Article

Applied Psychological Measurement
2021, Vol. 45(7-8) 477–493

DOI: 10.1177/01466216211042809
journals.sagepub.com/home/apm


Partial Measurement Invariance: Extending and Evaluating the Cluster Approach for Identifying Anchor Items

Steffi Pohl[1], Daniel Schulze[1], and Eric Stets[1]

Abstract

When measurement invariance does not hold, researchers aim for partial measurement invariance by identifying anchor items that are assumed to be measurement invariant. In this paper, we build on Bechger and Maris's approach for identification of anchor items. Instead of identifying differential item functioning (DIF)-free items, they propose to identify different sets of items that are invariant in item parameters within the same item set. We extend their approach by an additional step in order to allow for identification of homogeneously functioning item sets. We evaluate the performance of the extended cluster approach under various conditions and compare its performance to that of previous approaches, that are the equal-mean difficulty (EMD) approach and the iterative forward approach. We show that the EMD and the iterative forward approaches perform well in conditions with balanced DIF or when DIF is small. In conditions with large and unbalanced DIF, they fail to recover the true group mean differences. With appropriate threshold settings, the cluster approach identified a cluster that resulted in unbiased mean difference estimates in all conditions. Compared to previous approaches, the cluster approach allows for a variety of different assumptions as well as for depicting the uncertainty in the results that stem from the choice of the assumption. Using a real data set, we illustrate how the assumptions of the previous approaches may be incorporated in the cluster approach and how the chosen assumption impacts the results.

Keywords

measurement invariance, differential item functioning, scale indeterminacy, cluster analysis, anchor items

[1]Freie Universität Berlin, Berlin, Germany

Corresponding Author:
Steffi Pohl, Methods and Evaluation/Quality Assurance, Freie Universität Berlin, Habelschwerdter Allee 45, Berlin 14195, Germany.
Email: steffi.pohl@fu-berlin.de

Introduction

In order to meaningfully compare scores on a latent construct over time or across groups, the measures need to be on a common scale (e.g., Meredith, 1993). This is operationalized as measurement invariance in structural equation modeling (SEM) or the absence of differential item functioning (DIF) in item response theory (IRT). When evaluating a new scale, investigating measurement invariance has become a standard in psychometrics (AERA et al., 2014). If measurement invariance does not hold for a data set, a viable option is to strive for partial measurement invariance. For this purpose, researchers seek to identify a subset of items for which the assumption of measurement invariance holds and to use this subset as anchor items for group comparisons (Byrne et al., 1989). In this paper, we focus on models with one item parameter (i.e., the question of uniform DIF) and the two-group case.

While DIF is well defined in theory (e.g., Lord, 1977), it is not easy to identify in practice. This difficulty is due to scale indeterminacy (Lord, 1980). In both IRT and SEM models, the scale of the latent variable is not uniquely defined. Fixing the scale of the latent variable in single-group models is straightforward and does not threaten the conclusions of the analysis. However, fixing the scale does play an important role for group comparisons, comparisons over time, and the investigation of DIF. Depending on which scaling restrictions are used in each group, different assumptions are made. For example, fixing the item difficulty of the first item to zero in both groups implies that the first item is DIF-free. If the assumptions made for scale identification do not hold, group comparisons will be biased (e.g., W.-C. Wang, 2004). This issue cannot be resolved statistically. Instead, approaches aiming at identifying anchor items make different assumptions about DIF. Usually, when choosing an approach in applied research, the assumptions made are not evaluated for plausibility and are often not even mentioned despite their considerable impact on the results.

In this paper, (a) we extend the approach by Bechger and Maris (2015) by a step to identify homogeneous item clusters and (b) evaluate the performance of the approach and compare it with previous approaches.

Previous Approaches for Identifying Anchor Items

Several procedures have been developed for identifying uniform DIF, varying in their implicit or explicit assumptions (for an overview see, e.g., Magis et al., 2010 and Hidalgo & Gómez-Benito, 2010). Here, we focus on the strategies used to detect DIF-free items and the assumptions made in these approaches to tackle the scale indeterminacy issue. These strategies may be categorized into two types: (1) approaches that rely on assumptions about the average size of DIF (e.g., the equal-mean-difficulty approach) and (2) approaches that rely on assumptions about other items in the test (other-item approaches). Note, that each of these procedures provides a single set of anchor items and they all rely on inference statistics, which renders them dependent on the precision of item parameter estimation.

Equal-mean-difficulty approach. In the equal-mean-difficulty (EMD) approach, DIF is investigated by setting the mean difficulty of the items to be equal in both groups. Items displaying no DIF under this restriction are then chosen as anchor items for group comparisons. Underlying this procedure is the assumption that on average, items do not favor one group over the other (balanced DIF). This is a very strong assumption which has been considered implausible in many applications and whose violation leads to biased group comparisons (e.g., W.-C. Wang, 2004). Note, that EMD does not make any assumptions about the number of DIF-free items or about the presence of group mean differences.

Other-item approaches. There are different versions of the other-item approach (see Kopf et al., 2015a; 2015b and W-C. Wang, 2004, for a comprehensive overview). Here, we consider the iterative forward single-anchor approach, which was proposed by Kopf et al. (2015a) and demonstrated to outperform other approaches in simulation studies (Candell & Drasgow, 1988; Kopf et al., 2015a, 2015b; Lautenschlager et al., 1994; Park & Lautenschlager, 1990; W-C. Wang, 2004). The iterative forward approach involves testing DIF for each item multiple times. Differences in the difficulty of an item between two groups are tested by $k - 1$ models for k items. In each model, one of the other-item parameters is assumed to be measurement invariant across groups. This is done for each item and results in $k - 1$ DIF tests (using, e.g., likelihood ratio or Wald tests) per item. The results are subsequently aggregated for each item, where the aggregation can be based on whether the test within a model was significant, on the test statistic, or on the p-value (see, e.g., Kopf et al., 2015a, 2015b; W.-C. Wang, 2004). In order to eventually select a set of anchor items, Kopf et al. (2015a) suggest to iteratively building up the anchor item set from an initially empty set. The item which shows the lowest indication of DIF in the analyses is added to the anchor. In a next iteration, the same analyses are performed as before, however, this time with the item in the anchor being assumed to be measurement invariant. Kopf et al. (2015a) stop iteratively building up their anchor when the number of items in the anchor exceeds the number of further items detected to be DIF-free.

Kopf et al. (2015a) state that the iterative forward approach assumes that the majority of items is DIF-free. This assumption may be plausible in many, but not in all applications. Furthermore, although this assumption has been well delineated, its necessity and sufficiency for obtaining unbiased group comparisons has not been evaluated. We found only one publication that studied the performance of other-item approaches, when the minority of items is DIF-free (Woods, 2009). The authors showed that, under these conditions, performance deteriorates in terms of hit rate and mean bias. A direct comparison of the EMD and iterative other-item techniques has not yet been undertaken.

Allowing for Various Assumptions: The Cluster Approach

Bechger and Maris (2015) follow up on an idea formulated by Lord (1980) and propose an approach that identifies clusters of items that function similarly, highlighting multiple potential candidates for anchor item sets. Their approach allows for incorporating different assumptions and for evaluating the impact of the choice of assumption on the analyses results. While most of the previous studies aim at identifying *DIF of each single item*, Bechger and Maris (2015) argue that DIF cannot be identified for a single item, but the *functioning of an item can only be identified relative to another item*. This is inherent to the scale indeterminacy issue. Let us assume the competence of persons is scaled using a Rasch model (Rasch, 1960) and we want to compare the mean competence level of two groups (Group 1 and Group 2) using a multi-group design. The model equation of the two-group model is then given by

$$P\left(X_{pig}|\theta,\beta,g\right) = \frac{exp\left(\theta_{pg} - \beta_{ig}\right)}{1 + exp\left(\theta_{pg} - \beta_{ig}\right)} \tag{1}$$

with X_{pig} denoting the response of person p on item i, θ_{pg} the latent ability of person p, and β_{ig} the difficulty of item i in group g. Identification restrictions are needed for each group. Due to the scale indeterminacy problem, differences in item difficulties between groups cannot be meaningfully interpreted. However, what is unaffected by scaling restrictions are *relative item difficulties*

$$R_{ijg} = \beta_{ig} - \beta_{jg} \tag{2}$$

that is, the differences in item difficulties between any two items i and j within a group g. No matter which scaling restriction is used, the difference in item difficulty between any two items stays the same. In order to identify items that are measurement invariant to each other, Bechger and Maris (2015) consider the *difference in relative item difficulties* (DRIDs) between two groups

$$\text{DRID}_{ij} = R_{ij2} - R_{ij1} \tag{3}$$

Difference in relative item difficulties describes the extent to which the relation between item parameters within a group differs across groups. When $DRID_{ij}$ is (close to) zero, then the difference in item difficulties between the two items i and j are (almost) the same in both groups. As such, the DRID describes the degree to which two items are measurement invariant *to each other*.

Bechger and Maris (2015) propose a matrix of these DRIDs which they term ΔR-matrix. They recommend extracting item clusters with similar DRID by way of visual inspection of this matrix and construct a significance test for testing invariance of the identified item clusters. This results in different possible anchor item sets.

In our view, a strength of the approach is that it allows the imposition of any kind of assumption. First, researchers may look at the item content and base their decision for an anchor item set on substantive knowledge. Second, researchers may also make assumptions similar to those made in previous approaches, for example, that the largest cluster is DIF-free. Lastly, they may use each of the clusters for linking and as such depict the uncertainty in the results due to the choice of the anchor item set, which has several advantages: (i) it increases awareness of different possible solutions, (ii) it requires the researcher to explicitly state and justify the assumption made for choosing an anchor set and (iii) displays the amount of uncertainty that arises from the choice of anchor items. Although the cluster approach is very promising, so far it has not been used in applied research.

Research Question

Bechger and Maris (2015) have laid the groundwork for an approach to the issue of identifying invariant item clusters. Calculating DRIDs, graphs for visual inspection, and significance testing for predefined clusters are implemented in the R-package "dexter" (Maris et al., 2017). While the approach is very promising, it cannot be readily applied by substantive researchers. Clusters cannot always be easily identified by visual inspection when analyzing real data (see empirical example). In this paper, we aim to extend the approach proposed by Bechger and Maris (2015) to allow for identification of item clusters. This is necessary to facilitate application of the approach in practice.

Although the approach of Bechger and Maris (2015) is theoretically well derived and convincing, neither its performance nor its relation to other approaches has been investigated. One reason for this may be that the approach needs visual inspection for cluster identification, which makes it difficult to be implemented in a simulation study. Thus, our second aim is to evaluate the performance of the cluster approach under various conditions and compare its performance to that of previous approaches.

Extending the Cluster Approach

In order to make the approach by Bechger and Maris (2015) applicable to real data, we included a cluster identification step. To this end, we made use of the fact that the ΔR-matrix is

skew-symmetric, of rank 2. As per Bechger and Maris (2015, p. 324) item pair functioning can be evaluated using any row or column, we arbitrarily select the first. This reduces the data to a one dimensional vector.

For identifying clusters of invariant items, we propose to use a variant of the k-means algorithms: optimal k-means clustering by dynamic programming (H. Wang & Song, 2011). The advantage lies in its optimality for one dimensional vectors, where it is guaranteed to find a global optimum, which is not the case for heuristic k-means clustering (Lloyd, 1982).

The algorithm takes the entries of the first column of the ΔR-matrix sorted in non-descending order, which are denoted by $x_1, \ldots, x_n$. The aim is to cluster these n values into k clusters by minimizing the sum of squares of within cluster distances from the corresponding cluster mean (WSS). In order to guarantee optimality and to reduce the runtime for the algorithm, H. Wang and Song (2011) regard the sub-problem of finding the minimum WSS of clustering only the first $i \leq n$ numbers into $m \leq k$ clusters. They construct the matrix $D[i, m]$, in which the minimum WSS for each i and m is recorded. $D[i, m] = 0$, when $m = 0$ or $i = 0$. With j denoting the index of the smallest number in cluster m in an optimal solution to $D[i, m]$, the authors showed that the cluster solution for the first $j - 1$ values into $m - 1$ clusters must be optimal. This leads to the following recurrent equation

$$\boldsymbol{D}[i,m] = min_{m \leq j \leq i}\left\{D[j-1,m-1] + d\left(x_j,\ldots,x_i\right)\right\} \tag{4}$$

for $1 \leq i \leq n$, $1 \leq m \leq n$ and with

$$d\left(x_j,\ldots,x_i\right) = \sum_{h=j}^{i}\left(x_h - \overline{x}_m\right)^2 \tag{5}$$

and

$$\overline{x}_m = \frac{1}{n}\sum_{h=j}^{i} x_h \tag{6}$$

The authors then iteratively compute

$$d(x_1,\ldots,x_i) = d(x_1,\ldots,x_{i-1}) + \frac{i-1}{i}(x_i - \mu_{i-1})^2 \tag{7}$$

with

$$\mu_i = \frac{x_i + (i-1)\mu_{i-1}}{i} \tag{8}$$

and, thus, sequentially add further values to the previous solution. The start indices for each cluster are then identified from the solution in $D[i, m]$ by

$$\mathbf{B}[i,m] = argmin_{m \leq j \leq i}\left\{D[j-1,m-1] + d\left(x_j,\ldots,x_i\right)\right\} \tag{9}$$

for $1 \leq i \leq n$, $1 \leq m \leq n$.

There are different ways of determining the number of clusters to be extracted, we consider two: (1). Bayesian Information Criterion (BIC). (2). A threshold criterion which entails setting the maximum difference in DRID accepted between any two items within a cluster to a fixed number. The threshold can be chosen based on benchmarks of tolerable DIF as, for example, used with PISA (OECD, 2017). This procedure yields a solution that involves the least amount of clusters in

which DRID does not exceed the pre-specified threshold. A beta release of an R-package providing the functions for the analyses is available on GitHub (Schulze & Pohl, 2021b).

Investigating the Performance of the Approaches

Data generation. We generated data for comparison of latent mean scores across two groups. Person parameters θ_g for the two groups $g \in \{1, 2\}$ were independently drawn for each group from a normal distribution with $\theta_g \sim N(0, 1)$. This resulted in the true mean difference across groups being zero.[1] The measurement model in each group was assumed to follow the Rasch model with 24 items measuring a single latent construct. We used fixed item difficulty parameter sets with a range from −2 to 2, a mean of zero and a variance of 2.6 in each item cluster to cover the range of typical item parameter values in practice. The 24 items were distributed over three item clusters with one cluster representing DIF-free (FREE) items and two clusters representing DIF items.

In our simulation setup, we included the following factors (see Supplementary Figure S1 for a graphical depiction of some of the conditions): (1) Size of DIF-free cluster (1/3, 1/2, or 2/3 of the 24 items), (2) DIF size (0, 0.05, 0.1, 0.2, 0.4, or 0.8), (3) DIF unbalancedness (balanced and three degrees of unbalancedness), (4) sample size (500, 1000, or 2000 per group), (5) missing response proportion (0%, 20%, or 50%), and (6) location of missing values (on DIF items or DIF-free items), resulting in a 3*x*6*x*4*x*3*x*3*x*2 design. Note that for DIF sizes being zero, balancedness is irrelevant and for missing proportions of 0% the location of the missing values is irrelevant. We generated $r = 100$ data sets for each condition.

The setup includes both conditions in which the number of DIF-free items is the majority—an assumption for the iterative forward approach—as well as conditions in which it is not. In all conditions, the two DIF clusters consisted of the same number of items. The first cluster of items was generated to be DIF-free (FREE). Item difficulties of the second cluster of items (DIF1) were set to be $\beta_{i2} = \beta_{i1} + a$ and, thus, shifted by a constant a. We chose a so that it depicts a variety of conditions from no DIF ($a = 0$) to large DIF ($a = 0.8$). This corresponds to DIF categorizations used in some large scale assessments (Dorans & Holland, 1992). As such, the second set of items was more difficult in Group 2 as compared to Group 1 and disadvantaged Group 2. Item difficulties of the third cluster of items (DIF2) were set to be $\beta_{i2} = \beta_{i1} - a + c \cdot a$, with c depicting the amount of unbalancedness in DIF. c was set to be either 0, 0.5, 1.5, or 2. If c was zero, DIF was balanced. The larger c, the larger was the unbalancedness favoring Group 1. Note that with $c = 2$, the second and third set of items showed exactly the same amount of DIF. Thus, in this condition the two sets effectively made up one cluster. The condition with $a = 0$ and $c = 0$ is a control condition, in which all generated items are DIF-free. As most of the approaches rely on inferential statistics, we additionally varied the precision of item parameter estimation by inducing missing values. Missing values were induced as being missing completely at random either only in DIF-free items or in DIF items.

Analysis

We analyzed the generated data using the extended cluster, EMD and the iterative approaches. Multi-group Rasch models were estimated using full marginal-maximum likelihood framework of the mirt package, version 1.21 (Chalmers, 2012) in R (R Development Core Team, 2008). Applying the EMD approach, we fixed the average item difficulty in each group to be zero. DIF tests were then performed for each item by means of Wald tests. Items which were found to have non-significant group differences were used as anchor.

We implemented the iterative forward algorithm as described by Kopf et al. (2015a). We estimated the two-group model k times, fixing the difficulty of a different item to be the same across groups with each model estimation. Based on the suggestions by Kopf et al. (2015a), we

used the mean test statistic ranking criterion as DIF indicator. Using this criterion, first, the average test statistic across all $k - 1$ analyses was computed for each item. Subsequently, the median of these values across all items was computed.[2] We counted for each item how often each of the $k - 1$ test statistics for this item was above this median value and ranked the items according to this number. In the first iteration, the item with the highest rank (i.e., showing the lowest indications of DIF) was included in the anchor. Then, DIF analysis was again performed, however, with all items in the anchor being assumed to be measurement invariant in all analyses. This procedure was repeated until the number of items in the anchor exceeded the number of further items detected to be DIF-free.

When implementing the cluster approach, we first estimated the item difficulty parameters by constraining the latent means of both groups to be zero. Note that any parameter restriction would work as the DRIDs are invariant to scale identification. Difference in relative item difficulties were computed as described in equation (3). For cluster identification, we arbitrarily chose the DRID values of the first item and applied the optimal *k*-means algorithm using the R-package Ckmeans.1d.dp (H. Wang & Song, 2011). In order to determine the number of clusters to be extracted, we used BIC as well as threshold settings. In order to evaluate the impact of threshold choice, we varied the threshold for determining the number of clusters in all analyses to be 0.025, 0.04, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, or 0.8, covering the whole range of DIF sizes that were implemented in data generation.

We recorded the length of the anchor item set, the hit rate, and the bias in estimated mean difference between reference and focal groups as outcome measures. As in the extended cluster approach, different possible sets of items were identified, for depicting the results we only report the results of the 'best cluster', which we define as the cluster with the most DIF-free items. If more than one cluster shared the maximum number of DIF-free items, one of the clusters was chosen at random. The hit rate was computed as the number of DIF-free items relative to the number of items in the anchor item set or cluster.[3] Bias in the estimated mean difference between groups was evaluated by using the identified anchor items or the best cluster for linking in concurrent calibration. To this end, item parameters of the chosen item set were fixed to be the same in both groups.

Results

We did not encounter any convergence problems in implementing any of the approaches. We evaluated the impact of the different factors by running an ANOVA on hit rate and mean bias for each approach separately (see Table S1 and S2 in the Supplementary Material). We found a large impact of DIF size, balancedness, and number of DIF-free items. Thus, we present Figures that illustrate these results. Figures 1–3 show anchor length, the hit rate, and the bias in mean difference, respectively, exemplary in the condition with $N = 500$, no missing values, and the most extreme values of a and c. Results for all conditions are available in the Supplementary Material.

EMD and Iterative Forward

The iterative forward approach behaved similarly to the EMD approach with respect to all outcome variables, though yielding half as many anchor items as the EMD approach (see Figures 1–3). This is due to the stopping criterion of the approach, which is designed in such a way that not all items detected as DIF-free are included in the anchor.

Effect of DIF size, balancedness, and number of DIF-free items. As expected, we found that in the EMD and the iterative forward approaches DIF size, balancedness, and number of DIF-free items

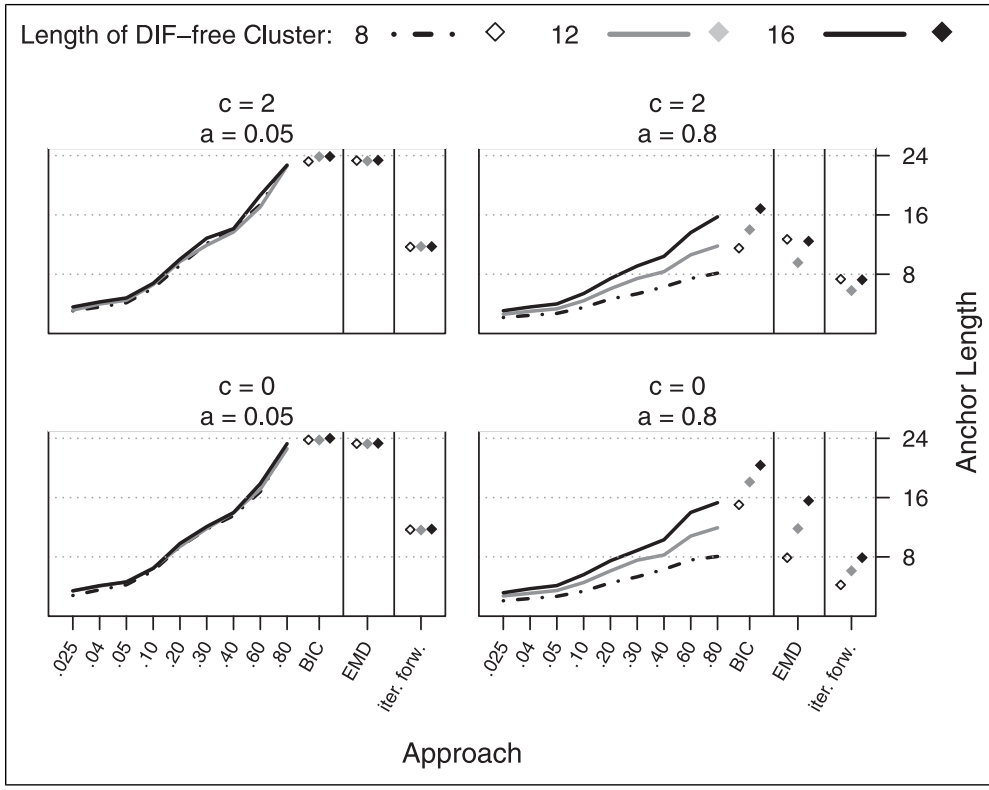

Figure 1. Anchor length with *N* = 500 and no missing values. Threshold sizes on *x*-axis. BIC = Clustering using the BIC to decide on cluster count. *Note*. EMD = Equal-mean-difficulties approach; Iter. forw. = Iterative forward approach; BIC = Bayesian information criterion.

had large effects on both hit rate (partial eta squared being 0.444/0.142 for DIF size, 0.488/0.315 for balancedness, and 0.729/0.481 for number of DIF-free items for the EMD/iterative forward approach) and mean bias (partial eta squared being 0.593/0.322 for DIF size, 0.536/0.344 for balancedness, and 0.124/0.091 for number of DIF-free items for the EMD/iterative forward approach). There was no bias in either of the methods with regard to balanced DIF and small DIF. Both methods displayed large bias with regard to unbalanced large DIF (up to more than 0.6 *SD*s). With an increase in DIF size, an increase in unbalancedness, and a decrease in the number of DIF-free items, hit rate decreased and mean bias increased (see main effects in Table S2 and the Figures in the Supplementary Material).

Interaction of DIF size, balancedness, and number of DIF-free items. There were noticeable two-way and three-way interaction effects of DIF size, balancedness, and number of DIF-free items on both hit rate and mean bias (see Table S2 in the Supplementary Material for ANOVA results). The effect of number of DIF-free items on hit rate and mean bias increased with unbalancedness of DIF. This effect increased even more with an increase in size of DIF, such that the lowest performance was found in conditions with low number of DIF-free items, large unbalancedness, and large DIF size.

Effect of sample size and missing values. An increase in sample size amplified the effects of unbalancedness and number of DIF-free items (see Figure S2, S8, S3, and Figure S9 in the

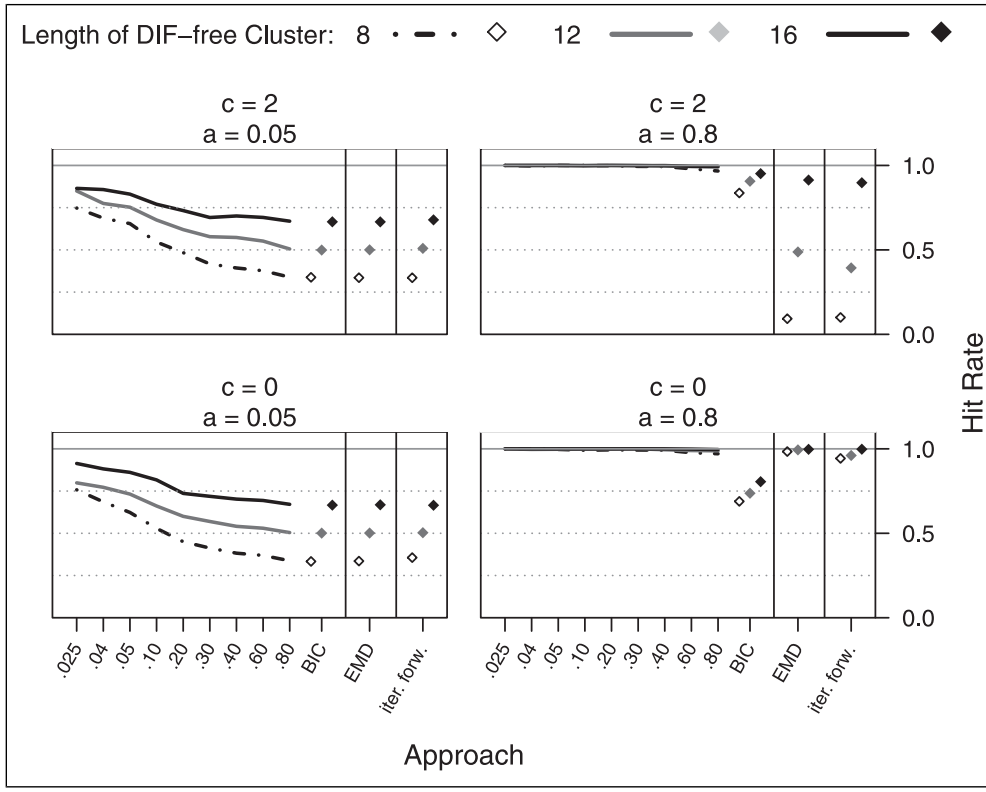

Figure 2. Hit rate with *N* = 500 and no missing values. Threshold sizes on *x*-axis. *Note*. BIC = Clustering using the BIC to decide on cluster count. EMD = Equal-mean-difficulties approach; Iter. forw. = Iterative forward approach; BIC = Bayesian information criterion.

Supplementary Material), that is, hit rate decreased and mean bias increased in conditions with unbalanced DIF and low number of DIF-free items. As expected, for both approaches, the effect of missing values strongly depended on the type of items in which the missing values occurred. When missing values occurred in DIF items (see Supplementary Material Figure S2 and Figure S29), hit rate decreased with missing rate, while it increased when missing values occurred in DIF-free items (see Supplementary Material Figure S2 and Figure S38). Missing values had little to no effect on mean bias for both approaches, regardless on whether the missing occured on DIF or DIF-free items.

Extended cluster approach

The extended cluster approach with BIC resulted in less accurate results than the cluster approach with any threshold settings in almost all conditions (see Figures 1–3). In most conditions, the BIC criterion returns all items as a single cluster (see Supplementary Material). As such, we do not consider it any further.

In almost all conditions, the cluster approach with any of the cluster selection settings resulted in a hit rate of the best cluster at least as high as for the other two approaches. Furthermore, the bias related to using the best cluster in the cluster approach was lower than of any of the other approaches in all conditions. The varied factors had no or only a small impact on mean bias (partial eta squared <0.005).

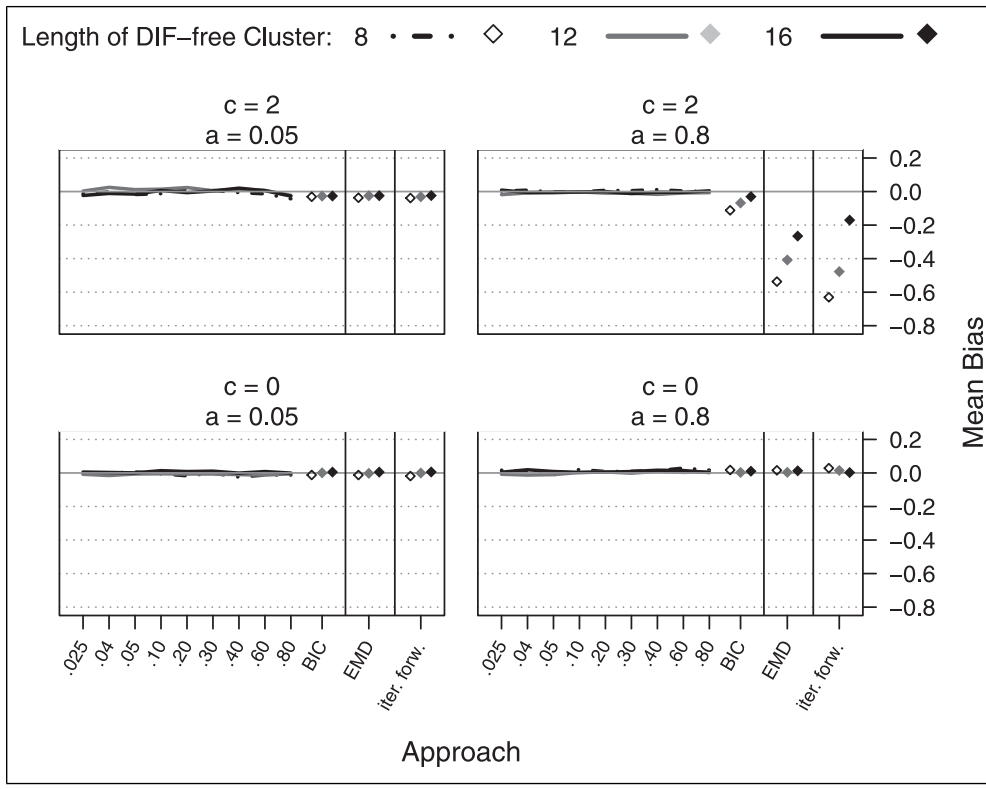

Figure 3. Bias in estimated mean difference with $N = 500$ and no missing values. Threshold sizes on *x*-axis. *Note*. BIC = Clustering using the BIC to decide on cluster count. EMD = Equal-mean-difficulties approach; Iter. forw. = Iterative forward approach; BIC = Bayesian information criterion.

Effect of DIF size, threshold, and number of DIF-free items. The extended cluster approach produced unbiased mean difference estimates in almost all conditions. Deviations from the true mean difference were only found in conditions with large DIF and at the same time threshold settings that were larger than the true DIF. The bias can be avoided by using threshold settings which are smaller than the true DIF. As expected, anchor length increased with increasing threshold, thus, resulting in the most optimal results regarding accuracy and low link error (i.e., through a large number of DIF items) when thresholds were chosen that are about the size of the true DIF.

Hit rate showed main effects as well as second order interactions of DIF size, threshold settings, and number of DIF-free items. Larger hit rate was achieved with larger number of DIF-free items and low threshold settings. Small thresholds resulted in small anchor item sets (see Figure 1). Specifically, in contrary to the other approaches, the cluster approach was affected by neither unbalancedness nor small number of DIF-free items.

Effect of sample size and missing values. An increase in sample size slightly amplified the effect of the other factors (see Supplementary Material Figure S2, S8, S3 and S9 for an illustration). In most conditions, hit rate slightly decreased with an increase in missing values, regardless of the type of items in which the missing values occurred (see Supplementary Material Figure S2, Figure S29, and Figure S38).

Empirical Study: Illustrating the Impact of Choice of Assumption

In order to illustrate the impact of using the different approaches in practice, we applied them to data from the reading competence assessment (Gehrer et al., 2013) in the National Educational Panel Study (NEPS), a longitudinal large scale educational study in Germany (Blossfeld, Roßbach, & von Maurice 2011). Specifically, we selected data from a linking study that aimed to link the assessment of grade 9 students to the adult sample (Pohl & Carstensen, 2013; Pohl et al., 2015).

We analyzed responses to items designed for grade 9 that were taken by the grade 9 main sample (G9; $n = 13{,}897$) and the adult link sample (AD; $n = 501$). We selected the 27 dichotomous items from the ninth grade test for our investigation. Pohl et al. (2015) found a considerable amount of DIF among the groups (ranging from −1.3 to 1.3 logits) when using a model that constrained the mean item difficulty to be equal across groups. Note that, these results do not tell us which items exhibit DIF, rather they only tell us that the items do not function similarly. In order to facilitate linking the two groups, we aim at identifying a set of anchor items that may be assumed to be measurement invariant.

Methods

We followed the analyses of Pohl et al. (2017). To identify anchor items, the same three approaches as above were utilized. We used the threshold criterion and set the maximum heterogeneity within a cluster to 0.6 logits when determining the number of clusters in the cluster approach–a size of DIF that was considered acceptable in the NEPS (Pohl & Carstensen, 2013). As a comparison, we also used the BIC as selection criterion for the cluster approach. To estimate the group mean difference, we set the item difficulties of the respective anchor item set or cluster to be equal across the two groups. The mean reading competence in the group of grade 9 students was fixed to zero, while the mean in the group of adults was estimated. The R code for the cluster approach can be found in the Supplementary Material.

Results

Figure 4(a) shows the DRID for the empirical data (Pohl et al., 2017). The EMD approach identified four items that are assumed to be DIF-free (black lines in Figure 4(b)). Using these as anchor items resulted in an estimated mean difference between grade 9 students and adults of 0.29 logits. The iterative forward approach identified 10 items that are assumed to be DIF-free (black solid and dashed lines in 4C), of which five were chosen for the anchor (solid black lines). Using the anchor items for linking resulted in an estimated group mean difference of 0.36 logits. The cluster approach with BIC resulted in one cluster that included all items. This is in line with the simulation results, that showed that the BIC is hardly able to distinguish between different clusters. We therefore did not consider this solution any further.

The cluster approach with threshold setting resulted in a five-cluster solution (see Figure 4(d)). The clusters consisted of two to 11 items, depicted by different colors and line types. The maximum difference in DRID between any two items within a cluster was 0.6, as determined by the cluster selection criterion. The estimated mean differences in the latent reading ability between the ninth graders and the adults differed considerably depending on the cluster chosen for anchoring: It was −0.85 logits for the first cluster (grey solid line); −0.25 logits for the second cluster (black solid line); 0.38 logits for the third cluster (grey dashed line); 0.88 logits for the fourth cluster (black dashed line); and 1.56 logits for the fifth cluster (grey dashed dotted line). The estimated mean difference varied between −0.85 logits (Cohen's $d = -0.57$) and 1.56 logits

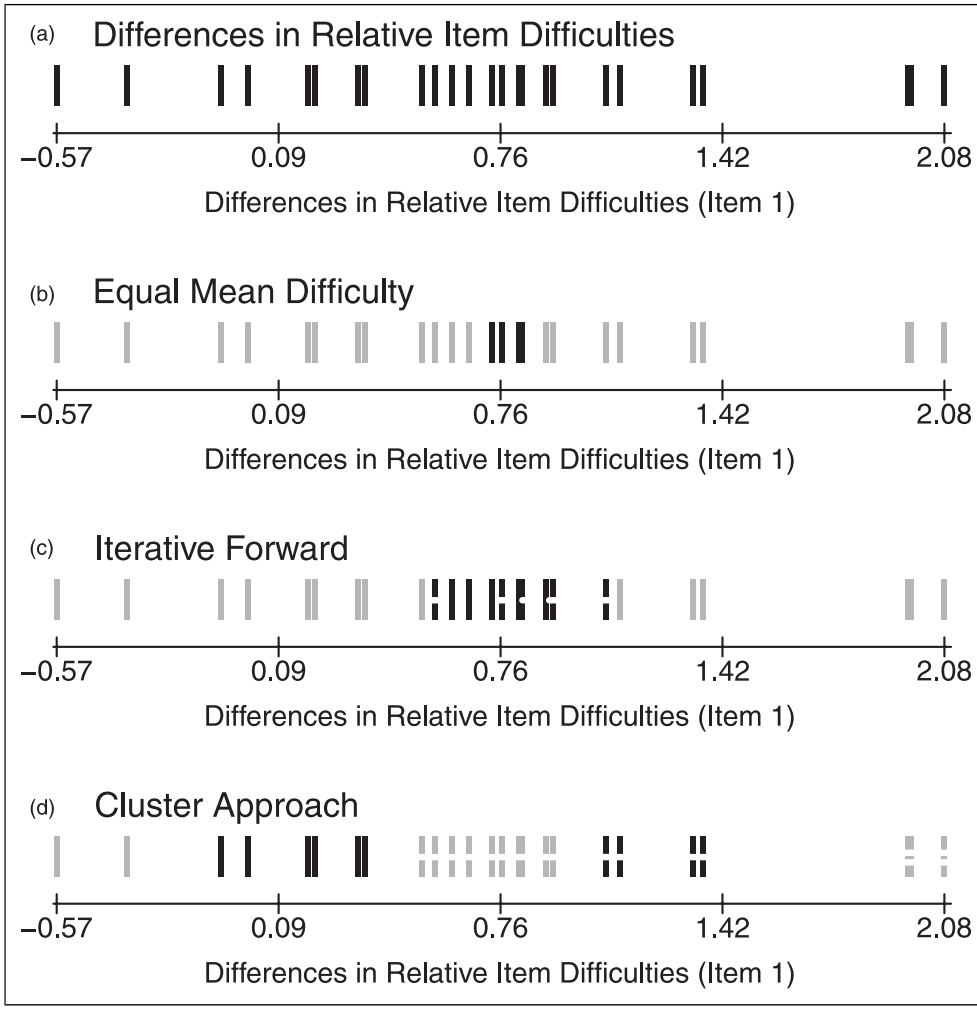

Figure 4. DRIDs of all items relative to item 1 (a) and anchors identified in the empirical data example when using the three approaches (b to d). In b and c, black solid lined items correspond to the anchor, black dashed items are further items found to be DIF-free. In d items with the same symbol belong to the same cluster.

(d = 1.05); this is not only a considerable difference in size but would also lead to a complete reversal of the conclusions drawn. Using one of the first two clusters for linking resulted in a higher estimated mean reading ability in adults, while choosing cluster 3, 4, or 5 suggested that grade 9 students have on average a higher reading competence than adults. From the data we cannot infer which of the clusters, if any, represents DIF-free items.

Researchers may now (a) choose the cluster for linking based on the content of the items. If there is no cluster to be favored from a substantive point of view, (b) researchers may decide which assumptions to make. Note that, one may apply similar assumptions as in the EMD or as in the iterative forward approach for choosing clusters within the cluster approach. If one assumes unbalanced DIF, one may use the by item-number weighted average of the estimated mean differences across all clusters (being 0.35 logits). One may also assume that the largest group of

items is DIF-free and chose the largest cluster (here cluster 3). Importantly, in this example, the largest cluster contains similar items as the anchor item set of the iterative forward approach, resulting in a similar estimated mean difference. Thus, the cluster approach allows to incorporate a variety of different assumptions. In addition, the cluster approach also identifies other possible item sets that may also be considered as possible anchor item sets. If researchers cannot plausibly make any assumptions, (c) they can use each (or some) of the clusters for linking and depict the uncertainty in the results.

Discussion

In this paper, we extended the approach of Bechger and Maris (2015) with the aim of making it accessible to substantive researchers. An advantage of our procedure is that it provides communicable criteria for the identification of clusters. We showed that BIC is less suitable for determining the number of clusters identified in the cluster approach and suggest setting thresholds instead. When setting a threshold, researchers will need to consider accuracy and number of items in the link, that is, size of linking error. Accuracy is guaranteed for thresholds similar or lower than the true DIF. A low linking error results for larger anchor item sets. As such, optimal results regarding both, accuracy and linking error can be achieved, when the threshold is set slightly lower than the true DIF. When working with real data, we suggest to choose the maximum amount of DIF one is willing to tolerate for setting the threshold. These thresholds may be informed by benchmarks of DIF that are commonly used in large scale assessments (e.g., OECD, 2017).

This paper provides the first comprehensive results on the performance of the cluster approach when applied to Rasch models. In contrast to the EMD and iterative forward approaches, the cluster approach always provided a cluster that resulted in unbiased mean differences in all conditions. Previous approaches performed well, when either the amount of DIF was small or DIF was balanced, they failed to identify the anchor items and to recover the true mean difference when DIF was unbalanced and large. Although the identification of the correct anchor items became less accurate (i.e., showed lower hit rates) when the amount of DIF was small, bias in mean differences was negligible. This was true for all approaches.

The cluster approach's ability to yield a cluster that results in unbiased mean differences, comes at the price of having to choose between different anchor sets. In practice, we do not know which of the identified clusters represents DIF-free items or whether there are DIF-free items at all. One might argue that the cluster approach merely shifts the methodological issue from finding DIF-free items to making a decision about *which* cluster represents a set of DIF-free items. The approach has, however, distinct advantages. First, the cluster approach does not presume that there are any DIF-free items. Second, in contrast to previous approaches, even in challenging conditions (i.e., large unbalanced DIF), the cluster approach results in unbiased results when choosing the right cluster. Third, using the cluster approach allows researchers to imply a variety of different assumptions. Researchers may still assume that the largest cluster is DIF-free (similar to what is proposed by the iterative forward approach). They may also assume that DIF is balanced (as in the EMD approach) and use all item clusters for linking. Conceptually, the cluster approach is flexible in terms of its ability to incorporate different assumptions. More importantly, the cluster approach enables researchers to base their choice on other criteria, such as content of the items. Regardless of the procedure used, researchers need to make their assumptions explicit. If applied researchers cannot tell from the item features which cluster to choose for linking, they can still present the results using each of the extracted clusters, thus depicting the uncertainty that arises from the choice of the anchor set. Fourth, the cluster approach has favorable statistical properties. While an increase in sample size and a decrease in missing values does not always increase the performance

of the EMD or iterative forward approach, it does enhance the performance of the cluster approach in all conditions.

Our results are in line with prior analyses of the EMD (W.-C. Wang, 2004) and iterative forward (Kopf et al., 2015b) approaches. Our study extends previous work on these approaches by (a) including conditions that violate the assumptions underlying them, (b) comparing several aspects of their performance, (c) evaluating mean bias, and (d) evaluating the statistical properties of the approaches in the presence of missing values. We showed that the EMD is to some extent robust to violations of balancedness of DIF when the relative number of DIF-free items is large. We also showed that the iterative forward approach does not necessarily need to rely on the assumption that the majority of items is DIF-free, but also works when a minority of items are DIF-free and DIF is balanced. Furthermore, we showed that even for low hit rates, in some conditions mean bias may still be negligible. Additionally, we showed that the EMD and the iterative forward approaches are suitable for the same conditions. The two approaches differ in the length of the anchor as well as sensitivity to sample size and missing values.

In practice, researchers often do not have viable reasons for making well-informed assumptions before running the analyses. As they implemented (and possibly developed) the scale, they usually assume measurement invariance for all items. If they had reasons to assume otherwise, they would probably have altered the scale before the assessment. In fact, although evaluating measurement invariance is nowadays state of the art and applied in many analyses, in their papers, substantive researchers hardly mention or discuss the plausibility of the assumptions they make (Doebler, 2019; Schroeders & Gnambs, 2018). Instead, in practice, researchers often run different models and decide upon anchor items based on statistical and theoretical reasons after analyses have been performed. One may argue that by encouraging examination of different assumptions, the cluster approach may motivate HARKing, that is, hypothesizing after the results are known (Kerr, 1998). However, this can also happen when using any other approach. The cluster approach enhances making the decisions transparent.

Researchers may also want to ascertain a broad content coverage and, thus, make use of more than one cluster for linking. An extension of this approach to Bayesian model averaging (Schulze et al., 2021) allows the researcher to use information from all clusters and at the same time quantifies the uncertainty in cluster selection in the posterior distribution. It is also possible to incorporate prior knowledge in aggregating the results, reflecting partial knowledge, or believes about plausibility of anchor item sets.

One difference between the EMD and iterative forward approaches as compared to the cluster approach is that the former rely on parameters of inference statistics for item selection. Thus, the results depend on the standard error (i.e., missing values and test targeting). This dependence is not always favorable as items that are badly targeted to the population and have many missing values are more likely get selected in the anchor set. As the cluster approach only relies on effect estimates, it is less affected by these factors. However, the drawback of this is that it does not take the uncertainty of parameter estimation into account. In future research, one may combine the advantages of the different anchor item selection strategies.

In the proposed approach group, comparisons are made based on a chosen set of anchor items. Strobl et al. (2021) propose an alternative approach, which is based on a similar principle as the scale alignment in SEM (Muthén & Asparouhov, 2014). Their approach may also identify local optima, which corresponds in principle to the cluster approach. However, instead of identifying different homogenous sets of items, the algorithm provides a set of different anchor points.

In this paper, we focused on uniform DIF, that is on item difficulties. An extension of this approach for nonuniform DIF, that is for simultaneously considering both item difficulty and item discrimination, has been presented by Pohl and Schulze (2020). The proposed approach is also limited to dichotomous covariates (e.g., comparison between two time points or between two

groups). An extension of the approach to categorical covariates with more than two levels (e.g., school form) or continuous covariates (e.g., age) has been presented by Schulze and Pohl (2021a). For the other-item approaches, Huelmann et al. (2020) propose aggregation rules that allow to extend the approaches to the multi-group case.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback and the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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 iD

Steffi Pohl https://orcid.org/0000-0002-5178-8171

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Note that, due to scale indeterminacy, the setting of the true means will hardly impact the results. It will only be relevant for standard errors of item parameters, which will be larger with worse test targeting.
2. Kopf et al. (2015a) argue that the item with the median value must be a DIF-free item because they assume that the majority of items (i.e., more than 50%) are DIF-free.
3. The rate of DIF-free items not included in the anchor can be computed based on the information on test length, the number of truly DIF-free items and hit rate. Note, however, that this is less a problem than including DIF-free items in an anchor, as unbiased estimates may be achieved also with small anchors.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*: American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing.

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

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

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. https://doi.org/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. https://doi.org/10.18637/jss.v048.i06

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement*, *43*(4), 303–321. https://doi.org/10.1177/0146621618795727

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization (ETS Research Report No. RR-92-10)* (*1992*, pp. i-40). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*(2), 50–79.

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

Huelmann, T., Debelak, R., & Strobl, C. (2020). A comparison of aggregation rules for selecting anchor items in multigroup DIF analysis. *Journal of Educational Measurement*, *57*(2), 185–215. https://doi.org/10.1111/jedm.12246

Kerr, N. L. (1998). HARKING: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/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. https://doi.org/10.1177/0146621614544195

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. https://doi.org/10.1177/0013164494054001003

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

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). Lawrence Earlbaum Associates.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. 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. https://doi.org/10.3758/BRM.42.3.847

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2017). Dexter: R Manual. https://CRAN.R-project.org/package=dexter

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

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

OECD (2017). *PISA 2015 technical report*: OECD.

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. https://doi.org/10.1177/014662169001400205

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National educational panel study–many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*(2), 189–216.

Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). Measuring competencies across the lifespan - challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Dependent data in social sciences research: Forms, issues, and methods of analysis* (pp. 281–308). Springer. https://doi.org/10.1007/978-3-319-20585-4_12

Pohl, S., & Schulze, D. (2020). Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. *Psychological Test Assessment and Modelling*, *2*(62), 281–303.

Pohl, S., Stets, E., & Carstensen, C. H. (2017). *Cluster-based anchor item identification and selection (NEPS Working Paper No. 68)*: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

R Development Core Team. (2008). *R: A language and environment for statistical computing*.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Institute of Educational Research.

Schroeders, U., & Gnambs, T. (2020). Degrees of freedom in multigroup confirmatory factor analyses: Are models of measurement invariance testing correctly specified? *European Journal of Psychological Assessment*, *36*(1), 105–113. https://doi.org/10.1027/1015-5759/a000500

Schulze, D., Reuter, B., & Pohl, S. (2021). *Approaching partial measurement invariance via bayesian model averaging*. Manuscript submitted for publication.

Schulze, D., & Pohl, S. (2021a). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 219–228. https://doi.org/10.1080/10705511.2020.1771186

Schulze, D., & Pohl, S. (2021b). *measurementInvariance*. https://github.com/Dani-Schulze/measurementInvariance.

Strobl, C., Kopf, J., Kohler, L., Oertzen, T. v., & Zeileis, A. (2021). Anchor point selection: An approach for anchoring without anchor items. *Applied Psychological Measurement*, *45*(45), 214–230. https://doi.org/10.1177/0146621621990743

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. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic pProgramming. *The R Journal*, *3*(2), 29–33. https://doi.org/10.32614/RJ-2011-015

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