Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 281-303

Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF

Steffi Pohl[1] *& Daniel Schulze*[2]

Abstract

Comparing scale scores across groups or time requires the assumption of measurement invariance. In order to still facilitate comparisons when measurement invariance does not hold, researchers can strive for partial measurement invariance by identifying respective anchor items. Recently, the cluster approach (Bechger & Maris, 2015; Pohl & Schulze, 2020) has been suggested for the identification of such anchor items regarding intercept parameters within the 1PL model only. We extend the cluster approach to intercepts and slopes in the 2PL model. The cluster approach acknowledges the scale indeterminacy problem and – in contrast to previous approaches – identifies multiple possible item clusters that may function as anchor items. This allows researchers to substantially consider various solutions as well as to depict the uncertainty of the results due to anchor item selection. Here, we evaluate the performance of the approach in a simulation study and illustrate its application in an empirical example.

Keywords: Measurement invariance, differential item functioning, cluster analysis, nonuniform DIF, scale indeterminacy

[1] *Correspondence concerning this article should be addressed to:* Steffi Pohl, PhD, Freie Universität Berlin, Arbeitsbereich Methoden und Evaluation/Qualitätssicherung, Habelschwerdter Allee 45, 14195 Berlin, Germany; email: steffi.pohl@fu-berlin.de

[2] Freie Universität Berlin

Introduction

In psychological and educational studies, researchers are often interested in comparisons of constructs across groups or over time. If constructs are measured via an item set, one prerequisite for a meaningful comparison is measurement invariance (MI, Borsboom, 2006; Kolen & Brennan, 2004; Meredith, 1964, 1993; Millsap, 2010). Investigating MI has become a standard in psychometrics when constructing and evaluating a new scale or for comparison of constructs over groups or time (AERA, APA, & NCME, 2014). MI does not always hold in empirical applications, i.e., differential item functioning (DIF) is present (e.g., Kreiner & Christensen, 2014; Zumbo, 2007). In this case, assuming partial MI is an option to still facilitate comparisons across groups or time by identifying a subset of items called anchor items, for which the assumption of MI holds. There is plenty of literature on identifying anchor items (for an overview see, e.g., Hidalgo & Gómez-Benito, 2010; Magis, Béland, Tuerlinckx, & De Boeck, 2010). Previously proposed approaches yield a single solution and implicitly or explicitly make strong assumptions on the nature of DIF that heavily impact which items are included in the anchor (Pohl & Schulze, 2020).

Based on Lord's ideas (1977, 1980), the cluster approach to anchor items has been suggested (Bechger & Maris, 2015; Pohl & Schulze, 2020). Here, no assumptions are made in the first place, instead multiple possible solutions for anchor item sets (i.e., clusters) are revealed. As such, the cluster approach facilitates an explicit decision regarding assumptions and allows to evaluate and depict the uncertainty stemming from the choice of the anchor item cluster (Pohl & Schulze, 2020). In previous research, this approach provided unbiased estimates of group differences (Pohl & Schulze, 2020). However, so far this approach has only been proposed for uniform DIF, regarding only item intercept parameters. In this paper, we extend the approach to item slopes and as such account for nonuniform DIF.

Current approaches for identification of anchor items

In order to regard nonuniform DIF, we consider the 2PL model (Birnbaum, 1968), however we here refer to the slope-intercept parametrization:

$$P(X_{pig}|\theta, \beta, \alpha, g) = \frac{e^{\alpha_{ig}\theta_{pg}-\beta_{ig}}}{1+e^{\alpha_{ig}\theta_{pg}-\beta_{ig}}} \tag{1}$$

with X_{pig} denoting the item response of person p on item i in group g, θ_{pg} the latent ability, which follows a normal distribution in each group g, that is, $\theta_g \sim N(\mu_g, \sigma_g)$; β_{ig} the intercept, and α_{ig} the slope of the respective item i in group g. Here we will limit the number of compared groups to $g = 2$ (see Discussion for extensions to more than two groups).

When a global invariance test indicates the presence of measurement variance, there are various approaches for identifying anchor items. Three strategies commonly used in practice are the single-anchor, the all-other, and the equal-mean-difficulty approach.

The strategies for identifying anchor items are implemented through specific model identification constraints.

In the *equal-mean-difficulty* approach, the scale of the latent variable in each group is identified by fixing the mean of item difficulties (or item intercepts) across all items to zero and the product of item discriminations (or item slopes) to unity, that is

$$\sum_{j=1}^{n} \beta_{j1} = \sum_{j=1}^{n} \beta_{j2} = 0, \quad \prod_{j=1}^{n} \alpha_{j1} = \prod_{j=1}^{n} \alpha_{j2} = 1. \tag{2}$$

DIF is evaluated for each item by the differences in estimated item parameters across groups. By setting the average (or product) of the item parameters to be equal across groups, the equal-mean-difficulty approach poses the assumption that DIF is balanced. This implies that the test does not on average favor one group over the other and that the item parameters' DIF cancels out on test level. This is a rather strong assumption. If it is violated, the approach yields low bias in mean group differences only when DIF is small or with a large proportion of DIF-free items (e.g., Pohl & Schulze, 2020; W.-C. Wang, 2004).

In the *single-anchor approach*, the latent scale in each group is identified by fixing the item intercept of one item i to zero and the slope of the same item to unity, that is

$$\beta_{i1} = \beta_{i2} = 0, \quad \alpha_{i1} = \alpha_{i2} = 1. \tag{3}$$

DIF is then represented by (significant) group differences in the other, freely estimated item parameters. This procedure assumes that the item chosen for model identification is DIF-free. If this assumption cannot be motivated by theory, one can perform the analysis multiple times, each time using another item for scaling. This way, less strict assumptions are applied (Kopf, Zeileis, & Strobl, 2015a, 2015b; Pohl & Schulze, 2020). Assuming the test includes n items, $n - 1$ DIF tests per item would result. These significance tests can then be aggregated (see Kopf et al., 2015a, 2015b; W.-C. Wang, 2004 for a comprehensive overview).

In the *all-other approach*, DIF is tested for each item i by setting the item parameters of all other items to be equal across groups, that is,

$$\beta_{j1} = \beta_{j2}, \quad \alpha_{j1} = \alpha_{j2} \qquad \forall j \neq i. \tag{4}$$

For n items, this results in n models with one DIF test for each item. Thus, for the DIF test of a single item, it is assumed that all other items are DIF-free (A. S. Cohen, Kim, & Wollack, 1996). This assumption is violated when DIF occurs in at least one other item.

In the single-anchor and the all-other approaches, an anchor can also be built in a stepwise procedure (see e.g. Kopf et al., 2015a, 2015b; W.-C. Wang, 2004; Woods, 2009). This entails either iteratively building up the anchor (iterative forward) or purifying it (scale purification). The iterative forward procedure has been found superior in simulation studies (Candell & Drasgow, 1988; Kopf et al., 2015b, 2015a; Lautenschlager, Flaherty, & Park, 1994; Park & Lautenschlager, 1990). Kopf et al. (2015a) state as an assumption of the iterative forward approach that the majority of items are DIF-free. The iterative forward approach has by now also been extended to multiple-group settings (Huelmann, Debelak, & Strobl, 2019).

There is another line of research striving for a *simple structure* by identifying one set of anchor items. All of the respective approaches are in a sense based on the assumption that the majority of items is DIF-free. Several approaches have been developed within this line of research, which differ in the specific algorithm and scope.

Muthén and Asparouhov (2014) proposed the alignment method, where a loss function is used in order to identify the latent moments of the construct while minimizing measurement non-invariance in a way that there are few large non-invariant measurement parameters and many approximately invariant measurement parameters (see Pokropek, Lüdtke, & Robitzsch, 2020, for an extension). As such it makes the assumption that ’the number of non-invariant measurement parameters and the amount of measurement non-invariance can be held at a minimum’ (p.5). Thus, only if DIF-free items are the majority, this method yields unbiased latent moments. As a byproduct, the approach also provides ad-hoc testing of measurement invariance applying a kind of single-anchor approach. The approach can easily be applied to cases with a multitude of groups.

Robitzsch and Lüdtke (2018) proposed the use of regularization methods. These allow the detection of large DIF effects and set all small effects to zero through a penalty function. An advantage of this approach is that it allows for nonuniform DIF and categorical as well as continuous covariates.

The anchor point method by Strobl, Kopf, Hartmann, and Zeileis (2018) uses the Gini index from inequality research and applies it to DIF analyses. Compared to the iterative forward approach, this approach enables a more fine grained search for anchoring. In its current state, this approach is restricted to uniform DIF and the two-group case.

Cluster approach for uniform DIF

As was emphasized in the previous section, current DIF detecting methods rely on strong assumptions and only perform well under certain conditions. The assumptions have to be justified, which is not always feasible and/or done in practice. Based on the work of Lord (1977, 1980), Bechger and Maris (2015) presented an approach in which no assumptions are made in the first place. Instead several solutions of anchor item sets result. The approach allows for incorporating one out of several assumptions. These include all of the assumptions of the previous approaches in addition to decisions based on content knowledge of the items. Furthermore, the approach allows for evaluating the uncertainty of the group comparisons caused by anchor item selection (Pohl & Schulze, 2020).

By reframing the definition of DIF, Bechger and Maris (2015) make DIF tests of a specific item invariant to the choice of model constraints. Traditionally, DIF tests aim at detecting *DIF for single items*. DIF for the item difficulty (or in the slope-intercept parametrization for the item intercepts) β was accordingly defined as:

$$\text{DIF} = \beta_{i1} - \beta_{i2}, \tag{5}$$

with β_{i1} and β_{i2} denoting the difficulty (or intercept) of item i in group 1 and group 2, respectively (Lord, 1977). While DIF is clearly defined in theory, it can not be identified in applications due to scale indeterminacy: The estimated DIF depends on the scaling restriction chosen. If, for example, we opt for model identification through setting the first item's difficulty (or intercept) to zero in both groups, we assume that the first item shows no DIF. If we instead chose to fix the parameters of the second item for identifying the scale, we would translate the scale in each group by a different constant, resulting in different DIF values for the same item. If means of the two groups are fixed to, for example, zero, the assumption is made that there is no mean difference in the construct between groups. As DIF cannot be identified in practice, Bechger and Maris (2015) proposed to regard *DIF of an item relative to another item* instead:

$$\text{relative DIF} = R^{\beta}_{ij} = (\beta_{i1} - \beta_{j1}) - (\beta_{i2} - \beta_{j2}). \tag{6}$$

Note that there is not a single DIF value per item (as in previous approaches), but relative DIF values for each item compared to every item in the test. While DIF of an item (Equation 5) will be affected by the chosen identification restriction, relative DIF (Equation 6) is not. This useful property allowed Bechger and Maris (2015) to develop an approach that does not need any initial assumptions about DIF.

Bechger and Maris (2015) proposed to group the items according to their relative DIF. Within a set the items function similar. Each resulting item set is a possible candidate for anchoring. Due to scale indeterminacy, we cannot know from the data alone which of the item sets is DIF-free or whether there is a DIF-free set at all.

Bechger and Maris (2015) determined subsets of items through visual inspection of the whole $n \times n$ matrix of relative DIF values. This visual clustering of items works well with small simulated item samples. However, as the authors noted, this would not be sufficient for applied data analyses. Pohl and Schulze (2020) proposed to add a clustering step for identification of clusters. Instead of using the whole matrix of relative DIF, they made use of the fact that this matrix is skew-symmetric and of rank two. Regardless of the reference item for calculating relative DIF, the relative DIF of any two items can be obtained from every row (or column) of the matrix. As such, the clustering issue is reduced to clustering a unidimensional vector of distances.

Because of its property of optimality for unidimensional data, Pohl and Schulze (2020) proposed to use k-means clustering by dynamic programming (H. Wang & Song, 2011) to identify the clusters. k-means clustering in general requires that the number of clusters is specified. As the number of clusters is hard to know beforehand in applied data analysis, Pohl and Schulze (2020) proposed using a threshold criterion. The threshold represents the maximum within-cluster range of relative DIF 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 cluster differences in relative DIF exceed the given threshold.

Pohl and Schulze (2020) evaluated their approach in comparison to the equal-mean-difficulty and the iterative forward approaches for uniform DIF. The authors only found a minor impact of sample size and number of missing values in DIF-free items (i.e., standard error of item parameters) or of relative number of DIF-free items on the performance of the approach. The performance was slightly better when sample size was large and standard error was low. Performance was mainly impacted by DIF-size and threshold settings. Slight bias in mean differences occurred only for very large DIF and threshold settings that were about the size or larger than the present DIF. The authors suggested to set the threshold for cluster selection to a DIF size value one is willing to tolerate. In contrast to the previous approaches, the cluster approach also performed well in cases with concurrent unbalanced DIF, large DIF, and when the number of DIF-free items was small.

Previous approaches classify items into those that possibly display DIF and those that do not. The latter are then used as anchor items for estimating group differences. The cluster approach on the other hand offers multiple candidates for anchor item sets. This reflects the fact that without further knowledge, it cannot be stated which items are DIF-free in an absolute sense. Compared to the other approaches, the big advantage of this is that researchers can evaluate all possible solutions, choose between different assumptions and may also base their decision on content knowledge. They may also impose any of the assumptions made by the previous approaches for choosing an item cluster for anchoring. This explicates the variety of options and makes the assumptions transparent and available for discussion. In addition, researchers may evaluate the uncertainty stemming from the choice of an item cluster for anchoring (i.e., the choice of assumptions) by investigating the target parameters using each of the item clusters for linking. Although the cluster approach is not affected by the scale indeterminacy problem and does not need assumptions on item DIF in the first place, the researcher has to eventually choose a cluster for anchoring. Moving the decision regarding assumptions to a later stage in the analyses allows for evaluating various solutions and for deciding deliberately on assumptions. Of course, in order to avoid hacking of results, the choice of assumption needs to be well justified. Displaying results using the other clusters helps in depicting the uncertainty stemming from the choice of a cluster for anchoring.

Research question

The cluster approach for uniform DIF is promising due to the following three advantages: it is able to identify item clusters that result in unbiased parameter estimates under a wide range of conditions, allows for a broad set of assumptions including decision based on content knowledge, and allows for evaluating the impact of the choice of cluster on parameter estimates. However, so far it is only applicable to 1PL models for evaluating uniform DIF, although many studies use 2PL models.

2PL models are not only used for cognitive scales like competence measures in PISA (OECD, 2017) or PIAAC (Rammstedt et al., 2012), but also for non-cognitive scales like those in the measurement of personality (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) or attitudes (e.g., Himelfarb & Esipova, 2016). Next, we present an extension of the cluster approach to nonuniform DIF, evaluate its performance, and illustrate its use with the goal of making the cluster approach available to a wider range of applications.

The cluster approach for nonuniform DIF

The cluster approach for nonuniform DIF relies on the 2PL model and it is assumed that this model holds in each of the two groups (or time points) to be compared. We define relative DIF in item intercept β according to Equation 6 and relative DIF in item slope α as

$$R_{ij}^{\alpha} = [ln(\alpha_{i1}) - ln(\alpha_{j1})] - [ln(\alpha_{i2}) - ln(\alpha_{j2})] \tag{7}$$

that is the differences in log item slopes between any two items i and j within group g. We call the matrix of all R_{ij}^{α} values ΔS-matrix.

Due to the dependency of the relative DIF in intercept parameters β on constraints on item slope parameters α (see supplementary materials for a derivation of this dependence), an extension of the cluster approach is not straightforward. Depending on the restrictions used for identification of item slope parameters, relative DIF R_{ij}^{β} changes. This impedes the extension to a simple two-dimensional cluster problem.

With the goal of finding items that are measurement invariant in *both* item parameters, we decided to follow a two-step approach under consideration of the above mentioned difficulties. First, items are arranged in clusters that are homogeneous in item slopes. Second, these clusters are further investigated for items whose intercepts additionally function similarly. This rationale is in line with the sequence of MI testing proposed by Meredith (1993) and Millsap (2010). Here MI in item slopes is tested first (metric MI), followed by item intercepts (scalar MI).

In the first step of identifying item clusters with homogeneous slopes, we use a multiple-group model with an arbitrary identification restriction for each group (e.g., setting the mean and variance of the person variable to zero and one, respectively, in each group). Relative DIF in item slopes is then computed according to Equation 7. As the ΔS-matrix is skew-symmetric with a rank of two and the absolute position of the scale is arbitrary for relative DIF, we arbitrarily choose the first column and apply a $k-$means algorithm (H. Wang & Song, 2011) to that vector of relative DIF values in item slopes. Like in the approach for uniform DIF, the number of clusters needs to be determined, for which we propose the use of thresholds as described above. These determine how much the relative difference in log item slopes may vary between the items within a cluster. An alternative to thresholds is the specification of cluster number itself.

In the second step, we investigate the homogeneity of item intercepts for each cluster extracted in the first step. A model is estimated for each cluster, setting the item slopes for items from the respective cluster as well as the intercept of one arbitrary item of that cluster to be the same across groups (and the mean and the variance of the first group to be zero and one, respectively). For each model, relative DIF in item intercept parameters is then computed only for the items that correspond to the respective cluster from the first step (for which item slopes were set equal)[1]. Relative DIF for item intercepts is calculated according to Formula 6. Again, k-means clustering is applied to the relative DIF for item intercepts. This corresponds to the procedure for uniform DIF explained above, with the difference that it is only applied to a subset of items in accordance with the clustering from the first step. The approach results in final clusters that consist of items that are homogeneous in both item intercept and slope parameters.

For estimating mean differences across groups, the items of the chosen cluster are used as anchor items. Within the first group, latent mean and variance are fixed to zero and one, respectively, and item parameters of the anchor items are set to be equal across the two groups.

Evaluation of the approach in a simulation study

We designed a simulation study in order to evaluate the performance of the proposed approach. Our setup reflects the notion that DIF in item slopes (α-DIF) and item intercepts (β-DIF) can occur in the same item and that it is important to find anchor items that function similarly on *both* parameters.

Data generation

Data generation was set up according to a 2PL model with two groups ($N = 1000$ each). This reflects a lower bound in LSAs. As shown by Pohl and Schulze (2020), increasing N improved results. Person parameters θ_g for the two groups $g \in \{1, 2\}$ were independently drawn from a normal distribution with $\theta_1 \sim N(-0.5, 1)$ and $\theta_2 \sim N(0.5, 1)$[2]. We simulated a measurement instrument comprising 24 items assessing a single latent construct. These 24 items were distributed over four item sets with varying DIF properties (each with six items). The first set represented DIF-free items, the second displayed β-DIF, the third α-DIF, and the fourth DIF in both parameters. Item parameters were

[1] We used all items in the estimation of each measurement model in the second step in order to enhance power and identification. Note that in each analysis, we only investigated relative DIF in item intercepts of items that were identified as invariant in item slopes.

[2] Differences in person parameter distributions result in different test targeting and, thus, in different standard errors for item and person parameters. We refrained in varying person parameter distributions in this simulation as Pohl and Schulze (2020) have already shown that standard errors have only minor effects on the performance of the approach. However, we simulated a mean difference between groups of one standard deviation, which depicts a rather large but still realistic impact. Based on previous work (Pohl & Schulze, 2020), we expect performance of the cluster approach to be better with better test targeting.

randomly and independently drawn from a normal distribution. In case of item slopes, this was a truncated normal with $\alpha_{ig} \sim TN(1, 0.25, lower = 0.7, upper = 1.5)$. These values were chosen in order to achieve empirically reasonable item slopes. The item intercepts were drawn from $\beta_{ig} \sim N(0, 1)$.

In our simulation setup, we aimed at investigating a wide range of DIF situations under a 2PL model. We thus varied (1) DIF size a and b for item slopes and item intercepts, respectively, (2) balancedness of DIF, and (3) the mutual direction of α- and β-DIF. We decided to scrutinize minor, medium, and large DIF sizes, which were $a \in (0.05, 0.2, 0.5)$ for item slopes and $b \in (0.1, 0.5, 0.8)$ for item intercepts. β-DIF for an item i was defined as $\beta_{i2} - \beta_{i1}$, while α-DIF was defined as $ln(\alpha_{i2}) - ln(\alpha_{i1})$. Thus, the three α-DIF sizes reflected item slopes that are 1.05, 1.22, or 1.65 times larger (or 0.95, 0.82, or 0.61 times smaller) in the second group than in the first group. These DIF sizes correspond to the ranges chosen in previous simulations (González-Betanzos & Abad, 2012; W.-C. Wang & Yeh, 2003). Secondly, we varied balancedness of DIF in both parameters (depicted in Figure 1 for the exemplary case of $a = 0.2$ and $b = 0.5$), as unbalancedness has been shown to be more challenging in previous approaches (Pohl & Schulze, 2020). When DIF is balanced, there are as many items displaying positive DIF as there are items with the same amount of negative DIF. Thus, on average, there is no difference in item slopes or intercepts across groups. In the balanced condition, one set of items was affected by positive α- or β-DIF and one set of items by negative DIF of the same size. In the unbalanced condition, DIF does not cancel out and instead favors one group. Thirdly, we varied whether α- and β-DIF pointed in the same direction or not. This was achieved by varying DIF directions in the fourth item set, in which both parameters had DIF, with α- and β-DIF being either opposed or alike[3] (see Figure 1).

In total, there were five factors that were fully crossed: α-DIF size (with three values), β-DIF size (with three values), balancedness of α-DIF (two types), balancedness of β-DIF (two types), as well as directionality of α-DIF and β-DIF (two directions). This resulted in 72 conditions overall. For each condition, we generated $r = 200$ data sets.

Analysis

We analyzed the data using the proposed cluster approach for nonuniform DIF under various combinations of threshold settings for α- and β-DIF. Threshold settings were chosen to cover the whole range of DIF. For the α-thresholds we used 20 values ranging from 0.02 to 0.4 by steps of 0.02 and for β-thresholds 20 values ranging from 0.05 to 1.0 by steps of 0.05. α- and β-thresholds were fully crossed in the simulation.

[3] The various conditions were achieved by the following equations (also see formulas in the lower left panel of Figure 1): The item parameters of the first item set were the same in both groups. Item intercepts of the second set of items were set to be $\beta_{i2} = \beta_{i1} + b$, thus no DIF was introduced to item slopes. In the third set of items, item intercepts were set to be the same as in the first group, while the item slopes in the second group were $\alpha_{i2} = e^{ln(\alpha_{i1})+a\cdot c}$, with $c \in (-1, 1)$. In the fourth set of items, the item parameters in the second group were derived by $\alpha_{i2} = e^{ln(\alpha_{i1})+a\cdot c\cdot d_\alpha}$ and $\beta_{i2} = \beta_{i1} + b \cdot d_\beta$, with $d_\alpha \in (-1, 1)$ and $d_\beta \in (-1, 1)$. Parameters c, d_α, and d_β described the balancedness and mutual direction conditions.

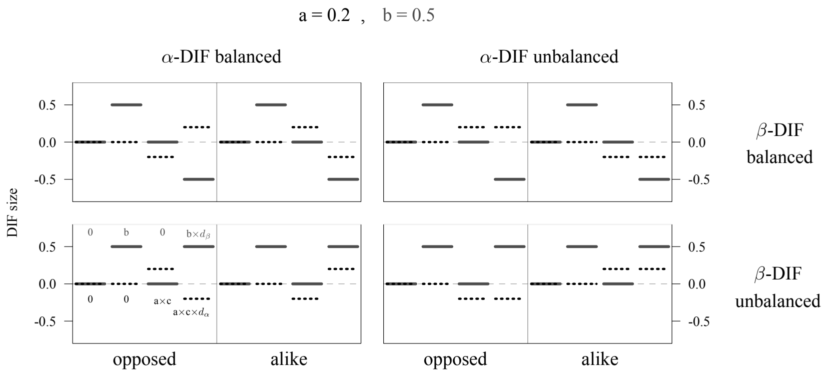

Figure 1:
Data generating conditions in the simulation study with the five simulated factors: α-DIF size, β-DIF size, balancedness of α-DIF and β-DIF, and mutual direction of α-DIF and β-DIF. Each of the four columns within one graph represents an item set of six items. Solid lines indicate DIF in item intercepts and dotted lines describe DIF in item loadings within the respective item set.

In order to evaluate and aggregate the results, we chose the cluster with the largest number of DIF-free items identified by our approach. This cluster was labeled *focus cluster* and was used for further analyses. If there were more than one cluster with the same number of DIF-free items, one of the clusters was chosen at random. We then evaluated cluster length, hit rate, as well as bias in the estimation of latent mean differences and variance ratios. Cluster length gives the number of items in the focus cluster. Hit rate was computed as the percentage of DIF-free items on the total number of items in the focus cluster. Bias in the estimated mean difference and bias in the estimated variance ratio across groups were determined by applying a multiple-group IRT-model to the data using mirt (Chalmers, 2012). We set the mean and variance of ability in the first group to zero and one, respectively, and fixed the item parameters of the items from the focus cluster to equality in both groups.

Results

Since we have far more results than can be presented in detail within the main body of the manuscript, we will only depict results that illustrate the mechanism. The whole range of results for all conditions is given in the supplementary material.

Figures 2 and 3 show the hit rate and the cluster length, respectively, for the focus cluster in one of the challenging conditions of unbalanced α-DIF, unbalanced β-DIF, and opposite direction of DIF. The black square with the cross indicates the size of simulated DIF. Cluster length and hit rate depended on α-DIF, β-DIF along with their chosen thresholds. Higher hit rates were achieved for greater DIF on both parameters. Smaller thresholds resulted in higher hit rates (see Figure 2), but also in smaller clusters (see Figure 3). If thresholds were chosen to be about the size of the induced DIF in both parameters, the focus cluster contained fewer items than there were DIF-free items (i.e., six in our simulation). This effect was even larger for smaller DIF sizes. The results were similar in the other conditions (see supplementary material for the respective figures).

Bias in mean differences between the groups is depicted for the conditions of unbalanced α-DIF, opposite direction of DIF, and balanced β-DIF in Figure 4, and for the case of unbalanced β-DIF in Figure 5. The results are similar in the other conditions (see supplementary materials) with slightly more bias in unbalanced than in balanced β-DIF conditions. There was hardly any bias in mean differences in any of the conditions. Only in cases of unbalanced β-DIF, when the β threshold was larger than the true DIF, bias in mean difference occurred. The bias became larger with greater thresholds exceeding the true DIF size. This is due to the fact that DIF items were included in the focus cluster that deviated from the true cluster in one direction. The bias was negligible for small β-DIF and considerable for large β-DIF. Note that for a β-DIF of 0.8 we only present thresholds up to 1; bias would presumably become larger for larger thresholds. In case of balanced β-DIF, mean bias only occurred in conditions with large α-DIF, at least medium β-DIF, and a threshold for item intercepts exceeding the true β-DIF.

There was hardly any bias in variance estimation in the second group in any of the conditions and thus, no bias in the ratio of the variances across groups. Unsystematic slight bias for some conditions occurred up to an absolute maximum value of 0.1 on the log scale. Bias in the ratio of variances in person parameters across groups is depicted for the conditions of unbalanced β and α-DIF and opposite direction of DIF in Figure 6. The same lack of any pattern was present in the other conditions (see supplementary materials).

Illustration of the approach in an empirical example

The national educational panel study (NEPS, Blossfeld, Roßbach, & von Maurice, 2011) is a longitudinal large scale study in Germany. It focuses on educational development over the whole life span, using different tests for every age group. Here, we will use cross-sectional data of ninth-graders' mathematical competence and compare persons with and without migration background. Mathematical competence was assessed by 22 items comprising the four content domains of quantity, space and shape, change and relationships, and data and chance (Duchhardt & Gerdes, 2013). For the illustration of our approach, we used 20 dichotomous items. In NEPS, having a migration background includes first generation immigrants up to persons having at least two grandparents not born in Germany (Olczyk, Will, & Cornelia, 2014). In total, we had a sample consisting of $n_1 = 10,253$ students without and $n_2 = 4,223$ (29%) students with migration background. 3.3% of all responses to the math competence items were missings. For analyses, we used a full information maximum likelihood estimator for a two group 2PL model, as implemented in the R package mirt (Chalmers, 2012).

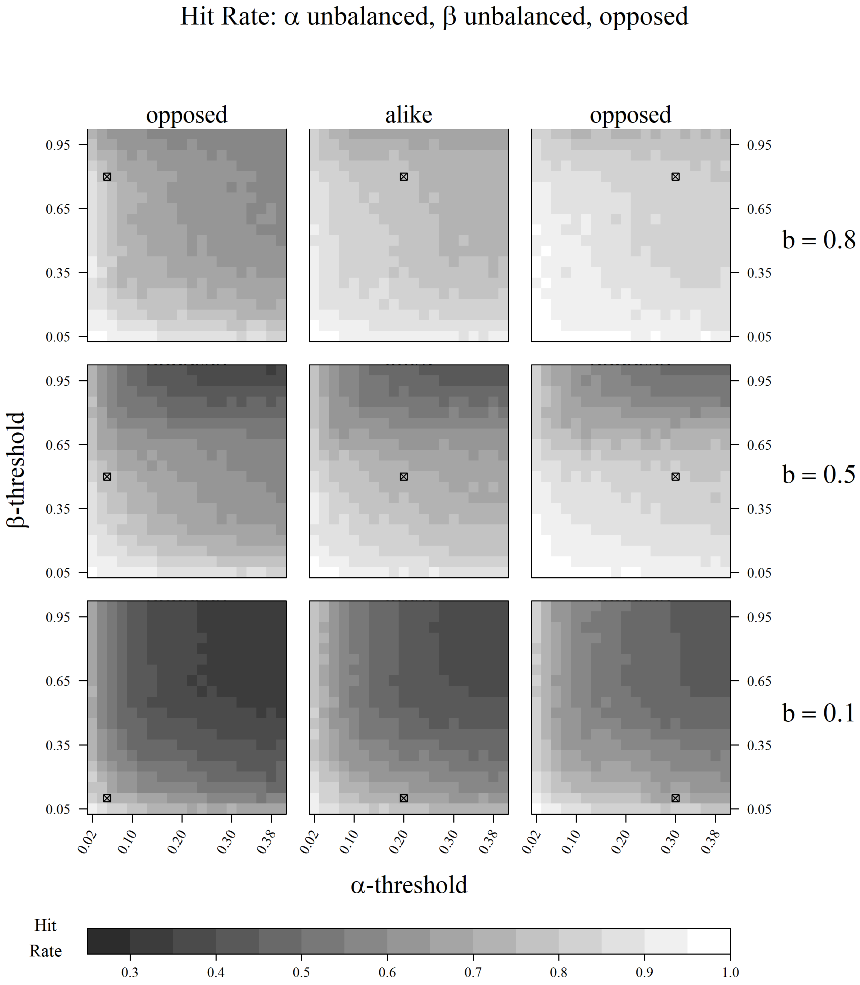

Figure 2:
Hit rate for the focus cluster in the condition of unbalanced α-DIF, unbalanced β-DIF, and opposed direction of DIF. The black square with the cross depicts the α- and β-DIF generated in the respective condition.

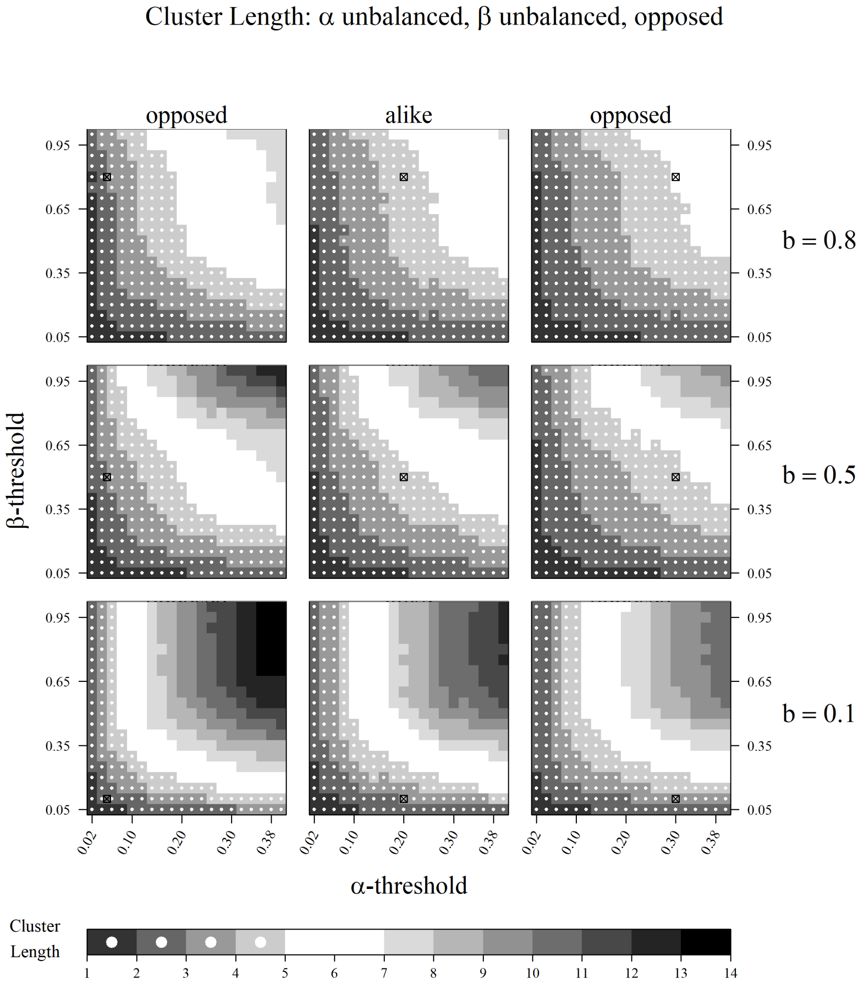

Figure 3:
Cluster length for the focus cluster in the condition of unbalanced α-DIF, unbalanced β-DIF, and opposed direction of DIF. The black square with the cross depicts the α- and β-DIF generated in the respective condition.

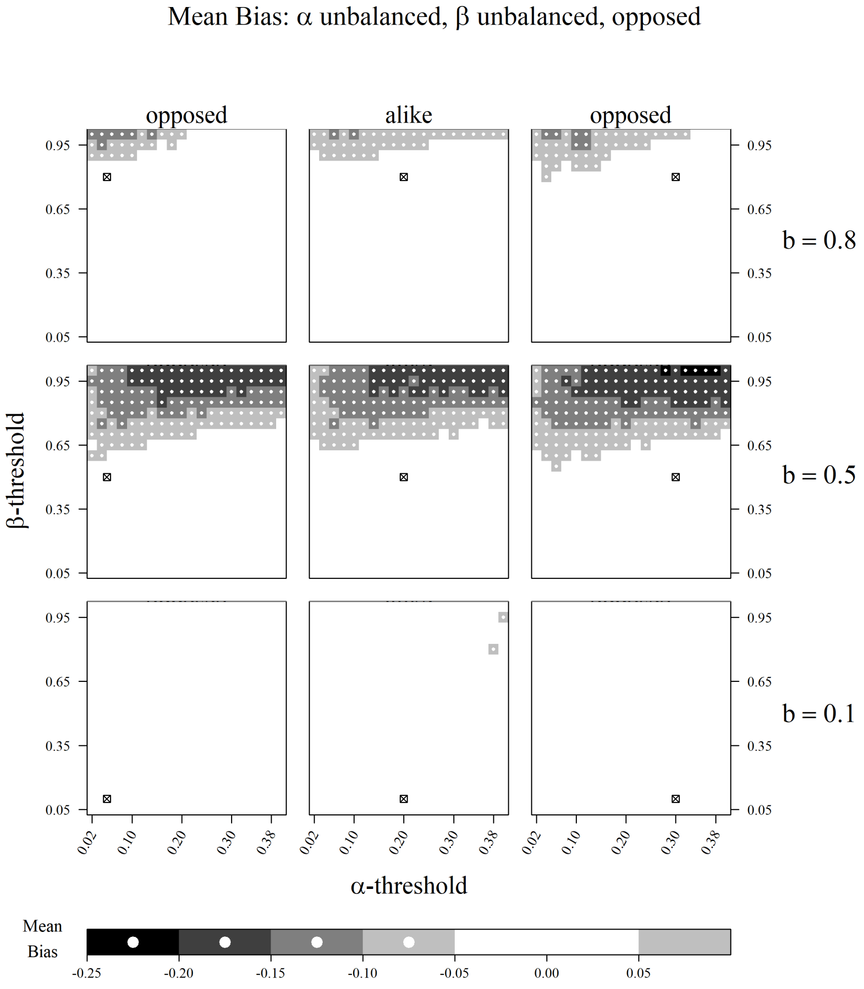

Figure 4:
Bias in estimated mean difference for the focus cluster in the condition of unbalanced α-DIF, unbalanced β-DIF, and opposed direction of DIF. The black square with the cross depicts the α- and β-DIF generated in the respective condition.

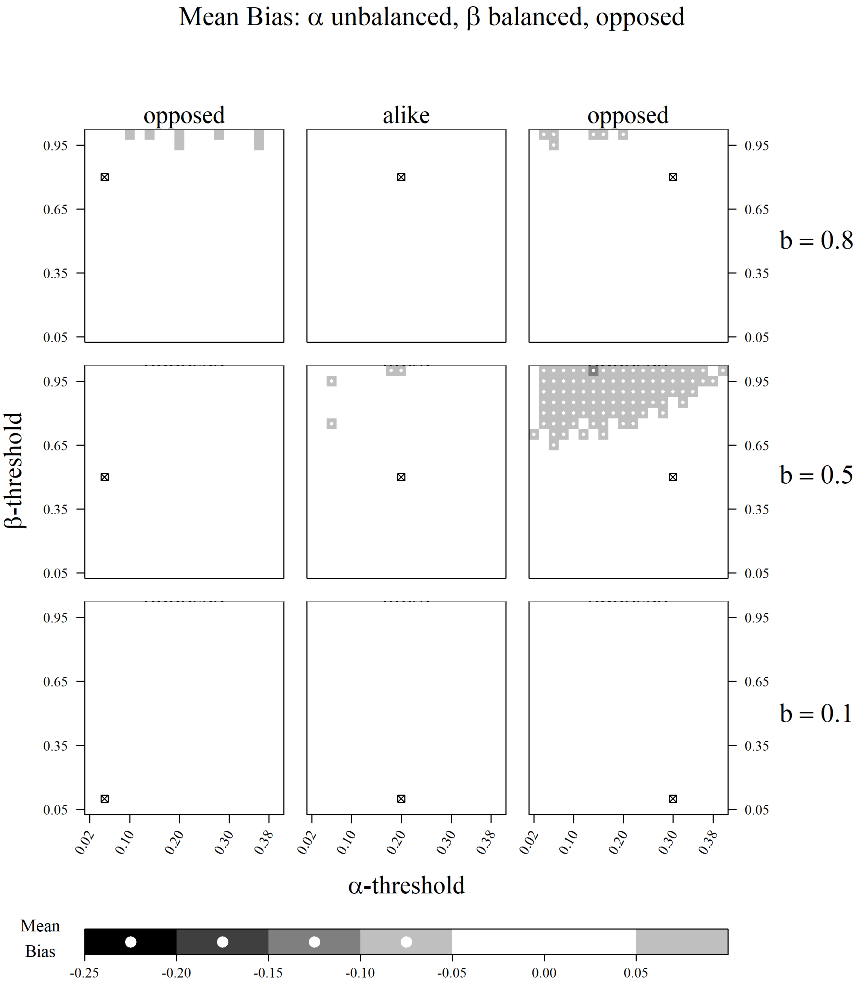

Figure 5:
Bias in estimated mean difference for the focus cluster in the condition of balanced α-DIF, unbalanced β-DIF, and opposed direction of DIF. The black square with the cross depicts the α- and β-DIF generated in the respective condition.

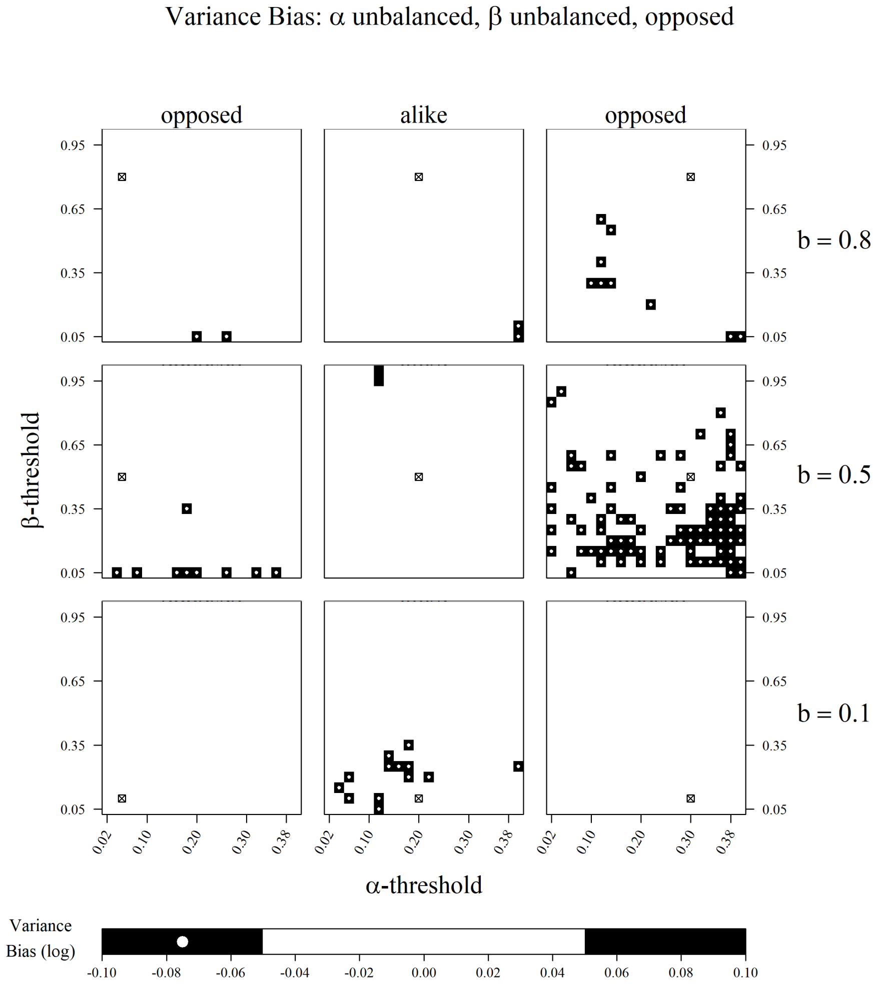

Figure 6:
Bias in estimated relation of variances for the focus cluster in the condition of unbalanced α-DIF, unbalanced β-DIF, and opposed direction of DIF. The black square with the cross depicts the α- and β-DIF generated in the respective condition.

First we checked MI globally by comparing an unrestricted baseline model to a fully restricted model (all intercepts and slopes were equal across groups). A significant violation of full MI was found ($\chi^2(38) = 145.06, p < .001$, Baseline-BIC = 319609.7, Restricted-BIC = 319390.7).

We applied the cluster approach for nonuniform DIF with thresholds (α-threshold = 0.2, β-threshold = 0.4) that can be considered to reflect small DIF (Dorans & Holland, 1992; Pohl & Carstensen, 2012). We found three item clusters for the math competence scale. The top of Figure 7 shows the results from the first step. Two clusters of items were identified that were homogeneous in item discriminations (up to a tolerance of 0.2). In the next step, the model was estimated twice, first by setting the item discriminations of the items in the first cluster (depicted in dark grey) to be equal across groups, second, by setting the item discriminations of the items in the second cluster (depicted in light grey) to be equal. For each analysis in the second step, relative DIF in item difficulties was computed and clustered. The results of this step are shown in the middle row of Figure 7. The items from the second cluster of the first step (light grey scale) were homogeneous in difficulty parameters (up to a tolerance of 0.4 logits) and made up one of the final three clusters. The items from the first cluster of the first step (dark grey scale) were split into two clusters which differed in relative DIF of item difficulties. The last row in Figure 7 summarizes the results, depicting the three item clusters. The items in each of these clusters function similarly in terms of item discrimination *and* item difficulty. The clusters did not match the content domains of the items, thus ruling out multidimensionality stemming from content as a source of DIF.

When using each of these clusters as anchor for estimating latent differences in math competence between migrants and non-migrants, slightly different estimates resulted for the unstandardized mean differences (dot-dashed line marked cluster: -0.40 logits, solid line marked cluster: -0.38 logits, dashed line marked cluster: -0.53 logits) as well as for the proportion of variances between the two groups (dot-dashed line marked cluster: 0.67, solid line marked cluster: 0.92, dashed line marked cluster: 1.01). Depending on the cluster chosen for linking, the estimated standardized mean difference in math competence was -0.59 when using the dot-dashed line marked cluster, -0.41 with the solid line marked cluster cluster, and -0.52 when using the dashed line marked cluster. Although all three effects indicate a substantially lower math competence for ninth-graders with migration background, the effect sizes vary from small to medium (J. Cohen, 1992).

While previous approaches would have identified only one group of presumably DIF-free items which function similarly and would have used them for linking, the cluster approach depicts different solutions. One may choose one of these clusters based on content or apply any of the previously mentioned assumptions (e.g., that the largest cluster is DIF-free). Alternatively one may report the results using each of the clusters, thus depicting the uncertainty stemming from the anchor item choice.

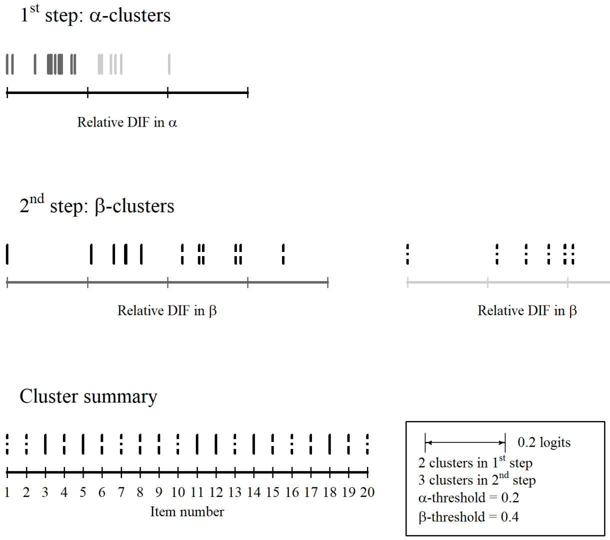

Figure 7:
Results of the cluster approach for nonuniform DIF in the empirical example on mathematical competences. Gray scale and line type indicates the cluster an item was assigned to in the respective step. The same gray scale and line type represents the same cluster.

Discussion

Summary

DIF analysis has become a common tool in scale construction and when assessing MI for score comparisons in applied research. If MI does not hold for the whole scale, researchers strive for partial MI. Traditionally, DIF analysis aimed at identifying a subset of DIF-free items by posing some assumptions on the occurrence of DIF. In practice, researchers are not always aware of these often implicit assumptions, rendering the analysis' results questionable. In the cluster approach, assumptions about the occurrence of DIF need to be made explicit. The approach yields multiple clusters of items with items functioning similarly within each cluster, but functioning differently between clusters. Faced with several options, the researcher may now make a conscious decision regarding the anchor set used for substantive analyses.

We extended the cluster approach of Bechger and Maris (2015) and Pohl and Schulze (2020) to 2PL models, as these are most prevalent in educational and psychological research. Our simulation study provided strong support for the algorithm's validity and ability to deal with various situations arising in DIF analysis. The generated data contained 75% items that displayed DIF on at least one parameter. Despite this challenge, the proposed two-step algorithm was reliably able to identify an item cluster, that, when used as anchor, yielded negligible bias in means and variance ratio differences across groups. In order to choose an appropriate threshold, one has to balance bias and efficiency: small thresholds yield small clusters, which are less biased but also less efficient. We propose to use a threshold as large as the DIF-size one is willing to tolerate.

In the empirical example, DIF analysis of a math competence test yielded three item clusters, each of which being a candidate for anchoring. While previous approaches would have only presented one solution, the proposed approach depicts all possible solutions and thus allows for applying one of different assumptions or even depicting the uncertainty stemming from the choice of an item cluster for anchoring.

Limitations and outlook

One may argue that choosing between assumptions after the clusters have been extracted poses the risk of choosing the cluster that results in the most favorable results. While similar risks are also present when using previous approaches (e.g., implementing different approaches or different items as anchor items and choosing between them depending on the results), this risk is more pronounced in the proposed approach. However, this decision is made much more transparent in the cluster approach. In practice researchers facing a MI problem usually do not know in advance which or how many items are affected by DIF. If they knew, they would alter the measurement instrument before data collection. Thus, they are usually uncertain which assumption they should pose. As with other statistical procedures and as discussed within the open science movement, transparency in the analyses and good scientific practice are warranted. So far, assumptions made by the approaches for identifying anchor items have hardly been stated nor has their plausibility been discussed in applied research. The cluster approach makes these assumptions transparent and requires some argumentation for the choice of an assumption. Furthermore, depicting the results of using all clusters may help in judging the uncertainty stemming from the choice of an item cluster for anchoring.

In many applications comparisons across multiple groups (e.g., countries) must be made and DIF may occur on continuous variables (e.g., age). Some previous approaches allow for categorical covariates with more than two groups (iterative forward approach, Huelmann et al., 2019; alignment method, Asparouhov & Muthén, 2014) or allow continuous covariates (regularization method, Robitzsch & Lüdtke, 2018; non-linear MIMIC models, Bauer, 2017). Thus, extending the cluster approach to multiple groups and continuous covariates is called for (see Schulze & Pohl, in press).

In contrast to the cluster approach presented here, the ’simple structure’ approaches assume that the majority of items is DIF-free and aim at identifying one set of anchor items. The cluster approach will usually yield multiple anchor item sets. Nevertheless, there are some ties to be found between these lines of research. Strobl et al. (2018) showed that their anchor point approach relates to the cluster approach. The authors presume that local maxima may indicate other item clusters, which may be similar to the clusters in the cluster approach. Thus, there may be connections between the different lines of approaches and advances in extending the ’simple structure’ approach to categorical or continuous covariates may also apply to the cluster approach. In fact, bringing together the different lines of anchor item selection strategies may open up the way to a general framework in which the strengths of both research traditions can be integrated.

Supplements

The supplementary material can be found here: https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-2/Pohl_2020Q2_supplements.pdf

Acknowledgments

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).

We thank Timo Bechger, Carolin Strobl, Thorben Huelmann, and Alexander Robitzsch for discussions and helpful comments on our work as well as the HPC service of Freie Universität Berlin for support and computing time.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. doi: 10.1080/10705511.2014.919210

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. doi: 10.1037/met0000077

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. doi: 10.1007/s11336-014-9408-y

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. B. Lord & N. M (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), 176–181. doi: 10.1097/01.mlr.0000245143.08679.cc

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253–260. doi: 10.1177/014662168801200304

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. doi: 10.1207/S15327906MBR3604_03

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26. doi: 10.1177/014662169602000102

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98–101.

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (ETS Research Report No. RR-92-10). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1992.tb01440.x

Duchhardt, C., & Gerdes, A. (2013). *NEPS technical report for mathematics – Scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 22). Bamberg, Germany: University of Bamberg.

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*(4), 134–145. doi: 10.1027/1614-2241/a000046

Hidalgo, M. D., & Gómez-Benito, J. (2010). Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., p. 36 - 44). Oxford: Elsevier. doi: https://doi.org/10.1016/B978-0-08-044894-7.00242-6

Himelfarb, I., & Esipova, N. (2016). Commitment to Islam in Kazakhstan and Kyrgyzstan: An item response theory analysis. *The International Journal for the Psychology of Religion*, *26*(3), 252–267. doi: 10.1080/10508619.2015.1033899

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

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

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. doi: 10.1177/0013164414529792

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103.

doi: 10.1177/0146621614544195

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210–231. doi: 10.1007/s11336-013-9347-z

Lautenschlager, G. J., Flaherty, V. L., & Park, D.-G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement*, *54*(1), 21–31. doi: 10.1177/0013164494054001003

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). Lisse, Netherlands: Swets & Zeitlinger Publishers.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. doi: 10.1007/BF02289699

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

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9. doi: 10.1111/j.1750-8606.2009.00109.x

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

OECD. (2017). *PISA 2015 technical report.* Paris, France: OECD Publishing.

Olczyk, M., Will, G., & Cornelia, K. (2014). *Immigrants in the NEPS: Identifying generation status and group of origin* (NEPS Working Paper No. 41a). Bamberg, Germany: Leibniz Institute for Educational Trajectories.

Park, D.-G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement*, *14*(2), 163–173. doi: 10.1177/014662169001400205

Pohl, S., & Carstensen, C. (2012). *NEPS technical report – Scaling the data of competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg.

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

Pokropek, A., Lüdtke, O., & Robitzsch, A. (2020). An extension of the invariance alignment method for scale linking. *Psychological Test and Assessment Modeling*, *62*(2), 305–334.

Rammstedt, B., Ackermann, D., Helmschrott, S., Klaukien, A., Maehler, D. B., Martin, S., … Zabal,

A. (2012). *PIAAC 2012: Overview of the main results*. Münster, Germany: Waxmann.

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, Jena, Germany.

Schulze, D., & Pohl, S. (in press). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *x*, xxx–xxx.

Strobl, C., Kopf, J., Hartmann, R., & Zeileis, A. (2018). *Anchor point selection.* Presentation at the International Meeting of the Psychometric Society (IMPS) 2018 in New York City, New York.

Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, *3*(2), 29–33.

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*(3), 221–261. doi: 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*(6), 479–498. doi: 10.1177/0146621603259902

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi: 10.1177/0146621607314044

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*(2), 223–233. doi: 10.1080/15434300701375832