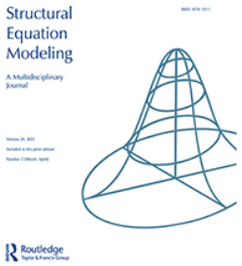
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Finding Clusters of Measurement Invariant Items for Continuous Covariates

Daniel Schulze & Steffi Pohl

To cite this article: Daniel Schulze & Steffi Pohl (2021) Finding Clusters of Measurement Invariant Items for Continuous Covariates, Structural Equation Modeling: A Multidisciplinary Journal, 28:2, 219-228, DOI: 10.1080/10705511.2020.1771186

To link to this article: https://doi.org/10.1080/10705511.2020.1771186

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2021, VOL. 28, NO. 2, 219–228
https://doi.org/10.1080/10705511.2020.1771186

Routledge
Taylor & Francis Group

Finding Clusters of Measurement Invariant Items for Continuous Covariates

Daniel Schulze and Steffi Pohl

Freie Universität Berlin

ABSTRACT

Measurement invariance (MI) cannot always be achieved for the full length of a test or questionnaire, leading researchers to seek for measurement invariant item subsets. Previous approaches either 1) can only analyze group covariates or 2) make implicit assumptions on the nature of MI and then yield a single item set. We provide a general approach for identifying invariant item sets via structural equation modeling that can be applied to dichotomous as well as continuous covariates. It yields multiple item clusters of close measurement invariant items and gives researchers comprehensive insight into the non-MI structure of a test. Furthermore, it allows making well-justified decisions on their assumptions when inferring on group differences or relationships with other variables. The proposed approach performed well in an extensive simulation study. We illustrate the usage and merits of the approach with empirical data on mood states.

KEYWORDS

Partial measurement invariance; cluster analysis; mimic model; moderated non-linear factor analysis; scale indeterminacy

Measurement invariance (MI) describes the stability of item properties given some covariate (e.g., gender, age, or time; Meredith, 1993). In latent variable modeling, MI is a basic prerequisite for sound subsequent comparative analyses (Byrne et al., 1989). More specifically, strong MI is necessary, i.e., the invariance of item loadings and intercepts (Reise et al., 1993). If a global test on strong MI fails, comparisons are still feasible as long as partial MI holds (Byrne et al., 1989; Van de Schoot et al., 2012). Under partial MI, some items are assumed to be measurement invariant, while others are not (also called differential item functioning, DIF; Osterlind & Everson, 2009). Only the invariant items can identify a common scale of a latent variable, while the parameters of non-invariant items may differ for covariates like gender, age, or across time. As long as there is at least one item for which strong MI holds, regressing the latent variable on the covariate provides unbiased results (e.g., in group comparisons; Chen, 2008). Our paper focuses on deriving partial MI models should global MI fail.

The bulk of MI analyses in the literature uses a multigroup framework, either within structural equation modeling (SEM, Vandenberg & Lance, 2000) or item response theory (IRT, Zumbo, 2007). Most analyses use manifest covariates (e.g., gender), while a growing number of studies address latent classes as source of DIF via mixture modeling (e.g., Sawatzky et al., 2018). In this paper, we will build upon previous work and extend it in two ways: (I) We will allow for the direct inclusion of continuous covariates like age, personality factors, or clinical scales. In the past, continuous covariates have usually been arbitrarily categorized in order to fit available DIF analysis methods (e.g., for age groups see McCrae et al., 2002). This approach has clear drawbacks (MacCallum et al., 2002), narrowing the applied researcher's scope of relevant covariates down to truly categorical ones. Here, for including continuous covariates, we will make use of multiple indicators, multiple cause SEM (MIMIC, Hauser & Goldberger, 1971) in a form presented by Bauer (2017). (II) MI analysis suffers from conceptual issues stemming from scale indeterminacy (Pohl & Schulze, 2020a; Steenkamp & Baumgartner, 1998). In order to deal with this issue, previous approaches make explicit or implicit assumptions on the nature of DIF. Assuming strong MI for the wrong items in partial MI models necessarily distorts the results of latent regression. We will illustrate and discuss scale indeterminacy and previous approaches in the next section.

The commonalities of both major psychometric schools, SEM and IRT, are at the core of this paper, as we combine and extend DIF analysis techniques from IRT with the capabilities of the SEM framework. We will derive the item cluster approach for continuous covariates, providing a powerful tool for MI analysis, as it handles continuous (as well as dichotomous) covariates for measurement models of all common item types. Additionally, the item cluster approach makes no fixed assumption on the nature of DIF in order to deal with scale indeterminacy, but instead enables the user to choose from all (reasonable) assumptions.

Scale indeterminacy

Scale indeterminacy impedes straightforward identification of DIF items. Consider the following standard SEM for observed response y_{ip}, of person p on item i and a single latent variable η:

$$y_{ip} = \alpha_i + \lambda_i \cdot \eta_p + \epsilon_{ip}, \quad (1)$$

where α are the item intercepts, λ the item loadings, and ϵ the item residuals. Scale indeterminacy means that the scale of the

CONTACT Daniel Schulze daniel.schulze@fu-berlin.de Department of Of Education and Psychology, Freie Universität Berlin, Berlin 14195, Germany.
This article has been republished with minor changes that do not affect the academic content.
Supplemental data for this article can be accessed on the publisher's website

latent variable is not uniquely defined, but instead must be fixed by some model constraint, e.g., by fixing the intercept of an item to zero and the loading of an item to unity. While identification constraints do not impose any assumptions in most applications of SEM, they imply strict assumptions in an MI analysis. In a two group case, constraints are to be set in both groups. If intercept and slope for an item are fixed to zero and unity, respectively, in each group, it is implied that MI holds true for that item. If that is not the case, the scale of the latent variable is shifted which will heavily bias DIF analysis of the other items. We illustrate the issue of scale indeterminacy by means of DIF of item loadings in an example for a two group case with three items (Figure 1). For loadings λ_{ig} in two groups $g \in (1, 2)$, DIF is commonly defined as

$$\text{DIF}_\lambda = ln(\lambda_{i1}) - ln(\lambda_{i2}) \tag{2}$$

The non-linear transformation with the natural logarithm is used due to the boundedness of loadings at $|1|$ (Rensvold & Cheung, 1998).

If the loading of item 1 is chosen as identifying constraint ("anchor") for the scale in both groups, we would make the assumption that item 1 is measurement invariant (Restriction A in Figure 1). As a consequence, we would conclude that item 3 has substantial DIF, while item 2 shows only marginal DIF. If instead item 3 is chosen as an anchor (Restriction B in Figure 1), we assume item 3 to be measurement invariant. In DIF analysis, items 1 and 2 would appear to have DIF. The arbitrary choice of an item for model identification thus implies strong assumptions in MI analyses. Accordingly, SEM MI literature suggests that anchor items should be chosen carefully on a theoretical basis (Sass, 2011; Steenkamp & Baumgartner, 1998). If a decision cannot be made from theory, one can resort to other approaches that make other assumptions on DIF.

Some approaches assume that the majority of items are DIF-free, e.g., the alignment method (Asparouhov & Muthén, 2014), the iterative forward approach (Kopf et al., 2015), or, in some sense, the factor ratio test[1] (Cheung & Lau, 2012; Rensvold & Cheung, 1998). This assumption might be easier to argue for than arguing for a single DIF-free item. On the downside, assuming that the majority of items are DIF-free is an implicit feature; if a researcher does not find it tenable, the method cannot be used.

Relative DIF and the item cluster approach

To address the issue of scale indeterminacy, we build upon an idea originally proposed for Rasch models in the IRT framework. Based on the work of Lord (1977), Bechger and Maris (2015) presented an approach for identification of MI of item intercepts in which no assumptions on DIF are necessary in the first place. The authors argue that due to scale indeterminacy, absolute DIF cannot be identified in practice. Instead, Bechger and Maris (2015) proposed regarding DIF of an *item i relative to another item j*. For the case of item intercepts α_{ig} in a two group model, relative DIF is defined as:

$$\text{relative DIF}_\alpha = (\alpha_{i1} - \alpha_{j1}) - (\alpha_{i2} - \alpha_{j2}). \tag{3}$$

When looking at the difference of item pair differences, the point of origin of the latent scale is irrelevant. Items with (approximately) equal values in relative DIF function similarly

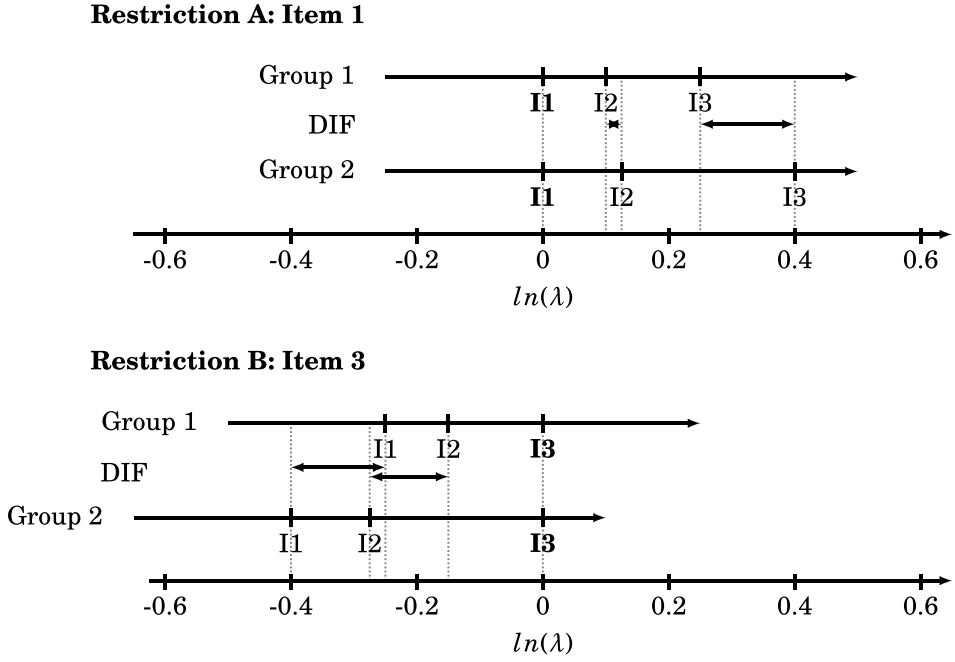


Figure 1. Identifying restrictions on loadings and their impact on the results of DIF analysis in a two group case. In the upper part, the loading of item 1 is constrained to 1 in both groups. In the lower part, the loading of item 3 is constrained to 1 in both groups.

[1]More precisely, this approach returns the largest item subset for which MI holds. This subset does not have to be the majority of items nor has there to be a single largest subset. Still, the factor ratio test yields a result in the spirit of the majority assumption.

to each other, but function differently from items with different relative DIF. Relative DIF of all item pairs can be represented as a $n \times n$ matrix for n items. Bechger and Maris (2015) suggested to visually inspect this matrix in order to find clusters of items functioning similarly to each other – an imprecise and cumbersome task when there are more than a few items. Instead, Pohl and Schulze (2020b) proposed and successfully used a tailored k-means cluster analysis to find subsets of items.

In k-means clustering, the number of clusters needs to be specified, however, this quantity will usually be unknown. Information criteria, like the Bayesian information criterion, proved to be rather unsuccessful for cluster selection in our case (Pohl & Schulze, 2020). Alternatively, the number of clusters could be increased until no cluster has significant residual heterogeneity left (e.g., by using the test of Bechger & Maris, 2015). As a significance test will depend not only on the residual heterogeneity of relative DIFs, but also on sample size, missingness, and model complexity, Pohl and Schulze (2020) instead proposed using a threshold criterion. The threshold represents the maximum within-cluster range of relative DIFs a researcher is willing to accept. The smaller the threshold, the more homogeneous and numerous the clusters are. The resulting cluster solution consists of the smallest number of clusters possible, in which no clusterwise differences in relative DIFs exceed the given threshold. In practice, the threshold's size can be informed by guidelines on DIF size (e.g., OECD, 2015) and expert knowledge.

The item cluster approach typically yields several item clusters. Items within each cluster function similarly to each other but different to items in other clusters. From the analysis only it cannot be decided, which cluster represents the *truly* DIF-free items, or if there is such a cluster at all. For group comparisons, the researcher will have to choose one of the identified item clusters as anchor item(s). This selection will imply an assumption on DIF. Any assumption is possible, e.g., a decision guided by item content; or the largest cluster could be chosen, which mimics the majority assumption of other approaches. In any case, the approach allows for a variety of assumptions and the researcher's assumption on DIF will have to be stated clearly and will thus be open for discussion. With the item cluster approach, the researcher does not have to choose a single item from n items but instead chooses from $k \ll n$ item sets – an easier task, from both a theoretical and a practical perspective.

So far, the cluster approach has been developed to account for DIF in intercepts (Pohl & Schulze, 2020) and slopes (Pohl & Schulze, 2020) in the case of dichotomous covariates. However, continuous covariates such as age are equally important in MI analysis and occur frequently. The next section describes current methods handling the generalization of MI analysis for continuous covariates.

Continuous covariates and MIMIC models

Various approaches in the SEM literature allow to include continuous covariates in MI analysis, like local SEM (Hildebrandt et al., 2016), penalized likelihood estimators (Robitzsch & Lüdtke, 2018), or MIMIC (B. O. Muthén, 1989). Here, we will focus on MIMIC models which assess the impact of the continuous covariate on the item parameters straightforwardly and have been thoroughly tested for the use in MI analysis. In MIMIC, the latent variable is both measured by observed items *and* is dependent on observed covariates. Muthén's (1989) approach incorporated a regression of the covariate on the item difficulties and the latent variable's mean, but not on loadings or variances. To this end, Muthén (1989) extended Equation 1 by incorporating a regression of the latent variable as well as the item intercept on a continuous covariate x:

$$\eta_p = \kappa + \beta \cdot x_p + \zeta_p \tag{4}$$

$$\alpha_{ip} = \alpha_{0i} + \alpha_{1i} \cdot x_p, \tag{5}$$

where κ is the latent intercept when $x_p = 0$, β is the change in η with changes in the covariates, ζ is the residual in the latent variable η, α_{0i} is the intercept of item i when $x_p = 0$, and α_{1i} is the change of item i's depending on x.

Muthén's (1989) approach is limited to testing invariance of item intercepts and has been rarely used (an exception is Marsh et al., 2013). Missing the capability to test strong MI, some researchers have instead arbitrarily dichotomized continuous variables and analyzed them with the common multigroup framework (e.g., Koomen et al., 2012; McCrae et al., 2002). As a remedy, Bauer (2017) introduced an extension of the MIMIC model that includes a regression of the covariate on the loadings λ:

$$\lambda_{ip} = \lambda_{0i} + \lambda_{1i} \cdot x_p, \tag{6}$$

where λ_{0i} is the loading of item i when $x_p = 0$, and λ_{1i} is the change of item i's when the covariate increases for one unit. Inserting Equation 5 and 6 in Equation 1 leads to

$$y_{ip} = \alpha_{0i} + \alpha_{1i} \cdot x_p + (\lambda_{0i} + \lambda_{1i} \cdot x_p) \cdot \eta_p + \epsilon_{ip}. \tag{7}$$

Additionally, Bauer (2017) allows for heteroscedasticity in η in the following nonlinear form:

$$var(\eta_p) = \phi \cdot e^{\gamma \cdot x_p}, \tag{8}$$

where ϕ is a baseline variance when $x_p = 0$ and γ describes the impact of the covariate on the variance of η.

With Bauer's (2017) model, item-wise analysis of strong MI with continuous covariates becomes feasible. In order to identify single items exhibiting DIF, Bauer (2017) proposes an iterative *all-other* strategy (for terminology, see Wang & Yeh, 2003): A baseline model without any DIF is estimated (i.e., setting $\alpha_{1i} = \lambda_{1i} = 0$). Then, n models with a single item i displaying DIF in intercept and loading are run sequentially, where $i = 1 \ldots n$. It is thus assumed that *all other* items are DIF free. Likelihood ratio tests are used to compare the increase in fit of each DIF model to the baseline model. The model with the strongest improvement defines the first DIF item. This item is excluded from any anchor set in the next step. The procedure is repeated until no further statistically significant increase in fit can be achieved. The items that remain in the anchor are interpreted as DIF-free items.

This strategy has two drawbacks: First, significance testing depends on sample size, generally resulting in some undesirable effects. E.g., items with larger standard errors of item

parameters, e.g., due to many missing values, will be more likely to be labeled "DIF-free."

Second, the issue of scale indeterminacy appears not only in multigroup SEM, but also in MIMIC-type MI analysis. In the proposed iterative approach of Bauer (2017), all other items but the one under investigation are assumed to be DIF-free in the first step. If we find significant differences to the baseline model, this could mean two things: A) the item under scrutiny has DIF or B) some of the other items *but* the item under scrutiny have DIF. Due to scale indeterminacy, we cannot decide between those two options by statistical means.

Research question

Current methods for MI analysis are restricted in one of two ways: They either comprise group comparisons only, or, when extended to continuous covariates, they rely on implicit assumptions on DIF that are not necessarily met in practice. Our main goal is to develop a general approach for the identification of anchor items that allows for continuous covariates and presents all possible options of anchor item sets. We bring together multiple strands of current developments in MI analysis. We build upon the approach proposed by Bechger and Maris (2015) and Pohl and Schulze (2020) and extend it by drawing on the MIMIC model of Bauer (2017). In the following, we will first define relative DIF in the presence of continuous covariates. We will then evaluate our approach in a simulation study and illustrate its application in an empirical example.

Item cluster approach for continuous covariates

The original definition of relative DIF in Equation 3 was set up for the comparison of two groups, but not for continuous variables. Still, we would like to argue that the comparison of two values along the scale of the covariate can be used to describe differences in item parameters. Of course, this assumes that the effect of the covariate on the item parameters is linear. Changing these points merely changes the relative DIF by some factor f. If the same two values of the covariate are used for all pairwise calculations of relative DIF (see Equation 3), the same factor f applies to all resulting relative DIF values, making it irrelevant to the clustering algorithm in the next step. We thus propose to always standardize the covariate and use the values $x_a = 0$ and $x_b = 1$ to calculate relative DIF. This a) simplifies the equations below and b) provides a standardized scale to the relative DIF values, which eases interpretation.

With respect to Equation 3 and 7, we define relative DIF of the loadings λ and intercepts α between two items i and j as:

$$\text{relative DIF}_\alpha = (\alpha_{ix_a} - \alpha_{jx_a}) - (\alpha_{ix_b} - \alpha_{jx_b}) \tag{9}$$

$$\begin{aligned}\text{relative DIF}_\lambda = &[ln(\lambda_{ix_a}) - ln(\lambda_{jx_a})] \\ &- [ln(\lambda_{ix_b}) - ln(\lambda_{jx_b})].\end{aligned} \tag{10}$$

Inserting Equation 5 in Equation 9, this leads to:

$$\begin{aligned}\text{relative DIF}_\alpha = &[(\alpha_{0i} + \alpha_{1i} \cdot x_a) - (\alpha_{0j} + \alpha_{1j} \cdot x_a)] \\ &- [(\alpha_{0i} + \alpha_{1i} \cdot x_b) - (\alpha_{0j} + \alpha_{1j} \cdot x_b)]\end{aligned} \tag{11}$$

$$= \alpha_{1i} \cdot x_a - \alpha_{1j} \cdot x_a - \alpha_{1i} \cdot x_b + \alpha_{1j} \cdot x_b \tag{12}$$

$$= \alpha_{1i}(x_a - x_b) - \alpha_{1j}(x_a - x_b). \tag{13}$$

Using Equation 6, relative DIF for the item loading λ can be depicted as:

$$\begin{aligned}\text{relative DIF}_\lambda = &[ln(\lambda_{0i} + \lambda_{1i} \cdot x_a) - ln(\lambda_{0j} + \lambda_{1j} \cdot x_a)] \\ &- [ln(\lambda_{0i} + \lambda_{1i} \cdot x_b) \\ &- ln(\lambda_{0j} + \lambda_{1j} \cdot x_b)].\end{aligned} \tag{14}$$

When choosing the points $x_a = 0$ and $x_b = 1$ of a standardized continuous covariate, Equations 13 and 14 simplify to

$$\text{relative DIF}_\alpha = \alpha_{1j} - \alpha_{1i}, \tag{15}$$

$$\begin{aligned}\text{relative DIF}_\lambda = &[ln(\lambda_{0i}) - ln(\lambda_{0j})] \\ &- [ln(\lambda_{0i} + \lambda_{1i}) - ln(\lambda_{0j} + \lambda_{1j})].\end{aligned} \tag{16}$$

In order to estimate relative DIF, we build upon the MIMIC model proposed by Bauer (2017) as described by Equation 7. For model identification, we set $\kappa = 0$ and $\beta = 0$ in Equation 4 and $\phi = 1$ and $\gamma = 0$ in Equation 8. This means that neither latent mean nor latent variance differs between two covariate values, which is a common model constraint. Note that this restriction has no impact on relative DIF$_\lambda$ as this is defined to be scale invariant (Bechger & Maris, 2015; Pohl & Schulze, 2020).

In a second step, we use the relative DIF values of both parameters to identify groups of homogeneously functioning items. To this end, we propose to use a k-means algorithm, though any other cluster algorithm could be used as well.[2] The clustering algorithm is fed with the first column of each relative DIF matrix instead of the whole matrix. As relative DIF matrices are skew-symmetric with a rank of two and the absolute position of the scale is arbitrary, a single column from both matrices contains all necessary information on relative DIF. Thus, the simple task of clustering a two-dimensional space remains (relative DIF of λ and α). In accordance with Pohl and Schulze (2020), the number of clusters can be determined in two different ways. 1) The number of clusters is directly specified. 2) The number of clusters is found by means of thresholds for the relative DIFs of both item parameters. To that end, we run a k-means routine with increasing cluster count, until relative DIF$_\lambda$ and DIF$_\alpha$ within each resulting cluster is below the respective threshold.

As described for the item cluster approach above a researcher then may choose between the resulting item clusters. The items of the chosen cluster can then be used as anchor items to identify the latent scale of a partial MI model. For this purpose, a model is specified in which the covariate has an impact on all item parameters but those of the chosen anchor items. The regression of the covariate on the latent variable η then reflects the unbiased effect, as long as an

[2]We additionally pretested hierarchical clustering as well as DBSCAN (Ester et al., 1996). Both performed equally well as k-means in order to recover DIF structure.

appropriate cluster has been chosen. The proposed approach is depicted in Figure 2.

Evaluation of the approach in a simulation study

For evaluation, we considered a measurement model containing a single latent variable measured by 15 continuous response variables and a single continuous covariate introducing DIF. The simulation included conditions varying the percentage and distribution of DIF items as well as DIF size in item loadings and difficulties. There was a baseline condition with no violation of MI for comparison. An overview over simulation conditions can be found in Table 1.

In all subsequent conditions but the DIF-free condition, the 15 items were distributed over four-item clusters with varying DIF properties. The first cluster represented DIF-free items, the second displayed DIF_λ, the third DIF_α, and the fourth DIF in both parameters. Our simulation setup reflects the notion that DIF in item loadings (DIF_λ) and item difficulties (DIF_α) can occur in the same item and that anchor items need to function similarly on *both* parameters to assure strong MI. We implemented three different DIF item proportions (20%, 60%, and 80%), see Table 1. The first condition depicts cases with the majority of items being DIF free, as assumed by some other approaches (see Introduction). The latter two are supposed to examine the approach's performance under harsh conditions. With 60%, the DIF-free item set was still the one with the highest cardinality (6 items), which was not the case with 80% DIF items. We scrutinized minor, medium, and large DIF sizes and chose the sizes according to common practice in multigroup MI testing (González-Betanzos & Abad, 2012; Wang & Yeh, 2003), i.e., $\text{DIF}_\lambda \in (0.05, 0.1, 0.2)$ for loadings and $\text{DIF}_\alpha \in (0.1, 0.3, 0.6)$ for difficulties.[3]

To summarize, the simulation design had three fully crossed factors: DIF_λ size (with three values), DIF_α size (with three values), and percentage of DIF items (with three values). In addition to the baseline no-DIF condition, this resulted in 28 conditions overall. For each condition, we generated $r = 100$ data sets.

Data generation

We generated data for $N = 500$ persons from the model described in Equation 4, 7, and 8 and incorporated the conditions from above. We simulated normally distributed data for the covariate x with $N(0, 1)$. The latent variable η was then generated following Equation 4 with an intercept of $\kappa = 1$, a slope of $\beta = 2$, and a determination coefficient of $P^2 = 0.25$. According to benchmarks on heteroscedasticity by Steiner et al. (2010), we induced heteroscedasticity with a variance ratio of 1.5, i.e., the latent variable's variance at the 97.5 percentile of the covariate was 1.5 times greater than at the 2.5 percentile.

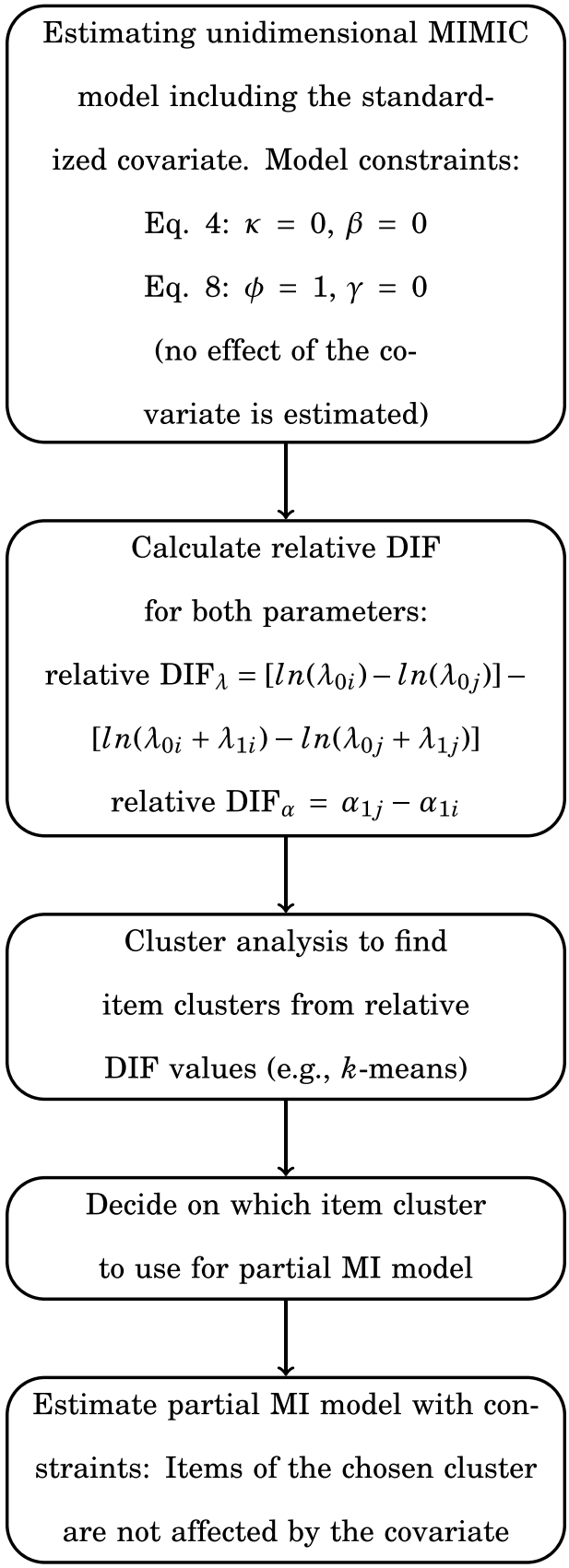


Figure 2. Item cluster approach for DIF analysis.

[3]To challenge the approach, all conditions portrayed *unbalanced* DIF. When DIF is balanced, the amount of positive DIF is equal to amount of negative DIF across all items. Thus, on average across the values of the covariate, there is no difference in loadings or difficulties. Such conditions have proven to be favorable for finding anchor item sets in the IRT literature (Kopf et al., 2015; Wang, 2004). On contrary, we used only positive and thus unbalanced DIF, thus challenging the approach.

Table 1. Setup of the simulation study displaying the conditions for DIF item proportions.

		20% DIF item		60% DIF item		80% DIF item	
Item	no DIF	λ	α	λ	α	λ	α
1							
2							
3							
4						DIF_λ	
5						DIF_λ	
6						DIF_λ	
7				DIF_λ		DIF_λ	
8				DIF_λ			DIF_α
9				DIF_λ			DIF_α
10					DIF_α		DIF_α
11					DIF_α		DIF_α
12					DIF_α	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$
13		DIF_λ		$2 \times DIF_\lambda$	$2 \times DIF_\alpha$	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$
14			DIF_α	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$
15		$2 \times DIF_\lambda$	$2 \times DIF_\alpha$	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$	$2 \times DIF_\lambda$	$2 \times DIF_\alpha$

A single latent variable was measured by 15 items, distributed over four DIF item clusters. In order to generate four completely distinct item clusters regarding DIF, DIF size on both parameters was twice as high in the fourth item cluster. Empty cells = no DIF. DIF_λ took one of the values 0.05, 0.1, or 0.2. DIF_α took one of the values 0.1, 0.3, or 0.6.

We randomly drew the initial item loadings from a truncated normal with $\lambda_i \sim TN(1, 0.25, \text{lower} = 0.7, \text{upper} = 1.5)$ for every repetition. These values were chosen in order to achieve reasonable item discriminations with regard to empirical data (e.g., see Empirical Example). The item difficulties α were randomly drawn from $N(0, 1)$. DIF_λ and DIF_α values were added to the items according to the respective condition. Finally, item responses are generated from Equation 7.

Analyses

We analyzed the data using the proposed item cluster approach under various combinations of threshold settings for DIF_λ and DIF_α. Threshold settings were chosen to cover the whole range of DIF. For DIF_λ-thresholds, we used values ranging from 0.05 to 1.0 by steps of 0.05 and values ranging from 0.02 to 0.4 by steps of 0.02 for the DIF_α-thresholds. DIF_λ- and DIF_α-thresholds were fully crossed in the simulation.

For evaluating and aggregating the results across repetitions, we chose the cluster with the highest hit rate, i.e., the percentage of truly DIF-free items on the total number of items in the cluster. This cluster was labeled *best cluster* and was used for further analyses. If there were two equally good clusters, one of the clusters was chosen at random. We then evaluated cluster length (i.e., number of items), hit rate, as well as bias in β, i.e., the regression parameter. We applied ANOVAs to depict the effects of the simulation conditions on these three criteria. This analysis model contained all main and interaction effects and effect size was expressed in terms of explained variance portions η^2.

We used R (R Development Core Team, 2018) for data generation and clustering, M*Plus*, version 7 (Muthén & Muthén, 1998–2012) for estimating the models and the R package MPlusAutomation (Hallquist & Wiley, 2018) as interface between R and M*Plus*.

Results

Table 2 gives η^2 for the main effects of the simulated conditions and outcome means for condition levels. Depictions of the detailed results can be found in the supplementary material.

Hit rate essentially depended on the share of DIF items in the test as well as the chosen thresholds. It naturally decreased considerably with higher percentages of DIF items, from an average hit rate of 0.93 for 20% DIF items to an average of 0.51 for 80% of DIF items. Furthermore, hit rate displayed a steep decrease when both thresholds in combination exceeded the true DIF size, but a lesser decrease when only one of the thresholds exceeded its respective DIF size (see figures in the Supplements).

A concurrent pattern could be observed for cluster length: It surpassed the true cluster length with increasing proportion of DIF items (see Table 2). Furthermore, the average cluster length changed proportionally to the thresholds. Cluster length was smaller than the truly DIF-free cluster when parameter thresholds were smaller than true DIF and vice versa. Inevitably, too long clusters included DIF items, which decreased the hit rate. In the no-DIF condition, cluster length depended only on the DIF_α-threshold (see top plots in the Supplements).

Most interesting for applied research is the bias in the latent regression parameter β. In general, only small or no average bias occurred in most conditions. The pattern of bias resembled the patterns found for hit rate and cluster length only in part. In the no-DIF condition, no bias was found,

Table 2. Main effects of the simulated conditions on outcome variables.

		Hit Rate	Cluster Length	Bias in Slope
DIF item proportion				
	η^2	*0.52*	*0.13*	*0.00*
	0.20	0.93	9.83	0.00
Level Means	**0.60**	0.71	8.44	0.00
	0.80	0.51	7.40	−0.01
α DIF size				
	η^2	*0.05*	*0.04*	*0.00*
	0.00	1.00	8.25	0.01
	0.10	0.64	9.67	0.00
Level Means	**0.30**	0.74	8.18	0.00
	0.60	0.75	8.00	−0.01
λ DIF size				
	η^2	*0.02*	*0.03*	*0.01*
	0.00	1.00	8.25	0.01
	0.05	0.67	7.85	0.01
Level Means	**0.10**	0.70	9.10	0.01
	0.20	0.75	8.89	−0.02
α threshold†				
	η^2	*0.17*	*0.36*	*0.01*
	0.05	0.95	3.37	−0.02
	0.25	0.76	7.34	−0.02
Level Means	**0.50**	0.69	9.10	−0.01
	0.75	0.66	10.19	0.01
	1.00	0.64	10.84	0.02
λ threshold†				
	η^2	*0.17*	*0.27*	*0.00*
	0.02	0.90	4.63	−0.02
	0.10	0.82	6.67	−0.01
Level Means	**0.20**	0.70	9.15	0.00
	0.30	0.65	9.97	0.00
	0.40	0.65	9.97	0.00

Italics = eta squared (explained variance proportion). See supplements for higher order effects. †Not all single steps are printed. α threshold increased by increments of 0.05, λ threshold by 0.02.

irrespective of the thresholds chosen. The same was true, when only 20% of all items presented DIF. Average bias increased up to |0.10| (60%-DIF-items) and |0.21| (80%-DIF-items) with increasing DIF on both parameters and when thresholds *were larger* than the true DIF size. This effect was more pronounced for the DIF_α-threshold than the DIF_λ-threshold (see figures in the Supplements).

In summary, the cluster approach was able to find a set of anchor items producing little to no bias in most conditions. The reproduction of the generated item clusters depended on the chosen thresholds, with thresholds close to or smaller than true DIF yielding unbiased results.

Illustration of the approach in an empirical example

Here, we illustrate our approach with data from the German multidimensional mood state (MDBF) questionnaire. The MDBF consists of 58 items assessing mood states by rating single adjectives and has a well-known factor structure (Steyer et al., 2004). In the following, we use the Unrest subscale which comprises 10 items with a Likert scale ranging from 1 (*not at all*) to 5 (*very much*). Although modeling Likert scales as ordinal variables is formally advantageous, we will treat this scale as continuous like the original authors did (Steyer et al., 2004) and as this provides a typical applied case in the analysis of questionnaire data. The item wording can be found in Table 3. For illustrating our approach we will investigate the relationship of Unrest with the covariate life satisfaction.

The sample of $N = 485$ participants (280 female, $M_{age} = 31.2$, $Min_{age} = 17$, $Max_{age} = 73$) was drawn in the course of test validation (Steyer et al., 2004). Life satisfaction was assessed by the Freiburger personality inventory (Fahrenberg & Selg, 1970). Its 12 items are rated as *true* or *not true* by the participants. According to Fahrenberg and Selg (1970), we used the standardized sum of affirmative answers as scale scores with higher values indicating higher life satisfaction. An initial global MI test via model comparison yielded a substantial violation of strong MI ($\chi(18) = 73.33$, $p < .001$, BIC-baseline = 14848, BIC-restricted = 14811). We thus applied the item cluster approach as proposed above. When using moderate thresholds ($DIF_\lambda = 0.2$ and $DIF_\alpha = 0.3$, González-Betanzos & Abad, 2012), four clusters emerged with two clusters consisting of only one item (see Table 3). The four clusters seem to depict subtle differences in item content: Compared to the large block of cluster B items, cluster C in contrast consists of a rather positive mood state (item 7 – exhilarated), and cluster D of a clearly negative one (item 10 – scared).

In order to investigate the relationship of Unrest and life satisfaction by means of partial MI models, one of the item clusters has to be chosen as anchor to identify the model. Researchers might either choose a cluster A) based on expert knowledge on the item contents (preferred from an applied research perspective), B) in the absence of expert knowledge make some assumption on DIF (e.g., choose the largest cluster, which is similar to the majority assumption of other approaches), or, C), in the absence of expert knowledge report a sensitivity analysis for which all clusters are used as anchors sequentially and all mean differences are reported, depicting the uncertainty in the results due to anchor item choice.

For illustrative purposes, we applied the last option. No matter which cluster was chosen as anchor, we found negative effects showing an decrease in Unrest with increasing life satisfaction (see second last column in Table 3). However, the effect size varied from negligible (cluster C) to strong and significant (cluster D). Heteroscedasticity varied as well, although significant heteroscedasticity could only be found using cluster B as anchor. The variance in Unrest decreased with a factor of 0.83 per additional life satisfaction unit (see last column in Table 3).

In comparison, we applied an all-other approach using likelihood ratio tests as described in Bauer (2017). Items 1, 5, 7, and 10 were detected having DIF, i.e., the items of cluster B were detected as non-DIF items (see Table 3) with a slope of −0.34 and a variance ratio of 0.83, both significant. This illustrates that previous MI approaches in MIMIC models approach identify only one possible anchor item set, in contrast to the item cluster approach.

Table 3. Items of the MDBF Unrest scale and results from DIF analysis with life satisfaction.

#	Item	relative DIF_λ	relative DIF_α	Cluster	β	e^ν
1	restless	0.00	0.00	A	−0.33	1.19
5	hyped up	0.06	0.17		$p < .001$	$p = .069$
2	nervous	−0.15	0.10	B	−0.34	0.83
3	irritated	−0.26	0.03		$p < .001$	$p = .013$
4	testy	−0.20	0.02			
6	irascible	−0.18	0.02			
8	excited	−0.29	0.08			
9	unsettled	−0.14	−0.06			
7	exhilarated	−0.25	0.31	C	−0.05	0.87
					$p = .710$	$p = .459$
10	scared	−0.76	0.01	D	−0.86	0.88
					$p < .001$	$p = .277$

DIF analysis results for four clusters found by the item cluster approach. Items of cluster B correspond to the result found by the all-other approach by Bauer (2017). MDBF = German multidimensional mood state questionnaire. Relative DIF_λ = loading DIF according to Eq. 10. Relative DIF_α = intercept DIF according to Eq. 9. β = slope from regressing Unrest on life satisfaction using the respective item cluster as constraint. e^ν = heteroscedasticity in terms of a variance ratio in the regression of Unrest on life satisfaction using the respective item cluster as constraint.

Discussion

When total MI of an instrument does not hold, partial MI poses a remedy. A critical issue in partial MI analysis is choosing items as anchor in order to establish a common scale. We proposed the item cluster approach by extending and combining modern advances in MI analysis from both major traditions, SEM and IRT. Our proposed approach can handle continuous as well as dichotomous covariates and allows for a variety of assumptions on DIF. In contrast to other approaches, the user can incorporate expert knowledge and theoretical considerations into the analysis, i.e., by choosing the size of thresholds and choosing the item cluster for a final partial MI model. Furthermore, our approach supports explicating assumptions on the nature of DIF. This strengthens the position of the researcher as an expert of her field by allowing and demanding explicit decisions at the same time.

E.g., content validity of the chosen item cluster has to be established and argued for, making decisions transparent.

The item cluster approach typically yields several item clusters as candidates for anchoring. The subsequent decision on which cluster to choose incorporates the assumption in DIF a researcher is willing to make. Alternatively, all item clusters could be used consecutively to estimate the desired effect and portray its uncertainty due to anchor item choice (see empirical example). Of course, all results should be reported in this case. Reporting only single results after calculating all might lead to hypothesizing after the results are known (HARKing).

In contrast, we illustrated the properties of an alternate approach to MI analysis in the empirical example. The *all-other* approach as described in Bauer (2017) is common in MI analysis and assumes that all of those items are DIF-free which are currently not under investigation (Wang & Yeh, 2003). Like other previous approaches, it is designed to yield a single anchor item set. In our example, this item set coincided with cluster B identified by the item cluster approach. Note that the result of an all-other approach will not always coincide with one of the clusters found by the item cluster approach. The first relies on significance testing, which depends on sample size and missing data proportions. The latter depends on the thresholds chosen.

By means of a simulation study, we were able to show the capabilities of the item cluster approach. Under a wide range of conditions, the algorithm was almost always able to retrieve an item cluster that yielded no or small bias in the latent regression parameter. This result is most important when applying the cluster approach with the goal to estimate the relation of the MI covariate with the latent variable. The chosen thresholds had little impact on bias, unless the percentage of DIF items was high ($\geq 60\%$) and DIF was strong.

In the simulation study, we only considered continuous covariates due to brevity, not dichotomous ones. The latter has been widely used in the previous literature (e.g., Van de Schoot et al., 2012) and the findings of the first can easily be transferred to dummy-coded dichotomous covariates. In the simulations results, we depicted the results of the best cluster of each iteration in order to be able to aggregate results. Thus, the results presented here are the best possible outcome of our approach. If other clusters are chosen as an anchor in applied analyses, bias in the latent regression would result. In other words, the simulation did not evaluate the effect of the final decision on which item cluster to choose but whether the approach is able to find an appropriate cluster. Again, we would like to emphasize the importance of bringing in expert knowledge on the items at this point. The results of the empirical example illustrate this point. Our analysis revealed four-item clusters among 10 mood state items. In applications, researchers have to choose from the item clusters. They might argue that a certain cluster has higher content validity than others and choose this cluster as an anchor. Or, like other approaches, they could argue for using the largest cluster, in the sense that most items could be expected to not portray DIF after careful test construction. In case no single tenable assumption can be made, a researcher may also depict the results from all anchor item sets separately. We followed this strategy in the empirical example and were thus able to portray the uncertainty in the regression results (e.g., in β).

The choice on DIF thresholds is crucial to our approach, as cluster count directly depends on it. The thresholds can be seen to represent the maximum amount of DIF a researcher is inclined to accept as negligible within an item cluster. Concluding from the simulation study, we suggest to set both thresholds to values just below the expected DIF size. Using thresholds that are larger than the true (and unknown) DIF size will decrease hit rate and induce bias in regression parameter estimates. Thresholds that are smaller than the true DIF size will lead to less bias, but slightly larger uncertainty in regression parameters.

Limitations and outlook

In this paper, we extended previous approaches by combining the identification of all possible anchor item sets with analyzing continuous covariates. The item cluster approach is thus capable of dealing with many common situations in applied data analysis. Some few open questions remain. First, the use of multiple covariates at once is a strength of MIMIC-type MI analysis. We did not consider multiple covariates in this paper. Future research should extend the approach and could cluster a higher-dimensional space to this end. The same technique could also be used to adopt the item cluster approach to polytomous covariates in the form of multiple dummy variables (e.g., school type or migration background). Further suggestions on how to deal with polytomous variables can be found in Huelmann et al. (2019), who evaluate various methods of aggregating DIF across a multitude of groups in the iterative forward approach.

Second, we assumed that the relation of the covariate with the item parameters is linear. This assumption could be abandoned by adding higher order terms of the covariate to the regression equations. From this perspective, generalizing the linearity assumption on a technical level seems related to the inclusion of multiple covariates.

Third, the estimates of DIF_α depend upon the item loadings λ (Pohl & Schulze, 2020). Thus, our approach does not fully solve the scale indeterminacy issue for item intercepts. Due to the dependency of DIF_α on λ during estimation, we think that this conceptual issue cannot be solved statistically.

The item cluster approach provides a thorough tool for MI analysis. It can be applied to dichotomous or continuous covariates and to dichotomous, ordinal and continuous response variables by means of already established SEM techniques. When used in test construction, it can inform the researcher on the structure of measurement non-invariance. When used in applied research, it leaves the researcher with more control on choosing an adequate item set to establish a partial MI model. With this work, we hope to raise awareness for the assumptions one has to make in MI analysis. The item cluster approach explicates these and allows users to incorporate various assumptions, making them open to discussion.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO1655/2-1).

ORCID

Daniel Schulze http://orcid.org/0000-0001-9415-2555
Steffi Pohl http://orcid.org/0000-0002-5178-8171

References

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

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005–1018. https://doi.org/10.1037/a0013193

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*, 167–198. https://doi.org/10.1177/1094428111421987

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. B. E. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)* (pp. 226–231), Portland, Oregon.

Fahrenberg, J., & Selg, H. (1970). *Das Freiburger Persönlichkeitsinventar FPI*. Verlag für Psychologie, Hogrefe.

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology*, *8*, 134–145. https://doi.org/10.1027/1614-2241/a000046

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in M plus. *Structural Equation Modeling*, *25*, 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *3*, 81–117. https://doi.org/10.2307/270819

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research*, *51*, 257–278. https://doi.org/10.1080/00273171.2016.1142856

Huelmann, T., Debelak, R., & Strobl, C. (2019). A comparison of aggregation rules for selecting anchor items in multi group dif analysis. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12246

Koomen, H. M., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the student-teacher relationship scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology*, *50*, 215–234. https://doi.org/10.1016/j.jsp.2011.09.001

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*, 22–56. https://doi.org/10.1177/0013164414529792

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Swets & Zeitlinger Publishers.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*, 1194. https://doi.org/10.1037/a0026913

McCrae, R. R., Costa, P. T., Jr, Terracciano, A., Parker, W. D., Mills, C. J., De Fruyt, F., & Mervielde, I. (2002). Personality trait development from age 12 to age 18: Longitudinal, cross-sectional and cross-cultural analyses. *Journal of Personality and Social Psychology*, *83*, 1456. https://doi.org/10.1037//0022-3514.83.6.1456

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585. https://doi.org/10.1007/BF02296397

Muthén, L., & Muthén, B. (1998–2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

OECD. (2015). *PISA 2015 technical report (Tech. Rep.)*. Organisation for Economic Co-operation and Development. http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications.

Pohl, S., & Schulze, D. (2020a). Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. *Psychological Test Assessment and Modelling*, *62* (2), 281–303. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-2/04_Pohl.pdf

Pohl, S., & Schulze, D. (2020b). *Partial measurement invariance: Extending and evaluating the cluster approach for identifying anchor items*. Manuscript submitted for publication.

R Development Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*, 1017–1034. https://doi.org/10.1177/0013164498058006010

Robitzsch, A., & Lüdtke, O. (2018). *A regularized moderated item response model for assessing differential item functioning. Talk given at the VIII*. European Congress of Methodology.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*, 347–363. https://doi.org/10.1177/0734282911406661

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*, *27*, 1745–1755. https://doi.org/10.1007/s11136-017-1680-8

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

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267. https://doi.org/10.1037/a0018719

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (2004). *Development of the multidimensional mood state questionnaire (MDBF). Primary*

data. (Version 1.0.0) [Data and documentation]. Center for Research Data in Psychology: PsychData of the Leibniz Institute for Psychology Information ZPID. https://doi.org/10.5160/psychdata.srrf91en15

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://doi.org/10.1177/109442810031002

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*, 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479–498. https://doi.org/10.1177/0146621603259902

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233. https://doi.org/10.1080/15434300701375832