum_{g=1}^{G} \rho(\hat{s}_g \hat{b}_{ig} - b_i + \mu_g),$$

where the estimated SDs $\hat{s}_g = \exp(\hat{s}_g)$ are inserted in the optimization function. Similar to the optimization function (14) for SDs, the original HL method uses the $L_2$ loss function.

As an alternative to the optimization function in (15), a modified optimization function in HL uses item intercepts $\hat{b}_{ig} = -\hat{\alpha}_{ig} \hat{b}_{ig}$ instead of item difficulties $\hat{b}_{ig}$ (see also [36]). The use of item intercepts is motivated by their role in IA. In this adaptation to HL, the group means $\mu$ and joint item intercepts $\nu = (\nu_1, \ldots, \nu_I)$ are determined by minimizing

$$\hat{H}_2(\mu, \nu) = \sum_{i=1}^{I} \sum_{g=1}^{G} \rho(\nu_{ig} - \nu_i - \frac{\hat{a}_{ig}}{\hat{s}_g} \mu_g).$$

There is a lack of research comparing the performance of the optimization functions $H_2$ and $\hat{H}_2$ in HL, particularly in comparing HL with these optimization functions to IA. As an exception, the author’s precursor article [36] included these two specifications, but they were only analyzed for a moderate or large number of groups, with a fixed DIF effect size and a fixed number of items. This article aims to provide additional insights into the performance of these two specifications for a smaller number of groups and more extended data-generating models.

6. Simulation Study 1: Dichotomous Items

6.1. Method

The 2PL model for $G = 3$ groups was used as the IRF in the data-generating model. The latent factor variable $\theta$ was normally distributed within each group, with group means $\mu_1 = 0$, $\mu_2 = 0.3$, and $\mu_3 = -0.5$. The SDs were $\sigma_1 = 1$, $\sigma_2 = 1.2$, and $\sigma_3 = 0.8$.

The simulation study was conducted for $I = 10$ and $I = 20$ items. Group-specific parameters $a_{ig}$ and $b_{ig}$ for each item $i = 1, \ldots, I$ and for groups $g = 1, 2, 3$ were based on fixed base item parameters, with DIF effects remaining constant across replications within a simulation condition. We now describe the item parameter choice in the $I = 10$ item condition. For the $I = 20$ condition, the item parameters were duplicated. For 10 items, the base item discriminations $a_i$ were all set to 1.00. The base item difficulties $b_i$ were set as $0.21, -1.47, 0.09, 0.55, -0.67, 0.77, 0.99, -1.75, 0.10, \text{ and } 1.17$, resulting in a mean item difficulty of 0.000 and an SD of 0.999. For the $I = 20$ condition, the item parameters were
duplicated. The base item parameters are also available at https://osf.io/kcnyu (accessed on 22 November 2024).

Fixed uniform DIF effects were assumed, meaning that DIF could occur in item difficulties, while item discriminations were not prone to DIF. Specifically, group-specific item difficulties \( b_{ig} \) were simulated as \( b_i + f_{ig} \delta \), where the factor \( f_{ig} \) was chosen as either 1 and \(-1\), indicating DIF items, or 0 for DIF-free items. In the first group, the factor was 1 for the first item and \(-1\) for the second item, while all other items did not exhibit DIF. In the second group, DIF effects were simulated for the third and fourth items with a factor of 1. In the third group, DIF effects were simulated for the fifth and sixth items with a factor of \(-1\). Note that the DIF effects were canceled out in the first group, but they were all positive in the second group and negative in the third group, resulting in unbalanced DIF. The size \( \delta \) of the DIF effect was set to 0, 0.3, and 0.6, representing no DIF, small DIF, and large DIF, respectively.

The sample sizes per group were selected as \( N = 250, 500, 1000, \) and 2000 to represent typical sample sizes in small- to large-scale testing applications of the 2PL model [67,68].

In each of the 4 (sample size \( N \)) \( \times 3 \) (DIF effect \( \delta \)) \( \times 2 \) (number of items \( I \)) = 24 cells of the simulation, 3000 replications were conducted. The 2PL model was first estimated separately in each of the three groups. Linking was performed in four specifications. First, IA was carried out using two approaches: the original method, based on untransformed item discriminations ([12]; method IA1, using optimization functions (11) and (13)), and an alternative method based on log-transformed item discriminations, similar to the approach used in HL (method IA2, using optimization functions (12) and (13)). In addition, we used the originally proposed HL ([26]; method HL1, using optimization functions (14) and (15)), which relies on item difficulties, and an HL implementation that employed item intercepts (method HL2, using optimization functions (14) and (16)). These four specifications were applied to both the \( L_{0.5} \) (i.e., \( p = 0.5 \)) and \( L_0 \) (i.e., \( p = 0 \)) loss functions, resulting in a total of 8 different linking methods. The \( L_2 \) loss function was not considered, as it has been shown in [36] to produce biased parameter estimates in IA and HL when unbalanced DIF is present. Furthermore, the simulated DIF effects were asymmetric, rendering the \( L_1 \) loss function ineffective in providing nearly unbiased linking estimates. For identification purposes, the distribution parameters in the first group were fixed at \( \mu_1 = 0 \) and \( \sigma_1 = 1 \), while \( \mu_g \) and \( \sigma_g \) were freely estimated for the second and third groups.

We evaluated the bias and the root mean square error (RMSE) for the \( \hat{\mu}_g \) and \( \hat{\sigma}_g \) estimates. Additionally, a relative RMSE was calculated, with IA1 using the \( L_{0.5} \) loss function (i.e., \( p = 0.5 \)) serving as the reference method (set to a relative RMSE of 100). To summarize the results across groups, the average absolute bias and average relative RMSE were computed for both the \( \mu_g \) and \( \sigma_g \) estimates across the two groups.

All analyses for this simulation study were carried out with the statistical software R (Version 4.4.1 [69]). The 2PL model was estimated with the function \sirt::rasch.mmi2()\ from the R package \sirt\ (Version 4.2-89 [70]). HL and IA were estimated with the \sirt::linking.haberman.lq()\ and \sirt::invariance.alignment()\ functions from the same R package, respectively. Replication material for this simulation study can be assessed at https://osf.io/kcnyu (accessed on 22 November 2024).

6.2. Results

Table 1 summarizes the average absolute bias and average relative RMSE for the estimated group means, \( \hat{\mu}_g \). When \( \delta = 0 \) (no DIF), all methods exhibited minimal bias, with values near 0.00 for larger sample sizes (\( N \)). As \( \delta \) increased, bias grew across all methods. IA1, IA2, and HL2 performed similarly in terms of bias, while HL1 demonstrated the poorest performance.
In terms of RMSE, IA2 closely resembled IA1 for \( p = 0.5 \) and \( p = 0 \). HL1 consistently showed the highest RMSE, although HL2 outperformed IA1 and IA2 in certain conditions with larger sample sizes. Increasing the number of items from 10 to 20 slightly reduced bias; however, bias converged to zero as sample sizes increased.

For the \( L_\delta \) loss function, \( p = 0 \) resulted in less bias than \( p = 0.5 \), but estimates based on \( L_0 \) were more variable, as reflected in higher RMSE values for most conditions. Overall, the HL2 method, as an alternative to HL1, was competitive with IA1 for estimating group means. HL2 offered particular advantages with large DIF effects (i.e., for \( \delta = 0.6 \)).

Table 1. Simulation Study 1 (dichotomous items): Average absolute bias and average relative root mean square error (RMSE) for estimated group means \( \hat{\mu}_g \) as a function of the DIF effect \( \delta \), number of items \( I \), and sample size \( N \).

| \( \delta \) | \( I \) | \( N \) | Average Absolute Bias | Average Relative RMSE |
|---|---|---|---|---|
|   |   |   | \( p = 0.5 \) & \( \hat{\mu}_g \) | \( p = 0 \) & \( \hat{\mu}_g \) | \( p = 0.5 \) & \( \hat{\mu}_g \) | \( p = 0 \) & \( \hat{\mu}_g \) |
| 0 | 10 | 250 | 0.02 0.02 0.00 0.02 0.02 0.02 0.00 0.02 | 100 101 | 114 102 | 122 121 | 140 123 |
|   |   | 500 | 0.01 0.02 0.00 0.02 0.01 0.01 0.00 0.01 | 100 101 | 113 101 | 118 119 | 136 118 |
|   |   | 1000 | 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 | 100 100 | 111 100 | 114 115 | 128 113 |
|   |   | 2000 | 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 | 100 100 | 108 99 | 108 108 | 118 106 |
| 0.3 | 10 | 250 | 0.07 0.07 0.06 0.07 0.06 0.07 0.06 0.07 | 100 101 | 117 102 | 121 122 | 146 121 |
|   |   | 500 | 0.06 0.06 0.06 0.06 0.06 0.06 0.06 0.06 | 100 101 | 115 101 | 118 121 | 144 119 |
|   |   | 1000 | 0.05 0.05 0.05 0.05 0.04 0.05 0.06 0.04 | 100 101 | 117 100 | 120 121 | 145 118 |
|   |   | 2000 | 0.03 0.03 0.03 0.03 0.02 0.02 0.04 0.02 | 100 100 | 128 98 | 112 112 | 158 108 |
| 0.6 | 10 | 250 | 0.10 0.10 0.11 0.10 0.08 0.08 0.10 0.08 | 100 101 | 120 101 | 121 122 | 152 120 |
|   |   | 500 | 0.06 0.06 0.08 0.06 0.03 0.03 0.07 0.03 | 100 100 | 129 100 | 119 169 | 164 115 |
|   |   | 1000 | 0.03 0.03 0.05 0.03 0.00 0.03 0.03 0.03 | 100 100 | 132 96 | 108 107 | 163 101 |
|   |   | 2000 | 0.02 0.02 0.02 0.02 0.00 0.00 0.00 0.00 | 100 100 | 122 95 | 101 101 | 128 95 |
| 10 | 250 | 0.09 0.09 0.10 0.09 0.07 0.07 0.10 0.06 | 100 100 | 117 101 | 120 120 | 150 119 |
|   |   | 500 | 0.06 0.06 0.07 0.05 0.02 0.02 0.05 0.02 | 100 100 | 121 98 | 108 108 | 147 105 |
|   |   | 1000 | 0.03 0.03 0.04 0.03 0.00 0.01 0.00 0.00 | 100 100 | 120 96 | 101 101 | 125 97 |
|   |   | 2000 | 0.02 0.02 0.02 0.02 0.00 0.00 0.00 0.00 | 100 100 | 111 96 | 96 96 | 105 92 |

Note. \( p = \) power used in the \( L_\delta \) loss function; IA1 = invariance alignment (IA) with untransformed item discriminations \( \hat{a}_{ig} \), as originally proposed by Asparouhov and Muthén (2014 [12]), using optimization functions (11) and (13); IA2 = IA based on log-transformed item discriminations \( \hat{a}_{ig} \), using (12) and (13); HL1 = Haberman linking (HL) based on item difficulties \( \hat{b}_{ig} \), as originally proposed by Haberman (2009 [26]), using (14) and (15); HL2 = HL based on item intercepts \( \hat{c}_{ig} \), using (14) and (16). \( \hat{b}_{ig} \) is the reference method for computing the relative RMSE. Absolute bias values larger than 0.025 are printed in bold font. Relative RMSE values larger than 104.5 are printed in bold font, while values smaller than 99.5 are printed in a gray background color.

Table 2 presents the average absolute bias and relative RMSE for the estimated SDs \( \hat{\sigma}_g \). Across all conditions, bias was small for all methods, with slightly better performance
observed at larger values of $I$ and $N$. This outcome was expected, as DIF was introduced only in item difficulties, not in item discriminations.

RMSE values were generally higher for the HL methods compared to the IA methods; however, these differences diminished as sample size $N$ increased. The $L_p$ loss function with $p = 0.5$ outperformed $p = 0$ in terms of relative RMSE.

Overall, HL2 performed similarly to, if not better than, IA1 for dichotomous items. Only a few conditions favored $p = 0$ over $p = 0.5$.

Table 2. Simulation Study 1 (dichotomous items): Average absolute bias and average relative root mean square error (RMSE) for estimated group standard deviations $\hat{\sigma}_g$ as a function of the DIF effect $\delta$, number of items $I$, and sample size $N$.

| $\delta$ | $I$ | $N$ | $p = 0.5$ | $p = 0$ | $p = 0.5$ | $p = 0$ |
|----------|-----|-----|-----------|---------|-----------|---------|
| $I$      | $N$ | IA1 | IA2 | HL1 | HL2 | IA1 | IA2 | HL1 | HL2 | IA1 | IA2 | HL1 | HL2 |
| 0.3      | 250 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 100 | 103 | 104 | 104 |
|          | 500 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 100 | 102 | 103 | 103 |
|          | 1000| 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 100 | 101 | 101 | 101 |
|          | 2000| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 100 | 99  | 99  |
| 0.6      | 250 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 100 | 102 | 104 | 104 |
|          | 500 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 100 | 102 | 102 | 102 |
|          | 1000| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 101 | 101 | 101 |
|          | 2000| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 100 | 99  | 99  |

Note. $p = \text{power used in the } L_p \text{ loss function}; \text{IA1 = invariance alignment (IA) with untransformed item discriminations } \hat{\alpha}_{ig} \text{ as originally proposed by Asparouhov and Muthén (2014 [12]), using optimization functions (11) and (13); IA2 = IA based on log-transformed item discriminations } \hat{\alpha}_{ig} \text{, using (12) and (13); HL1 = Haberman linking (HL) based on item difficulties } \hat{b}_{ig} \text{, as originally proposed by Haberman (2009 [26]); HL2 = HL based on item intercepts } \hat{\nu}_{ig} \text{, using (14) and (16).} \dagger \text{ The linking method IA1 with } p = 0.5 \text{ was the reference method for computing the relative RMSE. Absolute bias values larger than 0.025 are printed in bold font. Relative RMSE values larger than 104.5 are printed in bold font, while values smaller than 99.5 are printed in a gray background color.}

7. Simulation Study 2: Continuous Items

7.1. Method

In this second simulation study, HL and IA were compared for continuous items. The data-generating model followed the simulation from Asparouhov and Muthén [12].
Continuous item responses were simulated using a unidimensional factor model with $G = 3$ groups. The normally distributed factor variable was measured by either $I = 5$ or $I = 10$ normally distributed items.

The group means of the factor variable were set to 0, 0.3, and 1.0, while the group SDs were set to 1, 1.5, and 1.2, respectively. Item responses were generated with uniform DIF and nonuniform DIF. The following describes the item parameters in the condition with $I = 5$ items. Each group had one noninvariant item intercept, $\nu_{ig}$, and one noninvariant item discrimination, $a_{ig}$. For all groups, invariant item discriminations and residual variances of the indicator variables were set to $a_{ig} = 1$ ($i = 1, \ldots, 5$), and invariant item intercepts were set to $\nu_{ig} = 0$. The noninvariant item parameters in the first group were $\nu_{51} = 0.5$ and $a_{13} = 1.4$. The noninvariant item parameters in the second group were $\nu_{12} = -0.5$ and $a_{52} = 0.5$. The noninvariant item parameters in the third group were $\nu_{23} = 0.5$ and $a_{43} = 0.3$. In the $I = 10$ item condition, the item parameters were doubled.

The sample sizes $N = 250, 500, 1000, 2000$ were selected to reflect typical applications of the IA method [15,16].

In each of the $4 \times 2 = 8$ cells of the simulation, 3000 replications were performed. The unidimensional factor model was estimated in the first step using the R (Version 4.4.1 [69]) function `sirt::invariance_alignment_cfa_config()` from the R package sirt (Version 4.2-89 [70]). Subsequently, the linking methods HL1, HL2, IA1, and IA2, as described in Simulation Study 1 (see Section 6.1), were applied for the powers $p = 0.5$ and $p = 0$ in the $L_p$ loss function. Consistent with the previous simulation study, the average absolute bias and average relative RMSE were used as outcome measures. The method IA1 with $p = 0.5$ again served as the reference method for calculating the relative RMSE. As in Simulation Study 1, the HL and IA methods were implemented using the `sirt::linking.haberman.lq()` and `sirt::invariance.alignment()` functions from the same sirt package, respectively. Replication material for this simulation study is available at https://osf.io/kcnyu (accessed on 22 November 2024).

### 7.2. Results

Table 3 presents the average absolute bias and average relative RMSE for the estimated group means ($\hat{\mu}_g$) and group standard deviations ($\hat{\sigma}_g$). Both estimates improved in terms of bias as the sample size $N$ increased. For sample sizes of at least 1000, the bias across all methods became negligible (i.e., 0.01 or lower), particularly under $p = 0$. The HL2 method consistently outperformed HL1, especially at smaller $N$ and $I$, demonstrating the advantage of using item intercepts over item difficulties in HL. Additionally, HL2 outperformed both IA1 and IA2 across all conditions of this simulation study for $p = 0.5$ and $p = 0$. Lower RMSE values were observed for $p = 0$ compared to $p = 0.5$ across the four linking methods in many scenarios. The $L_0$ loss function performed less favorably than the $L_{0.5}$ loss function only at small sample sizes.

### Table 3. Simulation Study 2 (continuous items): Average absolute bias and average relative root mean square error (RMSE) for estimated group means $\hat{\mu}_g$ and estimated group standard deviations $\hat{\sigma}_g$ as a function of the number of items $I$, and sample size $N$.

| Par $I$ | $N$ | IA1 | IA2 | HL1 | HL2 | IA1 | IA2 | HL1 | HL2 | IA1$\dagger$ | IA2 | HL1 | HL2 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|----------|-----|-----|-----|
| $\hat{\mu}_g$ | 5 | 250 | 0.03 | 0.03 | 0.04 | 0.03 | 0.01 | 0.01 | 0.04 | 0.01 | 100 | 100 | 114 | 97 | 97 |
| | 500 | 0.02 | 0.02 | 0.02 | 0.02 | 0.00 | 0.00 | 0.02 | 0.00 | 100 | 100 | 116 | 96 | 102 | 124 |
| | 1000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 100 | 100 | 109 | 95 | 97 | 97 |
| | 2000 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 100 | 96 | 97 | 97 | 97 |
8. Discussion

In this paper, we compared the linking methods robust HL and IA for the $L_{0.5}$ (i.e., $p = 0.5$) and $L_0$ (i.e., $p = 0$) loss functions. It turned out that robust HL performed similarly, if not better than IA, if it was applied in a variant relying on item intercepts instead of item difficulties. Importantly, the originally proposed HL, which relies on item difficulties instead of item intercepts, was clearly inferior to IA. The findings for the loss function choice $L_{0.5}$ vs. $L_0$ were mixed. In our first simulation study, involving dichotomous items, RMSE values were larger for the $L_0$ loss function. In contrast, RMSE values were smaller for $L_0$ in the second simulation study, involving continuous items. However, in the two studies, the bias in estimated group means and SDs was smaller for $L_0$ than for $L_{0.5}$, indicating that $L_0$ always comes with higher variable parameter estimates.

As with any simulation study, our study also has several limitations. First, future studies might focus on comparing HL and IA for polytomous item responses. Second, we applied HL and IA to two-parameter models. The linking methods could also be applied in estimation settings for one-parameter models, namely, the Rasch model in IRT for dichotomous items [71] or the tau-equivalent measurement model for continuous items [72]. Alternatively, the item discriminations could be assumed as invariant across groups, while item intercepts are freely estimated in the first step. In the second step, only item intercepts are linked, reducing the uncertainty in estimated item discriminations during linking. Third, linking errors and total error estimation [73] could be developed and evaluated for IA.

In contrast to IA, both simulation studies revealed that additionally estimated item parameters in HL did not provide efficiency losses in the parameter estimates. Hence, HL might be preferred over IA for statistical reasons, and it has the advantage compared to IA that joint item parameters (i.e., that can be interpreted as item parameters in the partial invariance situation) are simultaneously estimated. In contrast, a subsequent estimation step is required in IA to determine joint item parameter estimates and DIF effects for items.
Finally, if robust HL or IA methods are utilized in the linking procedure, items with large DIF effects are essentially excluded from group comparisons. In this way, it is presupposed that the construct being measured will not be changed by eliminating these items from comparisons involving the factor variable. This view implies that these DIF items are deemed as construct-irrelevant. However, it might be dangerous to treat these DIF items with robust linking functions in group comparisons if they are essentially part of the construct (i.e., they are construct-relevant; see [74–78] but see [79,80] for alternative views).

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This article only uses simulated datasets. Replication material for creating the simulated datasets in the simulation studies can be found at https://osf.io/kcnyu (accessed on 22 November 2024).

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

**Abbreviations**

The following abbreviations are used in this manuscript:

- 2PL: Two-parameter logistic
- CFA: Confirmatory factor analysis
- DIF: Differential item functioning
- HL: Haberman linking
- IA: Invariance alignment
- IRF: Item response function
- IRT: Item response theory
- MML: Marginal maximum likelihood
- RMSE: Root mean square error
- SD: Standard deviation

**References**

1. Bartholomew, D.J.; Knott, M.; Moustaki, I. *Latent Variable Models and Factor Analysis: A Unified Approach*; Wiley: New York, NY, USA, 2011. [CrossRef]
2. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NJ, USA, 2021. [CrossRef]
3. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *arXiv* 2021, arXiv:2108.08604. [CrossRef]
4. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 1993, 58, 525–543. [CrossRef]
5. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
6. Mellenbergh, G.J. Item bias and item response theory. *Int. J. Educ. Res.* 1989, 13, 127–143. [CrossRef]
7. Bauer, D.J. Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *Br. J. Math. Stat. Psychol.* 2023, 76, 435–461. [CrossRef]
8. Holland, P.W.; Wainer, H., (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [CrossRef]
9. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics*; Volume 26: Psychometrics; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 125–167. [CrossRef]
10. Thissen, D. A review of some of the history of factorial invariance and differential item functioning. *Multivar. Behav. Res.* 2024, *epub ahead of print*. [CrossRef]
11. Wells, C.S. *Assessing Measurement Invariance for Applied Research*; Cambridge University Press: Cambridge, UK, 2021. [CrossRef]
12. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* 2014, 21, 495–508. [CrossRef]
13. Asparouhov, T.; Muthén, B. Penalized structural equation models. *Struct. Equ. Model.* 2024, 31, 429–454. [CrossRef]
14. Muthén, B.; Asparouhov, T. IRT studies of many groups: The alignment method. *Front. Psychol.* 2014, 5, 978. [CrossRef] [PubMed]
15. 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]

16. Cieciuch, 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: Oxfordshire, UK, 2018; pp. 571–592. [CrossRef]

17. 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.* 2019, 26, 724–744. [CrossRef]

18. Fischer, R.; Karl, J.A. A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Front. Psychol.* 2019, 10, 1507. [CrossRef]

19. Han, H. Using measurement alignment in research on adolescence involving multiple groups: A brief tutorial with R. *J. Res. Adolesc.* 2024, 34, 235–242. [CrossRef]

20. Lai, M.H.C. Adjusting for measurement noninvariance with alignment in growth modeling. *Multivar. Behav. Res.* 2023, 58, 30–47. [CrossRef]

21. 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]

22. 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] [PubMed]

23. Sideridis, G.; Alghamdi, M.H. Bullying in middle school: Evidence for a multidimensional structure and measurement invariance across gender. *Children* 2023, 10, 873. [CrossRef]

24. Tsousis, 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]

25. Wurster, S. Measurement invariance of non-cognitive measures in TIMSS across countries and across time. An application and comparison of multigroup confirmatory factor analysis, Bayesian approximate measurement invariance and alignment optimization approach. *Stud. Educ. Eval.* 2022, 73, 101143. [CrossRef]

26. 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]

27. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; (Research Report No. RR-09-40); Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]

28. Lee, W.C.; Lee, G. IRT linking and equating. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test*; Irving, P., Booth, T., Hughes, D.J., Eds.; Wiley: New York, NY, USA, 2018; pp. 639–673. [CrossRef]

29. Sansivieri, V.; Wiberg, M.; Matteucci, M. A review of test equating methods with a special focus on IRT-based approaches. *Statistica* 2017, 77, 329–352. [CrossRef]

30. Höft, L.; Bernholt, S. Longitudinal couplings between interest and conceptual understanding in secondary school chemistry: An activity-based perspective. *Int. J. Sci. Educ.* 2019, 41, 607–627. [CrossRef]

31. Liu, G.; Kim, H.J.; Lee, W.C.; Kim, Y. *Comparison of Simultaneous Linking and Separate Calibration with Stocking-Lord Method*; (CASMA Research Report Number 57); Center for Advanced Studies in Measurement and Assessment, University of Iowa: Iowa City, IA, USA, 2024. Available online: https://tinyurl.com/2bj6pbwn (accessed on 20 December 2024).

32. Moehring, A.; Schroeders, U.; Wilhelm, O. Knowledge is power for medical assistants: Crystallized and fluid intelligence as predictors of vocational knowledge. *Front. Psychol.* 2018, 9, 28. [CrossRef]

33. Neuenschwander, M.P.; Mayland, C.; Niederbacher, E.; Garrote, A. Modifying biased teacher expectations in mathematics and self-concept, and interest from grade 5 to 7. *Contemp. Educ. Psychol.* 2023, 55, 55–65. [CrossRef]

34. Sewasew, D.; Schroeders, U.; Schiefer, I.M.; Weirich, S.; Artest, C. Development of sex differences in math achievement, self-concept, and interest from grade 5 to 7. *Learn. Individ. Differ.* 2021, 87, 101995. [CrossRef]

35. Avvisati, F.; Le Donné, N.; Paccagnella, M. A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Meas. Instrum. Soc. Sci.* 2019, 1, 8. [CrossRef] [PubMed]

36. Robitzsch, A. $L_P$ loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* 2020, 3, 246–283. [CrossRef]

37. Birnbaum, A. Some latent trait models and their use in inferring an examinee’s ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; MIT Press: Reading, MA, USA, 1968; pp. 397–479.

38. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]

39. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 1981, 46, 443–459. [CrossRef]

40. Glas, C.A.W. Maximum-likelihood estimation. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 197–216. [CrossRef]
41. Basilevsky, A.T. Statistical Factor Analysis and Related Methods: Theory and Applications; Wiley: New York, NY, USA, 2009. [CrossRef]
42. De Boeck, P. Random item IRT models. Psychometrika 2008, 73, 533–559. [CrossRef]
43. Halpin, P.F. Differential item functioning via robust scaling. Psychometrika 2024, 89, 796–821. [CrossRef] [PubMed]
44. Magis, D.; De Boeck, P. Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. Multivar. Behav. Res. 2011, 46, 733–755. [CrossRef] [PubMed]
45. Robitzsch, A. Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: A comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. Symmetry 2021, 13, 2198. [CrossRef]
46. Robitzsch, A. Comparing robust linking and regularized estimation for linking two groups in the 1PL and 2PL models in the presence of sparse uniform differential item functioning. Stats 2023, 6, 192–208. [CrossRef]
47. Tutz, G.; Schauberger, G. A penalty approach to differential item functioning in Rasch models. Psychometrika 2015, 80, 21–43. [CrossRef] [PubMed]
48. Wang, W.; Liu, Y.; Liu, H. Testing differential item functioning without predefined anchor items using robust regression. J. Educ. Behav. Stat. 2022, 47, 666–692. [CrossRef]
49. Huber, P.J.; Ronchetti, E.M. Robust Statistics; Wiley: New York, NY, USA, 2009. [CrossRef]
50. Maronna, R.A.; Martin, R.D.; Yohai, V.J. Robust Statistics: Theory and Methods; Wiley: New York, NY, USA, 2006. [CrossRef]
51. Wilcox, R.R.; Keselman, H.J. Modern robust data analysis methods: Measures of central tendency. Psychol. Methods 2003, 8, 254–274. [CrossRef]
52. Giacalone, M.; Panarrello, D.; Mattera, R. Multicollinearity in regression: An efficiency comparison between $L_\infty$-norm and least squares estimators. Qual. Quant. 2018, 52, 1831–1859. [CrossRef]
53. Lipovetsky, S. Optimal $L_p$-metric for minimizing powered deviations in regression. J. Mod. Appl. Stat. Methods 2007, 6, 20. [CrossRef]
54. Livadiotis, G. General fitting methods based on $L_p$ norms and their optimization. Stats 2020, 3, 16–31. [CrossRef]
55. Sposito, V.A. On unbiased $L_p$ regression estimators. J. Am. Stat. Assoc. 1982, 77, 652–653.
56. Koenker, R.; Hallock, K.F. Quantile regression. J. Econ. Perspect. 2001, 15, 143–156. [CrossRef]
57. Koenker, R. Quantile regression: 40 years on. Annu. Rev. Econ. 2017, 9, 155–176. [CrossRef]
58. Robitzsch, A. Implementation aspects in regularized structural equation models. Algorithms 2023, 16, 446. [CrossRef]
59. Oelker, M.R.; Pößnecker, W.; Tutz, G. Selection and fusion of categorical predictors with $L_0$-type penalties. Stat. Model. 2015, 15, 389–410. [CrossRef]
60. Oelker, M.R.; Tutz, G. A uniform framework for the combination of penalties in generalized structured models. Adv. Data Anal. Classif. 2017, 11, 97–120. [CrossRef]
61. O’Neill, M.; Burke, K. Variable selection using a smooth information criterion for distributional regression models. Stat. Comput. 2023, 33, 71. [CrossRef] [PubMed]
62. Robitzsch, A. $L_0$ and $L_p$ loss functions in model-robust estimation of structural equation models. Psych 2023, 5, 1122–1139. [CrossRef]
63. Robitzsch, A. Examining differences of invariance alignment in the Mplus software and the R package sirt. Mathematics 2024, 12, 770. [CrossRef]
64. 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.; Palottie, 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]
65. Muthén, L.; Muthén, B. Mplus User’s Guide; Version 8.11; Muthén & Muthén: Los Angeles, CA, USA, 2024. Available online: https://www.statmodel.com/ (accessed on 22 November 2024).
66. Robitzsch, A. Implementation aspects in invariance alignment. Stats 2023, 6, 1160–1178. [CrossRef]
67. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J. (Eds.) Implementation of Large-Scale Education Assessments; Wiley: New York, NY, USA, 2017. [CrossRef]
68. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) A Handbook of International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis; Chapman Hall/CRC Press: London, UK, 2013. [CrossRef]
69. R Core Team. R: A Language and Environment for Statistical Computing; R Core Team: Vienna, Austria, 2024. Available online: https://www.R-project.org (accessed on 15 June 2024).
70. Robitzsch, A. sirt: Supplementary Item Response Theory Models, 2024. R Package Version 4.2-89. Available online: https://github.com/alexanderrobitzsch/sirt (accessed on 13 November 2024).
71. Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
72. Raykov, T.; Marcoulides, G.A. Introduction to Psychometric Theory; Routledge: London, UK, 2011. [CrossRef]
73. Monseur, C.; Berezner, A. The computation of equating errors in international surveys in education. *J. Appl. Meas.* **2007**, *8*, 323–335. Available online: [https://bit.ly/2WDPeqD](https://bit.ly/2WDPeqD) (accessed on 22 November 2024).

74. Camilli, G. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In *Differential Item Functioning: Theory and Practice*; Holland, P.W., Wainer, H., Eds.; Erlbaum: Hillsdale, NJ, USA, 1993; pp. 397–417.

75. Funder, D.C.; Gardiner, G. Misgivings about measurement invariance. *Eur. J. Pers.* **2024**, *38*, 889–895. [CrossRef]

76. Robitzsch, A.; Lüdtke, O. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Struct. Equ. Model.* **2023**, *30*, 859–870. [CrossRef]

77. Shealy, R.; Stout, W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* **1993**, *58*, 159–194. [CrossRef]

78. Welzel, C.; Inglehart, R.F. Misconceptions of measurement equivalence: Time for a paradigm shift. *Comp. Political Stud.* **2016**, *49*, 1068–1094. [CrossRef]

79. Fischer, R.; Rudnev, M. From Misgivings to Mise-en-scène: The role of invariance in personality science. *Eur. J. Pers.* **2024**, *epub* ahead of print. [CrossRef]

80. Meuleman, B.; Zöltak, T.; Pokropek, A.; Davidov, E.; Muthén, B.; Oberski, D.L.; Billiet, J.; Schmidt, P. Why measurement invariance is important in comparative research. A response to Welzel et al. (2021). *Sociol. Methods Res.* **2023**, *52*, 1401–1419. [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.