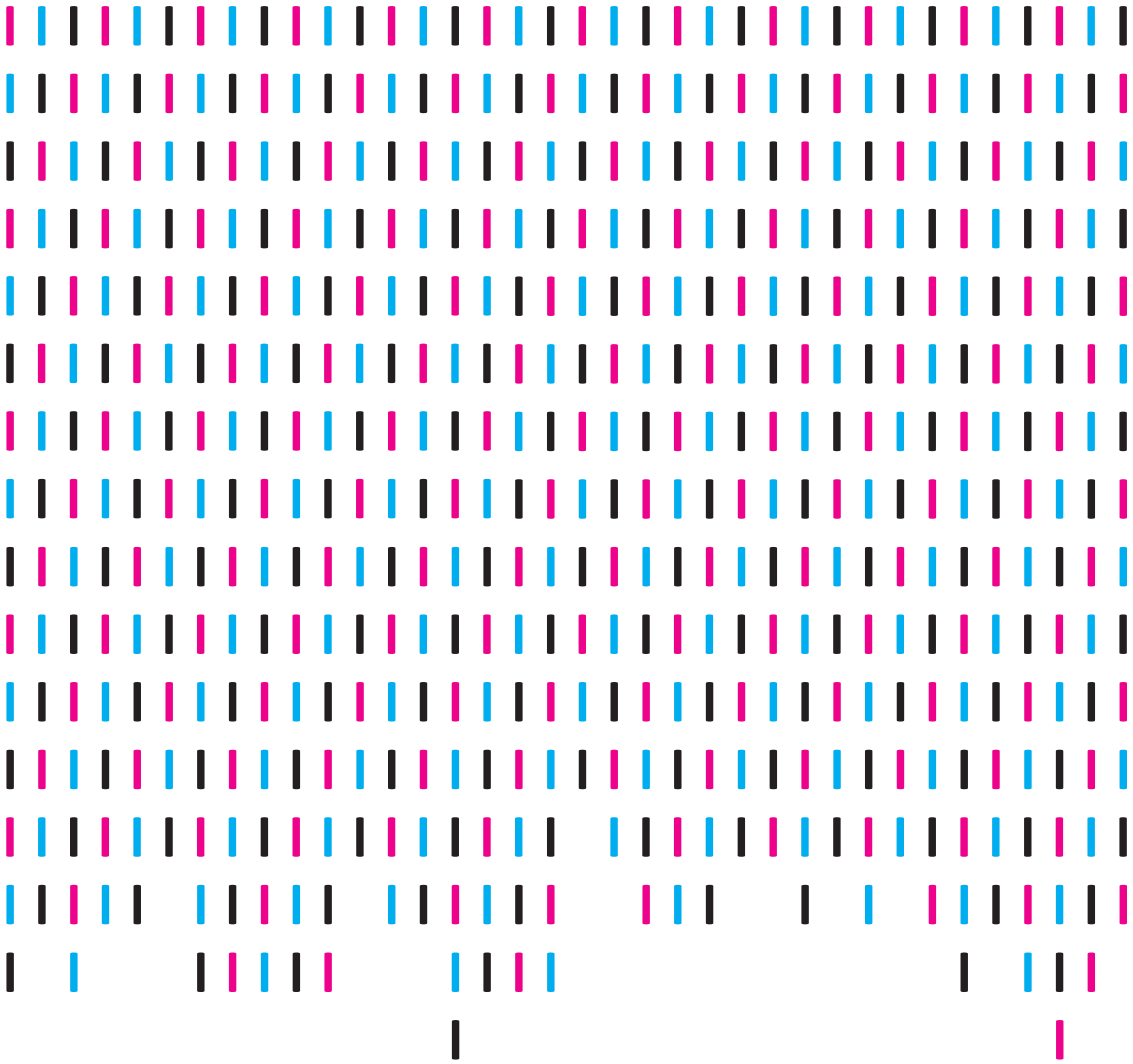

NEPS WORKING PAPERS

Steffi Pohl, Eric Stets, and Claus H. Carstensen

CLUSTER-BASED ANCHOR ITEM IDENTIFICATION AND SELECTION

NEPS Working Paper No. 68
Bamberg, March 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).

The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at **www.neps-data.de** (see section "Publications").

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

Cluster-based Anchor Item Identification and Selection

Steffi Pohl & Eric Stets, Freie Universität Berlin
Claus H. Carstensen, Leibniz Insitute for Educational Trajectories

E-mail address of lead author:

steffi.pohl@fu-berlin.de

Bibliographic data:

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

Cluster-based Anchor Item Identification and Selection

Abstract
In order to compare scores of latent variables across groups or measurement occasions, the respective items presented to both groups or at both measurement occasions need to be measurement invariant, that is, show no differential item functioning (DIF). In situations where this assumption is violated, researcher may strive for partial measurement invariance by identifying a set of items (anchor items) that are DIF-free. Different approaches for detecting DIF-free items exists. These either make the assumption of unbalanced DIF or the assumption that the majority of items is DIF-free. Recently, Bechger and Maris (2015) proposed an approach that instead of identifying DIF-free items identifies clusters of items that function similarly. As such they do not make the assumption of unbalanced DIF or that the majority of items is DIF-free. While this approach is very promising, it is not applicable, yet, for substantive research. 1. There is no clear criterion for the identification of clusters. 2. There are no criteria for choosing a cluster as anchor for linking purposes. (a) We propose two procedures for cluster identification, that are, the k-means clustering approach and the range-and-step-threshold approach. (b) We provide three selection criteria (cluster homogeneity, cluster accuracy, and cluster size) that may aid the choice of a cluster. For illustration, we apply the approach on data of a linking study in the National Educational Panel Study comparing reading competence between grade 9 students and adults. The paper closes with a discussion of the advantages as well as the limitations of the proposed methods and a delineation of further research areas.

Keywords
Item response theory, differential item functioning, partial measurement invariance, anchor items, clustering

1 Introduction

Many educational large-scale assessments as for example the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011) or the Programme for International Student Assessment (e.g., OECD (2012) aim at investigating differences in competence scores across groups or time.

In order to compare groups (or constructs across time), the measures need to be on a common scale. Within an item-response theory (IRT) framework, various methods can be used to achieve this (e.g., von Davier & von Davier, 2007). Particularly, an often employed design is the so called anchor-item design, where respondents of two non-equivalent groups are presented with the same items (e.g. Kolen & Brennan, 2004). It is assumed that the items presented to both groups are measurement invariant across groups, that is they show no differential item functioning (DIF). One way to analyze these data is to assume that some or all items have equal item parameters across groups, which forces the scales to be equal. These items are termed anchor items. If the invariance assumption for anchor items does not hold, the group parameter estimates will be biased and comparisons of groups become an artifact of the scaling procedure (e.g., Borsboom, 2006; Navas-Ara & Gómez-Benito, 2002; W.-C. Wang, 2004).

1.1 DIF-detection methods and their assumptions

Many DIF detection procedures have been developed (see, e.g., Magis, Béland, Tuerlinckx, & De Boeck, 2010 or Kopf, Zeileis, & Strobl, 2015a) but no consensus on a gold-standard has been found. Reasons for this can be found in the assumptions that are made by the procedures. One typically employed assumption is that the mean difficulty of the items is equal in both groups (equal-mean difficulty, EMD). Underlying this procedure is the assumption of balanced DIF. This means that on average the items do not favor one group over the other and the aggregated DIF effect over all items cancels out. However, this assumption has been considered unlikely and in circumstances where it is violated, biased group comparisons will emerge (e.g., W.-C. Wang, 2004).

In the presence of DIF in some items, partial measurement invariance may be an option, where only some of the items are selected as anchor items. This reduces the assumption of all common items showing no DIF to the assumption of only a set of items showing no DIF. The task of identifying these potential anchor items has sprouted development of a variety of procedures. One way of identifying potential anchor items is to test each item for DIF when constraining all other items to have equal item parameters across groups (e.g., Cohen, Kim, & Wollack, 1996). This approach has been shown to display increased type 1 error rates under conditions of unbalanced DIF, that is when the aggregated DIF over all items favors one group over another (Kopf et al., 2015a; W.-C. Wang, Shih, & Sun, 2012).

There are approaches that do not rely on the assumption of unbalanced DIF. One recently developed approach was put forward by Kopf, Zeileis, and Strobl (2015b) and is based on the work of W.-C. Wang (2004) and Shih and Wang (2009). It integrates rank-based item selection (Woods, 2009) and an iterative procedure (e.g., Candell & Drasgow, 1988; W.-C. Wang et al.,

2012) to build a set of anchor items. Different from W.-C. Wang et al. (2012), Kopf et al. (2015b) did not use a purification procedure, but rather iteratively built up the anchor. In simulation studies the methods show a good performance compared to previous methods (Kopf et al., 2015b; W.-C. Wang, 2004). However, its implicitly underlying assumption is – as the authors state – that the majority of items is DIF-free. This assumption does not need to hold in applications.

1.2 A different view on DIF: The cluster-based approach

Bechger and Maris (2015) claim that in empirical analyses DIF cannot be identified for the item, but only the relative position of two items can be identified. Instead of identifying DIF-free items, they propose an approach that identifies clusters of items that function similarly. As such they do not claim that *a specific item* shows DIF or not; they also do not claim that they can *identify DIF-free items*. In their approach no assumption of unbalanced DIF is made nor that the largest group of items is DIF-free. In fact, their approach does not even claim that there is a DIF-free item at all. In the following we will illustrate this approach. Note that this approach is the basis of the research presented in this paper. We will follow up on the idea of the cluster based approach and present ways in which this approach may be used for comparing groups or measurement occasions.

While most of the previous research aim at identifying *DIF an of item*, Bechger and Maris (2015) delineate 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 in the scale identification issue. Lets 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 multigroup design. The likelihood of the two-group model is then given by

$$P(\mathbf{X}=\mathbf{x}|\vartheta,\beta)=\prod_{p=1}^{n_1}\prod_{i=1}^{k}\frac{exp[x_{pi}((\vartheta_p+a)-(\beta_{i,1}+a))]}{1+exp[x_{pi}((\vartheta_p+a)-(\beta_{i,1}+a))]}+\prod_{q=n_1+1}^{n}\prod_{i=1}^{k}\frac{exp[x_{pi}((\vartheta_q+b)-(\beta_{i,2}+b))]}{1+exp[x_{pi}((\vartheta_q+b)-(\beta_{i,2}+b))]}$$

with $\mathbf{X}$ denoting the item response matrix, ϑ the latent ability, and $\beta_{i,g}$ the difficulty of item i in group g. All together there are k items in the test that are answered by n persons with group-wise sample sizes of n_1 in group one and $n-n_1$ in group two. a and b denote additive constants in group 1 and group 2, respectively.

In multiple-group Rasch-models, the model is identified up to an additive constant. This constant (here a and b) does not need to be the same across groups. As such two identification restrictions are needed and by adopting a restriction, the relative position of the scales can be shifted without changing the likelihood of the model. Hence, only the difference in item difficulties between items (relative item difficulties) are identified within a group. Due to these identification issue, *differences in item difficulties between groups* cannot meaningfully be interpreted. However, the *difference in relative item difficulties between the groups* is identified and can be interpreted. Bechger and Maris (2015) construct a matrix of these differences in

relative item differences across groups which they term ΔR-matrix. This matrix is given by

$$R_{ij}^{(g)} = \beta_{i,g} - \beta_{j,g} \tag{1}$$

$$\Delta R = R^{(1)} - R^{(2)}, \tag{2}$$

with i and j indicating items and g indicating the group. An example of the ΔR-matrix with artificial data of seven items is shown in Figure 1. The entries in the first column of Figure 1 depict the differences in relative item difficulties (DRID) across groups relative to item 1. The DRID for item 2 as compared to item 1 in this example is $\Delta R_{12} = R_{12}^{(1)} - R_{12}^{(2)} = (\beta_{i,1} - \beta_{j,1}) - (\beta_{i,2} - \beta_{j,2}) = -.08$. Thus, the difference in difficulties (DRID) between item 1 and 2 is 0.08 logits larger in group 2 as compared to in group 1.

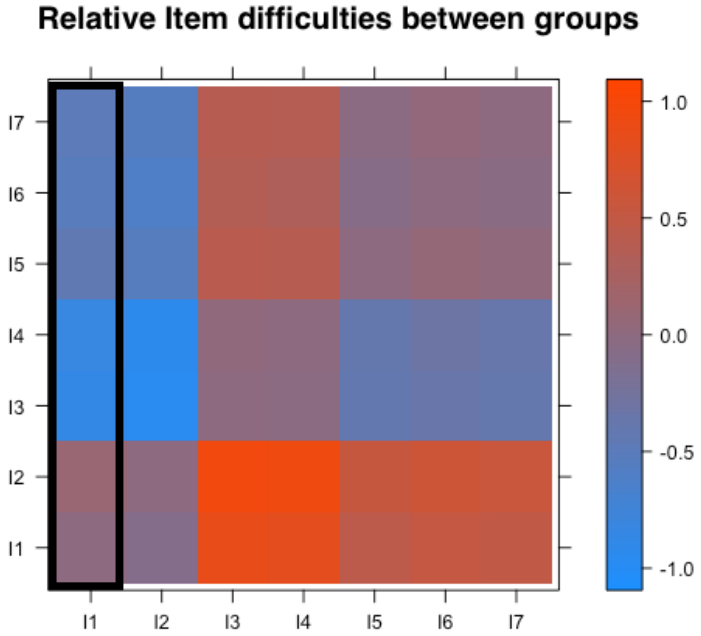

Figure 1: The ΔR-matrix showing the differences in relative difficulties (DRID) for seven example items.

In the next step clusters of invariant items need to be identified in such a way that items within a cluster are invariant in their item difficulties relative to the other items in the cluster. In the example of simulated data in Figure 1, three clusters of invariant items can be determined: cluster 1 consisting of items 1 and 2, cluster 2 consisting of items 3 and 4, and cluster 3 consisting of items 5, 6, and 7. Within the clusters there is hardly any DRID and the difference in item difficulties is rather homogeneous across the items in a cluster. Having identified clusters, Bechger and Maris (2015) construct a significance test to test sub-matrices of the ΔR-matrix belonging to one cluster for invariance. While for these simulated data the clusters can easily be identified, this is usually not the case in empirical data. An example of a ΔR-matrix from an empirical example using 27 items is shown in Figure 2 (The example is described in more detail later). In this example clusters may not easily be visually identified. For applying the cluster-based approach, a way to identify the clusters need to be developed. Having identified clusters, researcher need to choose one cluster for linking purposes.

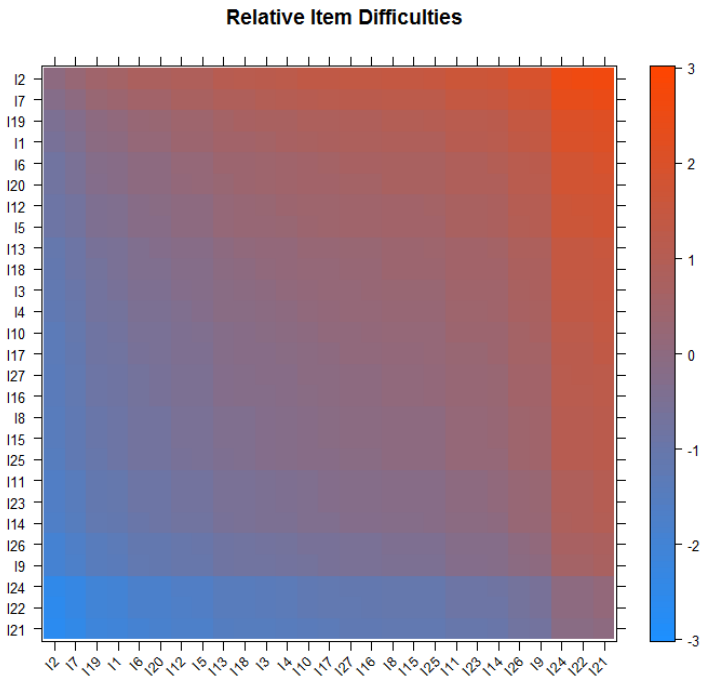

Figure 2: The ΔR-matrix for an empirical study.

1.3 Research Questions

Bechger and Maris (2015) have laid the groundwork for a promising approach to the detection of invariant item clusters. The authors made clear that DIF-free items cannot be identified from the data, only sets of invariant items can. While the approach is very promising, it cannot be readily applied by substantive researchers, yet. Specifically there is no clear criterion for the identification of clusters, but the identification of clusters is still based on visual inspection of the Δ*R*-matrix. Further, when multiple clusters are identified, it is not clear what criteria can be used by the researcher to choose one of the clusters as anchor items for linking purposes.

We approach these two open questions by (a) proposing two procedures for the identification of clusters of invariant items and (b) offering selection criteria that can guide the researcher in the process of selecting a cluster. Additionally, we illustrate the procedures on NEPS reading competency data, comparing grade 9 students and adults. Finally, we discuss our results and outline further ideas.

2 Procedures for cluster identification and selection

In the following we will not use the whole Δ*R*-matrix, but make use of the fact that the Δ*R*-matrix is skew-symmetric and of rank 1. Thus, all information is contained in a single row or column of the matrix. We arbitrarily select the first column of the matrix and further work with one dimensional data without loss of information. Figure 3 shows the DRID, that is the entries of the Δ*R*-matrix, comparing the difference in difficulties of all item to item 1 across the two

groups. Consequently the DRID of item 1 is zero in this Figure.

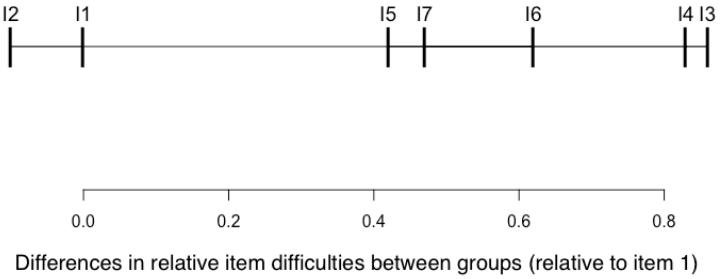

Figure 3: DRID values that are taken from the first column of the ΔR-matrix in Figure 1.

2.1 Cluster Identification

Items form invariant clusters of items, when their difference in relative difficulties between the groups are close together. Because under conditions of real data items are never absolutely DIF-free, the problem of how to identify these clusters arises. We propose two methods to do this: a) a optimal k-means clustering procedure for unidimensional data and b) a range threshold procedure based on sorted data.

2.1.1 Optimal k-means clustering

The Δ*R*-matrix is a distance matrix. While this already allows the application of a wealth of clustering algorithms, we selected a variant of the k-means algorithms that works on a single row/column of the matrix: optimal k-means clustering by dynamic programming (H. Wang & Song, 2011). The advantage lies in its optimality in one dimension, where it is guaranteed to find a global optimum, which is not the case for multidimensional k-means clustering. In its current implementation, the algorithm minimizes the squared sum of errors of items to the cluster center and selects the number of clusters based on BIC.

2.1.2 Range thresholds

Since the data are sortable, a straightforward idea is to define range threshold criteria for the length of a cluster. Such length can be set by the researcher depending, for example, on criteria for maximum tolerated DIF within a set of anchor items. Range thresholds may, for example, be based on DIF-criteria of the educational testing service (Zwick, 2012) or those in NEPS (Pohl & Carstensen, 2012). We propose a method to identify clusters of items that adhere to a given threshold for cluster length. This approach is graphically depicted in Figure 4. First, in order to make sure that items, which show invariance relative to many other items in a cluster, are grouped together, the item with the highest density of DRID values is picked as a starting point (see Figure 4 a)). The density is determined by a Gaussian kernel estimate of density. The item with the highest density is assigned to the first cluster. Subsequently, the item closest to the left or right end of the cluster is selected and it is checked whether its inclusion in the cluster

would exceed the threshold length of the cluster (see Figure 4 b)). If it does not, the item is added to the cluster and the items closest to the so found new cluster are checked. Figure 4 b) shows an excerpt of the DRID values in Figure 4 a). After the first item is chosen, eight further items are checked for inclusion. The arrows show the items that are closest to the right and left of the cluster, respectively. The item pointed at with the dark color is the closest one and checked for inclusion. This procedure is repeated until no further item can be added to the first cluster without breaking the thresholds. In the example in 4 b) this is the case after eight items have been included. If no item can be added to the cluster without exceeding the threshold, a new cluster is built. The starting point for the next cluster is selected by checking for the area of the highest density of DRID-values while excluding any items that have already been added to a cluster (see Figure 4 c)). Items are added to the next cluster in the same manner as described above. The procedure continues until all items are added to a cluster. Figure 4 d) shows the clusters identified for the example. There are four clusters, which are depicted by different symbols. It should be noted, that due to starting a new cluster based on the density, this form of clustering usually results in a large first cluster. This may be desirable as we aim at maximizing the homogeneity of the clusters.

2.2 Cluster Selection

After clusters of invariant items have been identified, the researcher needs to select one for linking purposes. One way of choosing a cluster is by expert judgement based on the item content. However, this is not always possible. And even if it is possible, additional criteria may aid the selection of a cluster. We propose to incorporated further information in the choice of a cluster. We propose to use the following quanitities as additional source of information for judging the clusters:

1. The *size of the cluster* in terms of the number of items that it contains is an important criterion because the stability of the link depends on the number of items in it. Given that the link items are truly DIF-free, sampling error in item parameter estimates will have less impact when using many anchor items.

2. *Cluster homogeneity* refers to the amount of DRID within a cluster. As usually the items within a cluster do not function perfectly the same, there is still some DRID left. A researcher might require the items that are used for linking to be as homogeneous as possible. This could be quantified by the within-cluster sum-of-squares.

$$h_c = \sqrt{\frac{\sum_{i \in c} (DRID_i - \bar{c})^2}{n_c}} \tag{3}$$

with

$$\bar{c} = \frac{\sum_{i \in c} DRID_i}{n_c} \tag{4}$$

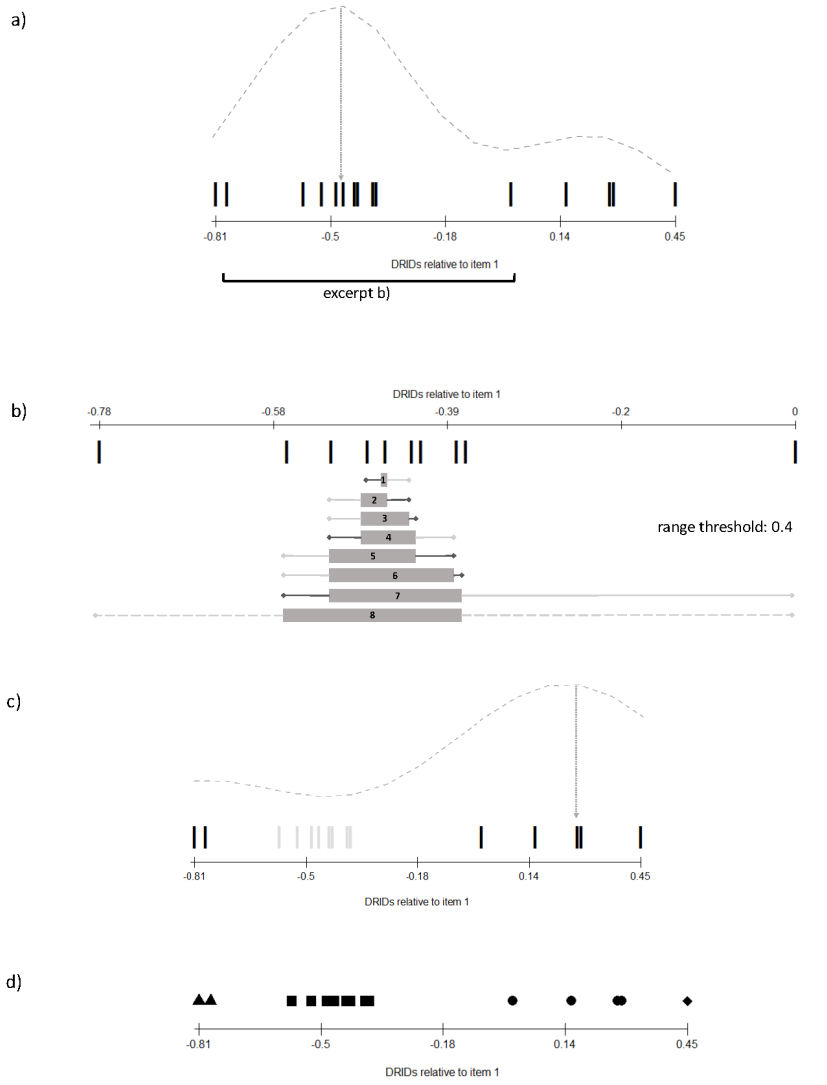

Figure 4: Illustration of the approach for identification of clusters using range thresholds.

where $DRID_i$ denotes the differences in relative item difficulty for item *i* and n_c denotes the number of items in cluster c. This measure is akin to the linking error (Monseur & Berezner, 2007). Lower values indicate higher homogeneity.

3. Finally, *precision of item parameter estimation* (i.e., item parameter standard errors) might provide important information when selecting a cluster of anchor items. The quality of the link will depend on the efficiency of the item parameter estimates. The efficiency of the items will depend on the test targeting as well as the amount of missing values. The precision of items may even differ between the groups. This is especially the case when the group means differ and as such the test targeting is different in each group. In these cases an item may have a low standard error in one group (because it is well targeted) but not in the other one. For linking purposes we will most often be interested in a low standard error (i.e. high precision) of item parameters in *both* groups. As such, we propose to use the size of the standard errors of the item parameter estimation of the items within a cluster in both groups as a criterion. In order to investigate this, the

precision of a cluster is evaluated for each group by averaging over the standard errors of the items within the cluster and converting the aggregated standard error in a precision measure, i.e.

$$prec_{gc} = \sqrt{n_c^{-1} \sum_{i \in c} \hat{s}_{ig}^{-2}}. \tag{5}$$

n_c^{-1} denotes the number of items in cluster c. s_{ig}^2 denotes the standard error for item i in group g, and c denotes a set of items belonging to a cluster. A high value indicates a high precision.

Generally, we cannot tell from the data which cluster is the correct one or if a DIF-free cluster even exists. However, the proposed criteria might either aid the decision and provide further information regarding the cluster chosen.

3 Empirical Study

We apply the cluster identification and selection procedures to empirical data from the NEPS reading competency assessment. The NEPS is a large-scale educational study that implements a multi-cohort longitudinal design to investigate competence development across the whole life span (Blossfeld et al., 2011). Specifically, we select data for reading competency (Gehrer, Zimmermann, Artelt, & Weinert, 2013) and link grade 9 students to an adult sample (Pohl & Carstensen, 2013; Pohl, Haberkorn, & Carstensen, 2015).

3.1 Design

To assess reading competency in grade 9 and in the adult population, separate tests are presented to each sample with no common items across the two groups. For illustration, we selected the 27 dichotomous items from the 9th grade test for our investigation. In order to allow for comparison of reading competence across cohorts, some form of linking is needed. To achieve a link between the grade 9 sample and the adult sample, the NEPS employs a link sample design to achieve overlap (Pohl et al., 2015). This design entails a separate sample of participants that was drawn from the adult population. This link sample is presented with the items of both, the grade 9 test and the adult test. Figure 5 illustrates the design. In this analysis we focus on the link between the grade 9 main sample (G9; $n = 13897$) and the link sample (AD; $n = 501$). As such we will only use the responses to the Grade 9 test items of these two groups.

Pohl et al. (2015) investigated the assumptions for linking in this design. They found that the items of both tests are unidimensional within the link sample. However, they also found a considerable amount of DIF between the groups when using a model that constrained the mean item difficulties to be equal. Figure 6 depicts the estimated DIF. Although theoretically all items

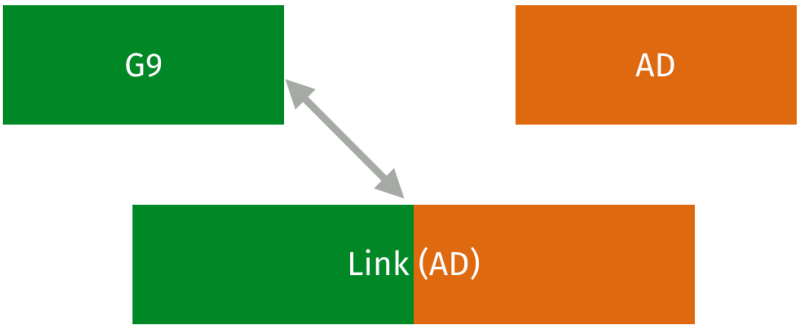

Figure 5: Link sample design of the NEPS reading competency assessment. A grade 9 main sample (G9) takes the grade 9 test (green) and an adult main sample (AD) take the adult test (orange). An additional link sample (of adults) takes both tests and is used for linking. Of concern in this example is the link between the grade 9 main sample and the link sample (grey arrow).

can be used for linking both scales, this result suggests a partial measurement invariance approach in which only a selection of items are used as anchors for linking.

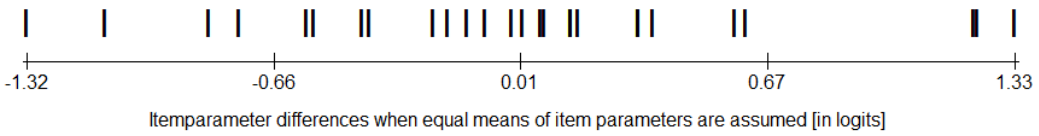

Figure 6: DIF of item difficulties between the grade 9 main sample and the adult link sample when equal mean difficulty is assumed between the groups.

3.2 Methods

We apply the cluster-based detection approaches outlined above to the data of the grade 9 sample and the link sample. Rasch-models were estimated under the marginal-maximum likelihood framework using the R software package (R Development Core Team, 2008) and the mirt-package version 1.17.1 (Chalmers, 2012). The latent distribution was assumed to be normal.

The optimal k-means algorithm was implemented using the R-package Ckmeans.1d.dp (H. Wang & Song, 2011). The appropriate number of clusters was selected by means of BIC. For the range threshold procedure, we selected 0.6 logits as range limit. The choice was made based on the criteria for judging DIF in the NEPS (Pohl & Carstensen, 2013).

Cluster selection criteria were calculated for each cluster identified. These include 1) the size of the cluster, 2) the homogeneity of the cluster (Formula 3), and 3) the precision for each cluster within groups (Formula 5).

3.3 Results

When applying the cluster-based approach to the NEPS data, the DRID-matrix in Figure 2 results. In contrast to the artifical example in Figure 1, in the empirical data there is no obvious clustering of items. Thus, there is a need for a procedure to identify clusters.

3.3.1 Cluster Identification

Based on BIC the optimal k-means clustering algorithm identifies only a single cluster with all items. When finding only a single cluster, generally, no cluster selection is necessary and the researcher might use all items for linking. However, this would result in rather invariant groups of items. We expect this to be a result of the BIC as selection criterion. The BIC may not necessarily be the most appropriate criterion and we will discuss alternative ones in the discussion.

The range threshold procedure detects five clusters (see Figure 7). The cluster consist of two to eleven items. Within a cluster, the DRID is rather homogeneous. We used range thresholds based on criteria of the NEPS. Other thresholds would have detected different clusters and it is up to the researcher to choose appropriate ones for the concrete case.

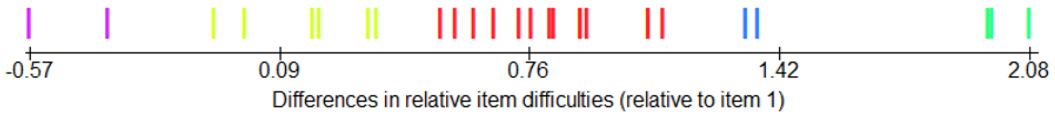

Figure 7: Relative differences in item difficulty for the grade 9 main and adult link studies. Colors illustrate the clusters.

3.3.2 Cluster Selection

As for the k-means solution only one cluster was identified, no selection is necessary. Therefore we do not further discuss selection criteria for this solution but for the cluster solution of the range threshold method. Using the range threshold criterion five clusters were identified. In order to link the grade 9 main sample to the adult link sample, one cluster needs to be chosen as anchor. One way of choosing a cluster is by experts judgement. We present some further

statistics that may either aid the selection process and/or provide further information on a selected cluster.

Table 1 shows the selection criteria for the five clusters. The cluster with the largest number of items is Cluster 3. This cluster may result in a rather stable link as compared to clusters with lower number of items (given the standard errors are the same). Cluster 4 and 5 are the most homogeneous ones, that is, showing the lowest amount of DRID. However, this comes at the cost of a rather small number of items. Note that homogeneity corresponds to some degree with the size of a cluster. Smaller clusters will more often be preferred as they span a lower range of DRID values. Regarding cluster precision, items in cluster 3 and 5 show a high precision in both groups. The items in these clusters are well targeted to the majority of persons in both groups and, thus, display low standard errors. As items of median difficulties will most often be the ones estimated with the highest precision, these are items that may be preferred for linking. Note that the presented criteria cannot solve the problem that we do not know whether there is a DIF-free cluster or which cluster is DIF-free. However, they may aid the selection by providing further information and may also be used for describing the selected cluster.

Depending on the clusters chosen for anchoring, the mean differences of the latent ability between the grade 9 main sample and the adult link sample varies between 0.85 logits (0.57 SD) and -1.56 logits (-1.05 SD) (see table 1). This is not only a considerable amount but also a complete reversal of the conclusions drawn. Using one of the first two clusters for linking would result in a mean reading ability being estimated larger in adults, while when choosing cluster 3, 4, or 5 we would conclude that grade 9 students have on average a higher reading competence than adults.

Table 1: Cluster selection criteria for the five clusters identified in the NEPS reading competency data. The column "mean differences in logits" shows the estimated mean differences of the latent ability, when the cluster was selected as anchor for linking.

Cluster	Size	Homogeneity	Precision G9	Precision adults	Mean difference in logits (G9-AD)
1 (purple)	2	0.022	4.88	2.36	0.85
2 (yellow)	6	0.150	5.94	2.53	0.25
3 (red)	14	0.466	6.32	2.48	-0.38
4 (blue)	2	0.000	5.77	2.31	-0.88
5 (green)	3	0.007	7.07	2.45	-1.56

Outlook and Further Ideas

The cluster-based identification of invariant items (Bechger & Maris, 2015) is a promising alternative to traditional anchor item selection procedures as it only relies on what can be identified from the model and as such explicates at which stage we make assumptions. It furthermore provides different solutions (clusters) of anchor item sets. This way it not relying on the assumption that the largest set of invariant items is DIF-free and also showing in which way results may be influenced by the choice of the anchor set. In order to enhance the applicability of the approach, we proposed an extension regarding cluster identification and cluster selection.

We presented two extensions for identifying clusters as well as criteria that might aid in the selection of a cluster. An advantage of our procedure is that it provides criteria for the identification and the selection of a cluster. These criteria are communicable to other researchers and make the assumptions made clear. The procedures work without making the assumption of balanced DIF or that the largest cluster is DIF-free. However, this comes at the expense of an additional cluster selection step. We illustrated the procedures using empirical data from the NEPS and show that the cluster selection choice is an important one with large implications for conclusions about group-level differences.

Applying the k-means clustering approach using BIC for cluster selection seems to be a rather conservative choice concerning the number of clusters chosen. As in our empirical case and some test cases we ran, only one cluster is identified. However, there are many other possibilities for the selection of *k* and the cost function. In further research these other options (e.g. k-medoids or silhouette methods) may be investigated in the light of the task at hand.

In the future, we would like to investigate the procedures more thoroughly using a simulation study. Also, we will investigate whether it is possible to integrate certain selection criteria (e.g. the cluster size) into the k-means clustering procedure (Nielsen & Nock, 2014). This would allow to unify the cluster identification and selection step to a certain extend and may optimize clusters.

A selection of a cluster may not always be feasible as no cluster may have enough support to be chosen as anchor. If a cluster cannot be selected as anchor based on substantive arguments, a research may link the data using each of the clusters (as was done here). This helps in determining the uncertainty that arises with the choice of a cluster. We would also like to resolve the model selection problem that is posed by the different clusters by means of Bayesian model averaging. This would allow to use all clusters and at the same time to quantify of the uncertainty in cluster selection in the posterior distribution.

Acknowledgements This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 – 5th Grade, doi:10.5157/NEPS:SC3:3.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

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

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (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*(Suppl 3), S176–S181. 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.

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

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.

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

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

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3).

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*(1), 9.

Nielsen, F., & Nock, R. (2014). *Optimal interval clustering: Application to Bregman clustering and statistical mixture learning.*

OECD - Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report*. Paris, France: OECD Publishing.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working paper No. 14)*. Bamberg: National Educational Panel Study.

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/Journal für Bildungsforschung 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.* New York: Springer.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Shih, C., & Wang, W. (2009). Differential item functioning detection using the multiple indi-

cators, multiple causes mimic method with a pure short anchor. *Applied Psychological Measurement*, *33*, 184–199.

von Davier, M., & von Davier, A. A. (2007). A Unified Approach to IRT Scale Linking and Scale Transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(3), 115–124. doi: 10.1027/1614-2241.3.3.115

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.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-Free-Then-DIF Strategy for the Assessment of Differential Item Functioning. *Educational and Psychological Measurement*, 0013164411426157. doi: 10.1177/0013164411426157

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

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30.