LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Julia Kopf, Achim Zeileis, Carolin Strobl

Anchor methods for DIF detection: A comparison of the iterative forward, backward, constant and all-other anchor class

Technical Report Number 141, 2013
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Anchor methods for DIF detection: A comparison of the iterative forward, backward, constant and all-other anchor class

Julia Kopf
Ludwig-Maximilians-Universität München

Achim Zeileis
Universität Innsbruck

Carolin Strobl
Universität Zürich

Abstract

In the analysis of differential item functioning (DIF) using item response theory (IRT), a common metric is necessary to compare item parameters between groups of test-takers. In the Rasch model, the same restriction is placed on the item parameters in each group in order to define a common metric. However, the question how the items in the restriction – termed anchor items – are selected appropriately is still a major challenge. This article proposes a conceptual framework for categorizing anchor methods: The *anchor class* to describe characteristics of the anchor methods and the *anchor selection strategy* to guide how the anchor items are determined. Furthermore, a new anchor class termed the *iterative forward* anchor class is proposed. Several anchor classes are implemented with two different anchor selection strategies (the all-other and the single-anchor selection strategy) and are compared in an extensive simulation study. The results show that the newly proposed anchor class combined with the single-anchor selection strategy is superior in situations where no prior knowledge about the direction of DIF is available. Moreover, it is shown that the proportion of DIF items in the anchor – rather than the fact whether the anchor includes DIF items at all (termed *contamination* in previous studies) – is crucial for suitable DIF analysis.

Keywords: Rasch model, anchoring, anchor selection, contamination, item response theory (IRT), differential item functioning (DIF), DIF analysis, item bias.

1. Introduction

The analysis of differential item functioning (DIF) in item response theory (IRT) research investigates the violation of the invariant measurement property among subgroups of examinees, such as male and female test-takers. Invariant item parameters are necessary to assess ability differences between groups in an objective, fair way. If the invariance assumption is violated, different item characteristic curves occur in subgroups. In this paper, we focus on *uniform* DIF where one group has a higher probability of solving an item (given the latent trait) over the entire latent continuum and the group differences in the logit remain constant (Mellenbergh 1982; Swaminathan and Rogers 1990).

A variety of testing procedures for DIF on the item-level is available (e.g., Lord 1980; Mellenbergh 1982; Holland and Thayer 1988; Thissen, Steinberg, and Wainer 1988; Swaminathan and Rogers 1990; Shealy and Stout 1993, for an overview see Millsap and Everson 1993). In the analysis of DIF using IRT, item parameters are to be compared across groups. Mostly, research focuses on the comparison of two pre-defined groups, the reference and the focal group. Thus, a common scale for both groups is required to assess meaningful differences in the item parameters. The minimum (necessary but not sufficient) requirement for the construction of a common scale in the Rasch model is to place the same restriction on the item parameters in both groups (Glas and Verhelst 1995). The items included in the restriction are termed *anchor items*.

The anchor method determines how many items are used as anchor items and how they are located. Consistent with the literature, we use the term *locate* as a synonym for selecting anchor items. The choice of the anchor items has a high impact on the results of the DIF analysis: If the anchor includes one or more items with DIF, the anchor is referred to as *contaminated*. In this case, the scales may be biased and items that are truly free of DIF may appear to have DIF. Therefore, the false alarm rate may be seriously inflated – in the worst case all DIF-free items seem to display DIF (Wang 2004) – and the results of the DIF analysis are doubtful, as various examples demonstrate (see Section 2).

Even though the importance of the anchor method is undeniable, Lopez Rivas, Stark, and Chernyshenko (2009, p. 252) claim that "[a]*t this point, little evidence is available to guide applied researchers through the process of choosing anchor items*". Consequently, the aim of this article is to provide guidelines how to choose an appropriate anchor for DIF analysis in the Rasch model.

In the interest of clarity, we introduce a new conceptual framework that distinguishes between the *anchor class* and the *anchor selection strategy*. Firstly, the *anchor class* describes the pre-specification of the anchor characteristics (such as a pre-defined anchor length). We review the all-other (used, e.g., by Cohen, Kim, and Wollack 1996; Kim and Cohen 1998), the constant anchor (used, e.g., by Thissen *et al.* 1988; Wang 2004; Shih and Wang 2009) and the iterative anchor class (referred to here as iterative backward and used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002). Furthermore, we introduce a new anchor class named the iterative forward anchor class. Secondly, the *anchor selection strategy* determines which items are chosen as anchor items. We discuss the all-other (introduced as rank-based strategy by Woods 2009) and the single-anchor selection strategy (proposed by Wang 2004). The complete procedure to choose the anchor is then called *anchor method*.

To derive guidelines which anchor method is appropriate for DIF detection in the Rasch model, we conduct an extensive simulation study. In our study, we compare the all-other, the constant, the iterative backward and the newly suggested iterative forward anchor class for

the first time. Furthermore, our study is to our knowledge the first to systematically contrast different anchor selection strategies that are combined with the anchor classes mentioned above. Altogether, nine different anchor methods are evaluated regarding their appropriateness for DIF analysis. Finally, practical recommendations are given to facilitate the process of selecting anchor items for DIF analysis in the Rasch model.

In the next section, technical details of the anchor process for the Rasch model are explained and illustrated by means of an instructive example and the framework of anchor classes, anchor selection strategies and anchor methods is introduced in detail. The simulation study is presented in Section 3 and the results are discussed in Section 4. The problem of contamination and its impact on DIF analysis are addressed in Section 5. Characteristics of the selected anchor items are discussed in Section 6. A concluding summary of the simulation results and practical recommendations are given in Section 7.

2. A conceptual framework for anchor methods

In the following, the anchor process is technically described and analyzed for the Rasch model. The item parameter vector is $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$, where k denotes the number of items in the test. In the following, it is estimated using the conditional maximum likelihood (CML) estimation due to its unique statistical properties, its widespread application (Wang 2004) and the fact that its estimation process does not rely on the person parameters (Molenaar 1995).

As the origin of the scale in the Rasch model can be arbitrarily chosen (Fischer 1995) – what is often referred to as *scale indeterminacy* – one linear restriction of the form

$$\sum_{\ell=1}^{k} a_\ell \tilde{\beta}_\ell = 0\,, \tag{1}$$

with constants a_ℓ holding $\sum_{\ell=1}^{k} a_\ell \neq 0$ is placed on the item parameter estimates $\tilde{\beta}_\ell$ (Eggen and Verhelst 2006). Thus, in the Rasch model only $k-1$ parameters are free to vary and one parameter is determined by the restriction. Note that equation 1 includes various commonly used restrictions such as setting one item parameter $\tilde{\beta}_\ell = 0$ or restricting all item parameters to sum zero $\sum_{\ell=1}^{k} \tilde{\beta}_\ell = 0$ (Eggen and Verhelst 2006). Without loss of generality, we here estimate the item parameter vector $\tilde{\beta}$ with the employed restriction $\tilde{\beta}_1 = 0$. The corresponding covariance matrix $\widehat{\mathrm{Var}}(\tilde{\beta})$ then contains zero entries in the first row and in the first column. It is irrelevant which particular restriction is chosen to identify the parameters, as the interpretation of the results is independent of the restriction employed. In the following, we discuss different restrictions for which the sum of a selection of items is set to zero and a is an indicator vector. These restrictions can be obtained by transformation using the equations

$$\hat{\beta} = A\tilde{\beta} \tag{2}$$

$$\text{and} \quad \widehat{\mathrm{Var}}(\hat{\beta}) = A\widehat{\mathrm{Var}}(\tilde{\beta})A^\top, \tag{3}$$

where $A = I_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot a^\top$, I_k denotes the identity matrix, 1_k denotes a vector of one entries and a is a vector with one entries for those elements a_ℓ that are included in the restriction and zero entries otherwise (e.g., $a = (1, 0, 1, 0, 0, \ldots)^\top$ including item 1 and item 3). Additionally, the entries of the rank deficient covariance matrix $\widehat{\mathrm{Var}}(\hat{\beta})$ in the row and in the column of the item that is first included in the restriction are set to zero.

While for the estimation itself, the choice of the restriction is arbitrary, for the anchor process a careful consideration of the linear restriction that is now employed in each group g is necessary. A necessary but not sufficient requirement in order to build a common scale for two groups is that the same restriction is employed in both groups (Glas and Verhelst 1995). Items in the restriction are termed *anchor items* and the restriction can be rewritten as

$$\sum_{\ell=1}^{k} a_\ell \hat{\beta}_\ell^g = \sum_{\ell \in \mathcal{A}} \hat{\beta}_\ell^g = 0, \tag{4}$$

where the set $\mathcal{A}$ is termed the *set of anchor items* or the *anchor*. The estimated and anchored item parameters are denoted $\hat{\beta}^g$. Equation 4 includes various commonly used anchor methods such as setting one item parameter $\hat{\beta}_\ell^g$ to zero ($\hat{\beta}_\ell^g = 0$, for one $\ell \in \{1, 2, \ldots, k\}$) for the so called constant single-anchor method or restricting all items except the studied item j to sum zero in each group ($\sum_{\ell \neq j} \hat{\beta}_\ell^g = 0$) for the so called all-other anchor method. The item parameters, estimated separately in each group, are transformed to the respective anchor method by using equation 2, where all items are then shifted on the scale by $-\frac{1}{|\mathcal{A}|} \sum_{\ell \in \mathcal{A}} \tilde{\beta}_\ell^g$. The covariance matrices are transformed using equation 3.

Furthermore, the anchor items should be DIF-free. Otherwise, if for example the single anchor item that is restricted to zero in both groups is a DIF item, the construction of a common scale fails as all other items are artificially shifted apart as will be shown in the following example (for a detailed illustration see also Wang 2004). This is termed artificial DIF by Andrich and Hagquist (2012). Unfortunately, since prior to DIF analysis it cannot be known which items are DIF-free, we face a somewhat circular problem, as pointed out by Shih and Wang (2009).

The construction of a common scale is now illustrated by means of an instructive example. The data set from a general knowledge quiz was conducted by the weekly German news magazine SPIEGEL in 2009. A thorough discussion and analysis of the original data set are provided in Trepte and Verbeet (2010) including a global DIF analysis by means of model-based recursive partitioning by Strobl, Kopf, and Zeileis (2010). From about 700,000 test-takers that answered each a total of 45 items from different domains, we select a subsample of $9,442$ test-takers (that obtained their A-levels in Germany) and four items from politics (listed below together with the correct answers) for the illustration of the anchor problem:

Item 1	Who determines the rules of action in politics according to the German Constitution? (The Bundeskanzler.)
Item 2	What is the role of the second vote in the elections for the German Bundestag? (It governs the seating in the German Bundestag.)
Item 3	How many people were killed by the RAF? (33)
Item 4	Indicate the location of Hessen on the German map.

As an exemplary illustration, let us suppose we want to test for DIF in the first item between the focal (foc) group of the test-takers that obtained their A-levels in the German federal state Hessen and the reference (ref) group of all remaining test-takers.

Figure 1 displays three different restrictions: The second item as constant single-anchor, the fourth item as constant single-anchor and all other items (item 2 to item 4) as anchor. The points represent the estimated item parameters from the reference (light points) and the focal group (dark points). The rectangles surround the anchor item(s).

In Figure 1 (left), item 2 is used as constant single-anchor and, thus, both estimated item parameters are set to zero. The results here display the estimated item parameters. The negligible difference in the item parameters of item 1, that we are currently interested in,

suggests no DIF in this item. To understand the DIF test results for item 1 in the next scenarios, it is also important to note that the large difference in item 4 implies DIF in this item. Since item 4 was the question to indicate the location of Hessen on the German map, it is plausible that this item 4 is a true DIF item since it was easier for test-takers that obtained their A-levels in Hessen. The item-wise Wald test (see Lord 1980 and equation 5 below) for item 1 does not display statistically significant DIF ($t = -.968$ with the corresponding p-value of .333). As a result, item 1 is classified as DIF-free.

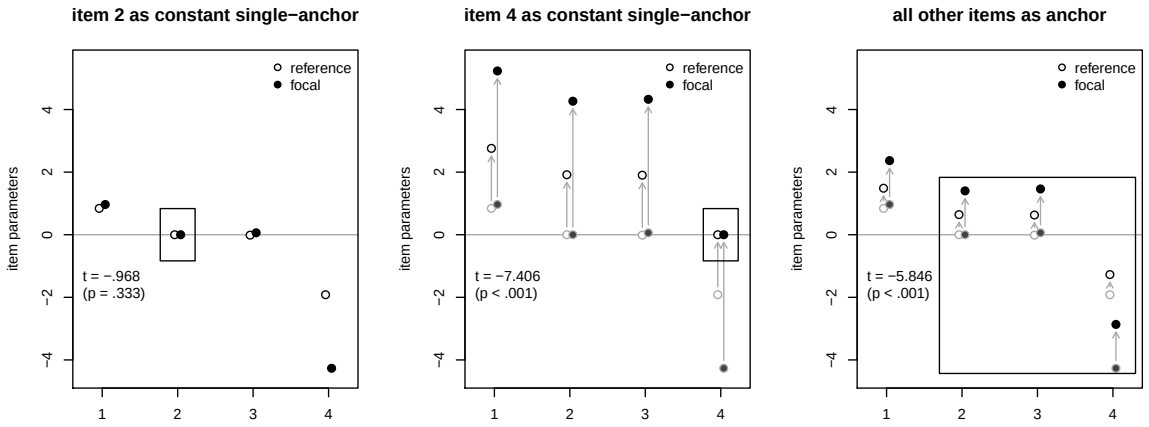

Figure 1: Different restrictions placed on the item parameters that are estimated using the Rasch model in each group.

In the next scenario in Figure 1 (middle), the (suspected) DIF item 4 is used as a constant single-anchor. Compared to the first scenario, the item parameters are now shifted upwards by the estimated difficulties of item 4 (approximately 2 for the reference and 4 for the focal group as indicated by the arrows). The difference that now occurs for item 1, that we are currently interested in, is likely to reflect the presence of artificial DIF. Here, the statistical test indicates DIF for item 1 ($t = -7.406$ with the corresponding p-value $< .001$). Hence, item 1 is classified as a DIF item.

In the last scenario in Figure 1 (right), all other items – except the currently studied item 1 – are used as anchor items. Compared to the first scenario where item 2 was set to zero, all items are shifted upwards by the average over the estimated difficulties of item 2, 3 and 4. Now, all item parameters are shifted apart. Compared to the second scenario, the scales are shifted apart less strongly since the scale shift is reduced from the estimated difficulties of the suspected DIF item 4 to the average over the estimated difficulties of item 2, 3 and 4 (including the presumed DIF-free items 2 and 3) as visible by the shorter arrows. However, the statistical test still classifies item 1 as a DIF item ($t = -5.846$ with the corresponding p-value $< .001$). This example illustrates the major impact of the anchor method on the results of the DIF analysis, since – depending on the anchor set – three different test statistics result in the DIF tests for item 1.

From a theoretical perspective and from our instructive example, it is obvious that an appropriate anchor is crucial for the results of the DIF analysis. Previous simulation studies have compared different selections of anchor methods (an overview over the existing literature will be given in Section 3.4). Empirical findings also show that, ideally, the anchor items should

be DIF-free. Otherwise, the *contamination* may lead to seriously augmented false alarm rates in DIF detection (see, e.g., Wang and Yeh 2003; Wang 2004; Wang and Su 2004; Finch 2005; Stark, Chernyshenko, and Drasgow 2006; Woods 2009) that "*can result in the inefficient use of testing resources, and [...] may interfere with the study of the underlying causes of DIF*" (Jodoin and Gierl 2001, p. 329). Naturally, the risk of contamination would suggest to use only few items in the restriction (i.e. a short anchor), but the simulation results also show that the statistical power increases with the length of a DIF-free anchor (Thissen *et al.* 1988; Wang and Yeh 2003; Wang 2004; Shih and Wang 2009; Woods 2009).

Therefore, in the following this trade-off between the false alarm rate (DIF-free items that are erroneously diagnosed with DIF) and the hit rate (DIF items that are correctly diagnosed with DIF) is analyzed for different anchor methods in a simulation study. As a statistical test for DIF we will focus on the item-wise Wald test here (Lord 1980). For the Wald test, the item parameters are first estimated separately in each group. The anchor methods are then used to obtain a common scale for the item parameters which are to be compared.

The rationale behind the Wald test is that DIF is present if the item difficulties are not equal across groups. The test statistic T_j for the null hypothesis $\mathrm{H}_0 : \beta_j^{\mathrm{ref}} = \beta_j^{\mathrm{foc}}$, where β_j^{ref} and β_j^{foc} denote the item difficulties for reference and focal group for item j and $\hat{\beta}_j^{\mathrm{ref}}$ and $\hat{\beta}_j^{\mathrm{foc}}$ the corresponding estimated item parameters using the anchor $\mathcal{A}^j$, has the following form:

$$T_j = \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}})}} = \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{ref}})_{j,j} + \widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{foc}})_{j,j}}}\,. \tag{5}$$

Note that, the estimated and anchored item parameters $\hat{\beta} = \hat{\beta}(\mathcal{A}^j)$, which can be calculated using equation 2, depend on the anchor and, hence, so does the test statistic $T_j = T_j(\mathcal{A}^j)$. The anchor set $\mathcal{A}^j$ may depend on the studied item (as is the case for the all-other method). If the anchor is constant regardless which item is tested for DIF, it is denoted $\mathcal{A}$ in the following.

2.1. Anchor classes

In our conceptual framework *anchor classes* describe characteristics of the anchor that answer the following questions: Is the anchor length pre-defined? If so, how many items are included in the anchor? Is the anchor determined by the anchor class itself or is an additional anchor selection strategy necessary? Are iterative steps intended to define the anchor?

In the *equal-mean-difficulty* anchor class (see, e.g., Wang 2004, and the references therein) all items are restricted to have the same mean difficulty (typically zero) in both groups, whereas in the *all-other* anchor class the sum of all items – except the item currently tested for DIF – is restricted to be zero. Both anchor classes have a pre-defined anchor length but no additional anchor selection is necessary as the items included in the restriction are already determined by the anchor class itself. The equal-mean-difficulty and the all-other class only differ in one anchor item and, therefore, essentially lead to similar results (cf. Wang 2004). Since it seems reasonable not to include the item currently tested for DIF in the anchor, only the all-other method is included in the following simulation study.

The *constant* anchor class includes a pre-defined number of the test items (e.g., 1 or 4 items according to Thissen *et al.* 1988) or a certain proportion of the test items (e.g., 10% or 20% according to Woods 2009) as anchor. The term *constant* reflects the pre-defined, constant anchor length. In our simulation study, we implemented the constant anchor class with one single anchor item as well as the constant anchor including four items, which is supposed

to assure sufficient power (cf. e.g., Shih and Wang 2009; Wang, Shih, and Sun 2012). The constant anchor class needs to be combined with an explicit anchor selection strategy.

The *iterative backward* anchor class includes a variety of iterative methods that have been suggested, discussed and combined with different statistical methods to assess DIF. Here, we focus on the commonly used re-linking procedure where one parameter estimation step suffices to conduct DIF analysis. Firstly, the scales of both groups are linked on (approximately) the same metric, e.g., by using the all-other anchor method. Then, the DIF items are excluded from the current anchor, the scales are re-linked using the new current anchor, the DIF analysis is carried out and the steps are repeated until two steps reach the same results (e.g., Drasgow 1987; Candell and Drasgow 1988; Park and Lautenschlager 1990; Kim and Cohen 1995; Hidalgo-Montesinos and Lopez-Pina 2002). This iterative procedure is referred to here as the *iterative backward* anchor class, since the method includes the majority of items in the anchor at the beginning. Then, it successively excludes items from the anchor.

The research of Wang and Yeh (2003), Wang (2004), Shih and Wang (2009) and Wang *et al.* (2012) made clear that the direction of DIF influences the results of the DIF analysis using all other items as anchor: If all items favor one group, DIF tests using all other items as anchor perform weak. Hence, in complex DIF situations such as unbalanced DIF, the initial step of the iterative-backward anchor class, that includes all other items as anchor, may lead to biased test results.

Inspired by the results of Wang and Yeh (2003), Wang (2004), Shih and Wang (2009) and Wang *et al.* (2012), we introduce another possible strategy to overcome the problem that the anchor selection is based on initially biased test results: The *iterative forward* anchor class. As opposed to the iterative backward class, we suggest to build the iterative anchor in a step-by-step forward procedure. Starting with one single anchor item, we link the scales and estimate DIF. Then, iteratively, as long as the current anchor length is shorter than the number of currently presumed DIF-free items, one item – located by means of the respective anchor selection strategy – is added to the current anchor and DIF analysis is conducted using the new current anchor. Unlike the iterative backward anchor class where items are successively excluded, now items are successively included in the anchor. An anchor selection strategy is again needed to guide which items are included in the anchor.

The rationale behind the *iterative forward* approach is similar to the common approach in confirmatory factor analysis (CFA) described by Stark *et al.* (2006), where one referent item is constrained to conduct DIF analysis. Here, we determine an order of candidate anchor items and guide the decision whether additional anchor items are included by DIF tests for all items (except the first anchor item, see Section 2.3) using the current anchor.

2.2. Anchor selection strategies

The anchor selection strategies discussed here are based on preliminary item analyses. This means that DIF tests are conducted to locate – ideally – DIF-free anchor items. The (non-statistical) alternative relying on expert advice and certain prior knowledge of DIF-free anchor items (Wang 2004; Woods 2009) will not often be possible in practice (for a literature overview where this approach fails see Frederickx, Tuerlinckx, De Boeck, and Magis 2010).

In our simulation study, we implemented different anchor selection strategies that provide a ranking order of candidate anchor items. One anchor selection strategy investigated in this article is the rank-based strategy proposed by Woods (2009) that we term all-other (AO) anchor selection strategy. Initially, every item $j = 1, \ldots, k$ is tested for DIF using all other items as anchor. This implies an anchor set $\mathcal{A}^j = \{1, \ldots, k\} \setminus j$ that depends on the studied

item j. The anchor is then chosen according to the lowest rank(s) of the resulting (absolute) DIF test statistics.

A selection strategy proposed by Wang (2004) – applied in a modified version for the MIMIC procedure by Shih and Wang (2009) – is called single-anchor (SA) selection strategy here and, to our knowledge, for the first time systematically compared with the all-other strategy using various anchor classes. With every item acting as single-anchor, every other item is tested for DIF. Again, the anchor sets $\mathcal{A}^j$ vary across the studied items and $k-1$ tests result for every item $j = 1, \ldots, k$ of the test. The anchor is then chosen from the items with the smallest number of significant results. If more than one item displays the same number of significant results, one of the corresponding items is selected randomly. A similar approach was suggested by Cheung and Rensvold (1999) in the context of factorial invariance to concurrently select a set of non-variant items that also relies on results testing every item by using every other item in the restriction.

2.3. Anchor methods

An *anchor method* results as a combination of an anchor class with an anchor selection strategy (in cases where the latter is necessary). The anchor methods to be investigated in this article are now presented with details of the implementation and summarized in Table 1.

The all-other anchor method (*all-other*) does not require an additional anchor selection. Every item is tested for DIF using all remaining items as anchor items. Here, the anchor set $\mathcal{A}^j = \{1, \ldots, k\} \setminus j$ depends on the studied item $j = 1, \ldots, k$ and one re-linking step is necessary for each item.

All remaining anchor methods consist of two steps: Firstly, the anchor selection is carried out to determine a ranking order of candidate anchor items and the proceeding of the anchor class is carried out to determine the final anchor. Secondly, the final anchor found in the first step is then used for the assessment of DIF. This procedure was termed DIF-free-then-DIF strategy by Wang *et al.* (2012). For the following anchor methods, the final anchor $\mathcal{A}$ is independent of which item is studied. Since $k-1$ parameters are free in the estimation, only $k-1$ estimated standard errors result (Molenaar 1995), the k-th standard error is determined by the restriction and, hence, only $k-1$ tests can be carried out and one item in the final assessment of DIF obtains no DIF test statistic. Thus, for the following anchor methods, the first item selected as anchor item is presumed DIF-free in the final DIF test and all remaining test items are tested for DIF using the final anchor $\mathcal{A}$.

The constant anchor class consisting of one anchor item or four anchor items can be combined with the all-other selection strategy (*single-anchor-AO*, *four-anchor-AO*). The anchor is selected according to the lowest rank(s) of the DIF test statistics resulting from the initial DIF tests for every item using the all-other method (Woods 2009). The final DIF test is conducted after re-linking the parameters using the final anchor.

The constant anchor class can also be combined with the single-anchor selection strategy (*single-anchor-SA*, *four-anchor-SA*). In this case, initially every item is tested for DIF by using every other item as single-anchor. The final anchor (consisting of either one item for the single-anchor-SA or four items for the four-anchor-SA method) is selected from the set of items displaying the smallest number of significant test results.

Furthermore, we implemented a suggestion of Wang (2004) that we refer to as the four-anchor next candidate (NC) method. In the *four-anchor-NC* method, the item that is selected by the SA-selection strategy functions as the current single-anchor and DIF tests are conducted (see Wang 2004, p. 249) similar to the single-anchor-SA method. In this step, one DIF test

Anchor class	Anchor selection	Combination	Initial step and anchor selection strategy
all-other	none	all-other cf. e.g., Woods (2009)	Initial step: Each item is tested for DIF using all remaining items as anchor. Selection strategy: No additional selection strategy is required.
constant	AO	single-anchor-AO Woods (2009)	Initial step: Each item is tested for DIF using all remaining items as anchor. Selection strategy: The item with the lowest absolute DIF statistic (AO) is chosen.
	SA	single-anchor-SA Wang (2004)	Initial step: Each item is tested for DIF using every other item as single-anchor. Selection strategy: The item with the smallest number of significant DIF tests (SA) is chosen.
	AO	four-anchor-AO Woods (2009); Wang *et al.* (2012)	Initial step: Each item is tested for DIF using all remaining items as anchor. Selection strategy: The four anchor items corresponding to the lowest ranks of the absolute DIF statistics from the initial step (AO) are chosen.
	SA	four-anchor-SA Wang (2004)	Initial step: Each item is tested for DIF using every other item as single-anchor. Selection strategy: The four anchor items corresponding to the smallest number of significant DIF tests (SA) are chosen.
	NC	four-anchor-NC proposed by Wang (2004)	Initial step: Each item is tested for DIF using every other item as single-anchor. Selection strategy: The first anchor is found as in single-anchor-SA; the next candidate anchor item (up to three) is found from tests using the current anchor if its result corresponds to the lowest non-significant absolute test statistic and is then added to the current anchor.
iterative backward	AO	iterative-backward-AO e.g., Drasgow (1987)	Initial step: Each item is tested for DIF using all remaining items as anchor. Selection strategy: Iteratively, all items displaying DIF are excluded from the anchor and the next DIF test with the current anchor is conducted.
iterative forward	AO	iterative-forward-AO	Initial step: Each item is tested for DIF using all remaining items as anchor. Selection strategy: As long as the current anchor is shorter than the number of currently presumed DIF-free items, the next item with the lowest rank in the initial step (AO) is added to the anchor.
	SA	iterative-forward-SA	Initial step: Each item is tested for DIF using every other item as single-anchor. Selection strategy: As long as the current anchor is shorter than the number of currently presumed DIF-free items, the next item with the smallest number of significant test results in the initial step (SA) is added to the anchor.

Table 1: Classification and nomenclature of the investigated anchor methods.

statistic results for every item except for the anchor. The next candidate anchor item is the item that displays "*the least magnitude of DIF*" (Wang 2004, p. 250) among all remaining items that we defined as lowest absolute DIF test statistic from the tests using the current single-anchor item. The candidate item is added to the current anchor only if its DIF test result is not significant (Wang 2004).[1] The next DIF test is conducted using the new current anchor and the next candidate item is selected again if it has the lowest absolute DIF test statistic among all remaining items and displays no significant DIF. These steps are repeated until either the next candidate anchor item displays DIF or the maximum anchor length (of four items in our implementation) is reached.

Technically speaking, this procedure is a combination of the constant and the iterative anchor class because it allows a varying anchor length but its length is limited to a pre-specified number of items. However, since in our simulation always four anchor items were selected for the final anchor, here we classify the anchor class as constant.

The iterative backward class is implemented using all-other items as anchor in the initial step and then excluding DIF items from the anchor (*iterative-backward-AO*) as it is widely used in practice (e.g., Bolt, Hare, Vitale, and Newman 2004; Edelen, Thissen, Teresi, Kleinman, and Ocepek-Welikson 2006). When items are excluded from the anchor, the scales are re-linked with the new current anchor and DIF tests are conducted. Items displaying DIF are again excluded from the anchor and the steps are repeated until the final anchor is found.[2] Note that the iterative backward class is not combined with the SA-selection since the latter provides only a ranking order of candidate anchor items, but no information which set of items should be used in the initial step.

The newly suggested iterative forward class can be combined with the all-other anchor selection strategy (*iterative-forward-AO*). In this case, DIF analysis for every item is conducted using the all-other method. The DIF test statistics are again ranked by their absolute value. The item corresponding to the lowest DIF test statistic is the first anchor item. As long as the anchor length does not exceed the number of currently presumed DIF-free items, the item with the next lowest absolute DIF test statistic from the initial step is included in the current anchor and DIF analysis is carried out using the new current anchor to check whether the anchor length exceeds the number of presumed DIF-free items. When the final iteration is reached, again, DIF analysis is conducted with the new final anchor, where, again, the first anchor item is presumed DIF-free.

Furthermore, the proposed iterative forward class can be also combined with the single-anchor selection strategy (*iterative-forward-SA*). In this case, initially every item is tested for DIF using every other item as single-anchor. The items are ranked according to the number of significant test results. The item with the smallest number of significant test results is chosen as anchor. As long as the anchor length does not exceed the number of currently presumed DIF-free items, the next (new) item displaying the smallest number of significant DIF tests from the initial step is included in the anchor and DIF analysis is carried out using the new current anchor to check whether the current anchor length exceeds the number of currently presumed DIF-free items. When the final iteration is reached, the DIF test results of the DIF analysis using the final anchor are returned and the first anchor item is presumed DIF-free.

[1] We employed a significance level of .05, but future research may investigate the additional proposition of Wang (2004) augmenting it to e.g., .30.

[2] In case all items were excluded from the anchor (which happened in only 2 out of 154,000 replications), one single anchor item was chosen randomly in our simulation study.

3. Simulation study

To evaluate which of the anchor methods presented in the previous section (for a brief description and nomenclature see again Table 1) are best suited to correctly classify items with and without DIF, an extensive simulation study is conducted. 2000 data sets are generated from each of 77 different simulation settings. For every data set, the item-wise Wald test (Lord 1980, see Section 2) – based on the currently investigated anchor method – is conducted at the significance level of .05 in the free R system for statistical computing (R Development Core Team 2011). A short description of the study design is given in the following paragraphs. Parts of the simulation design were inspired by the settings used by Woods (2009) and Wang (2004).

3.1. Data generating process

Each data set, that represents one of 2000 replications from one simulation setting, corresponds to the simulated responses of two groups of subjects (the *reference* (ref) and the *focal* (foc) group) in a test with $k = 40$ items.

- *Person and item parameters*

  The person parameters are generated from a normal ability distribution with a higher mean for the reference group $\theta^{\text{ref}} \sim N(0.5, 1)$ than for the focal group $\theta^{\text{foc}} \sim N(0, 1)$. Values assigned to the item parameters are replicated from the sequence of $\beta_j^{\text{ref}} \in \{-1, -0.5, 0, 0.5, 1\}$ in equal proportions.

- *DIF items*

  In case of DIF, the affected DIF items are chosen randomly to display uniform DIF by setting the difference in the item parameters of reference and focal group $\Delta_{\text{DIF}} = \beta_j^{\text{ref}} - \beta_j^{\text{foc}}$ to $+.5$ or $-.5$ (consistent with the intended direction of DIF).

- *IRT model*

  The responses in each group follow the Rasch model. They are generated in two steps: The probability of person i solving item j is computed by placing the corresponding item and person parameters in the Rasch model formula 6. The binary responses are then drawn from a binomial distribution with the resulting probabilities.

$$P\left(U_{ij} = 1 \,|\, \theta_i, \beta_j\right) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{6}$$

3.2. Manipulated variables

Three main conditions determine the specification of the manipulated variables: One condition under the null hypothesis where no DIF is present and two conditions under the alternative where DIF is present.

- *Sample sizes*

  The sample sizes in reference and focal group are defined by the following pairs $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250, 250), (500, 250), (500, 500), (750, 500), (750, 750), \ldots, (1500, 1500)\}$. Thus, both equal and different group sizes are considered.

- *Directions and proportions of DIF*

  Under the condition of the null hypothesis (*no DIF*), only the sample sizes are varied. The two remaining conditions represent the alternative hypothesis where DIF is present, but they differ with respect to the direction of DIF: The second condition represents *balanced DIF*. Here, each DIF item favors either the reference or the focal group but no systematic advantage for one group remains because the effects cancel out. For the third *unbalanced DIF* condition a systematic disadvantage for the reference group is generated such that every DIF item favors the focal group. In addition to the sample size, also the proportion of DIF is manipulated including the following percentages $p \in \{15\%, 30\%, 45\%\}$.

3.3. Outcome variables

To allow for a comparison of the anchor methods, the classification accuracy of the DIF tests is evaluated by means of false alarm rate and hit rate.

- *False alarm rate*

  For a single replication the *false alarm rate* is defined as the proportion of DIF-free items that are (erroneously) diagnosed with DIF. The estimated false alarm rate for each experimental setting is computed as the mean over all 2000 replications and, thus, corresponds to the *type I error rate*. Similarly, the standard error is estimated as the square root of the unbiased sample variance over all replications.

- *Hit rate*

  Analogously, for a single replication the *hit rate* is computed as the proportion of DIF items that are (correctly) diagnosed with DIF. The hit rate is only defined in conditions that include DIF items, namely in the balanced and unbalanced condition. The estimated hit rate and the standard error are again computed as mean and standard error over all 2000 replications and correspond to the *power* of the statistical test and its variation.

- *Further outcome variables*

  Moreover, the percentage of replications where at least one item in the anchor is a DIF item (*risk of contamination*) is computed over all replications of one setting. The average proportion of DIF items as compared to the overall number of anchor items (*degree of contamination*) is computed, too, for replications where the anchor is contaminated. Average false alarm rates are also computed separately for the tests based on a contaminated and for the tests based on a pure (not contaminated) anchor to allow for a more detailed interpretation of the results.

3.4. Research hypotheses

The simulation study is carried out to investigate the trade-off between the false alarm rate and the hit rate of DIF tests using the anchor methods introduced in Section 2.3.

Under the condition of the null hypothesis, we expect all anchor methods to yield well-controlled false alarm rates, since no DIF items and, therefore, no risk of contamination

exists (Wang and Yeh 2003; Stark *et al.* 2006; Woods 2009; González-Betanzos and Abad 2012).

Under the condition of balanced DIF, based on previous simulation studies (Wang and Yeh 2003; Wang 2004), we expect the all-other class to yield a well-controlled false alarm rate and a high hit rate.

Under the condition of unbalanced DIF, where all DIF items are simulated to favor one group, an augmented false alarm rate for the all-other method is expected (Wang and Yeh 2003; Wang 2004).

In both conditions of the alternative, balanced and unbalanced DIF, we anticipate the constant anchor class to show an increase in the false alarm and the hit rate when the anchor length rises from one to four items and the proportion of DIF items is high (Thissen *et al.* 1988; Woods 2009). Wang *et al.* (2012) also found that four anchor items combined with the IRTLRDIF procedure (Thissen 2001) yielded low power rates as might also be the case in this simulation with the Wald test.

González-Betanzos and Abad (2012) compared an iterative backward two-step procedure based on the AO-selection strategy to specific constant single-anchors, to a purification procedure based on a DIF-free constant single anchor and to the all-other method. The constant single-anchor items were selected from the set of known a priori DIF-free items. The iterative backward two-step procedure showed good performance. Due to the fact that one additional purification step improved the test results, the authors assumed improvements when further purification steps are added as we have implemented in this article. Accordingly, we expect the iterative backward anchor class to achieve high hit rates as they allow for a long anchor, but at the expense of an augmented false alarm rate especially in settings where the proportion of DIF items is high and DIF is unbalanced.

Little information is available on how well the anchor selection strategies perform, as Wang and Yeh (2003), Wang (2004) and Thissen *et al.* (1988) included only DIF-free items in the constant anchor class. This approach is only possible in simulation studies, however, where it is known by design which items are DIF-free. In practice, on the other hand, a set of DIF-free items prior to DIF analysis is usually not available (González-Betanzos and Abad 2012). Including only DIF-free items avoids the risk of contamination (for the consequences of contamination see Section 2) and, thus, leads to an advantage for the methods from the constant anchor class. However, in order to compare the anchor classes under realistic conditions where it is not known a priori which items are DIF-free, the methods from the constant anchor class should be investigated together with an anchor selection strategy.

Woods (2009) investigated the AO-selection strategy to locate a set of constant anchor items and found results suitable for DIF analysis and superior to the all-other method. However, Wang *et al.* (2012) investigated the constant anchor method based on the selection of four anchor items using the AO-selection strategy (here referred to as the four-anchor-AO method) and found that poor results occurred when DIF was unbalanced and no additional purification step was used. Therefore, we expect the four-anchor-AO method to perform well only in the condition of balanced DIF and poorly in the condition where DIF is unbalanced (Wang and Yeh 2003; Wang 2004; Shih and Wang 2009; Wang *et al.* 2012).

The SA-selection strategy proposed by Wang (2004) is (to our knowledge) implemented and combined with several anchor classes here for the first time. Since the SA-selection strategy relies on DIF tests using every item as single anchor, we anticipate the SA-selection strategy to outperform the AO-selection strategy if the sample size is large and DIF is unbalanced. When DIF is balanced, we expect the AO-selection strategy to be superior.

The newly suggested iterative forward class builds the anchor in a step-by-step forward procedure. In comparison with the iterative backward method, we expect the forward procedure to be superior when the SA-selection strategy is used and DIF is unbalanced since the initial step of the iterative backward procedure is built on biased test results. In comparison with methods from the constant anchor class, we anticipate higher hit rates because the anchor of the iterative forward procedure grows as long as the current anchor is shorter than the number of currently presumed DIF-free items and should, thus, include more than four items. As a drawback, we also expect higher false alarm rates since the risk of contamination increases with the anchor length. Furthermore, we anticipate the methods from the iterative forward class to show lower hit rates than the all-other method in the balanced case, because the latter uses all items – except the studied item – as anchor.

In summary, nine anchor methods are compared for eleven different sample sizes in one condition of the null hypothesis (no DIF) and two conditions of the alternative (balanced and unbalanced DIF) with three different proportions of DIF items (15%, 30% or 45%). The aim of this simulation study is to assess their appropriateness to identify DIF and DIF-free items and to evaluate the trade-off between the false alarm rate and the hit rate. Altogether, 77 different simulation settings result. Each setting is replicated 2000 times to ensure reliable results.

4. Results

4.1. Null hypothesis: No DIF

In the first condition, all items are truly DIF-free. Therefore, only the false alarm rate (proportion of DIF-free items that are diagnosed with DIF) is computed.

False alarm rates

The estimated false alarm rates are depicted in Figure 2 and (for equal sample sizes to save space) also reported together with their standard errors in Table 2 in the Appendix. As shown in Figure 2, all anchor methods hold the 5% level. While methods from the all-other, the iterative backward (iterative-backward-AO) and the iterative forward class (iterative-forward-SA, iterative-forward-AO) exhaust the significance limit, methods from the constant anchor class (constant single-anchors: single-anchor-AO and single-anchor-SA; constant four-anchors: four-anchor-AO, four-anchor-SA and four-anchor-NC) remain below that level. The constant single-anchors – that consist of an anchor with the constant length of only one item – display false alarm rates not exceeding 0.01, whereas the constant four-anchors display slightly higher false alarm rates (approximately 0.03 for the constant four-anchor-AO as well as for the four-anchor-SA method and about 0.045 for the constant four-anchor-NC method). Hence, DIF tests with an anchor method from the constant anchor class – especially the constant single-anchor methods – are over-conservative.

Summary

All anchor methods hold the significance level. Over-conservative test results are found for the constant anchor methods, whereas the newly suggested iterative-forward-AO and iterative-forward-SA methods together with the all-other and the iterative-backward-AO methods exhaust the significance level.

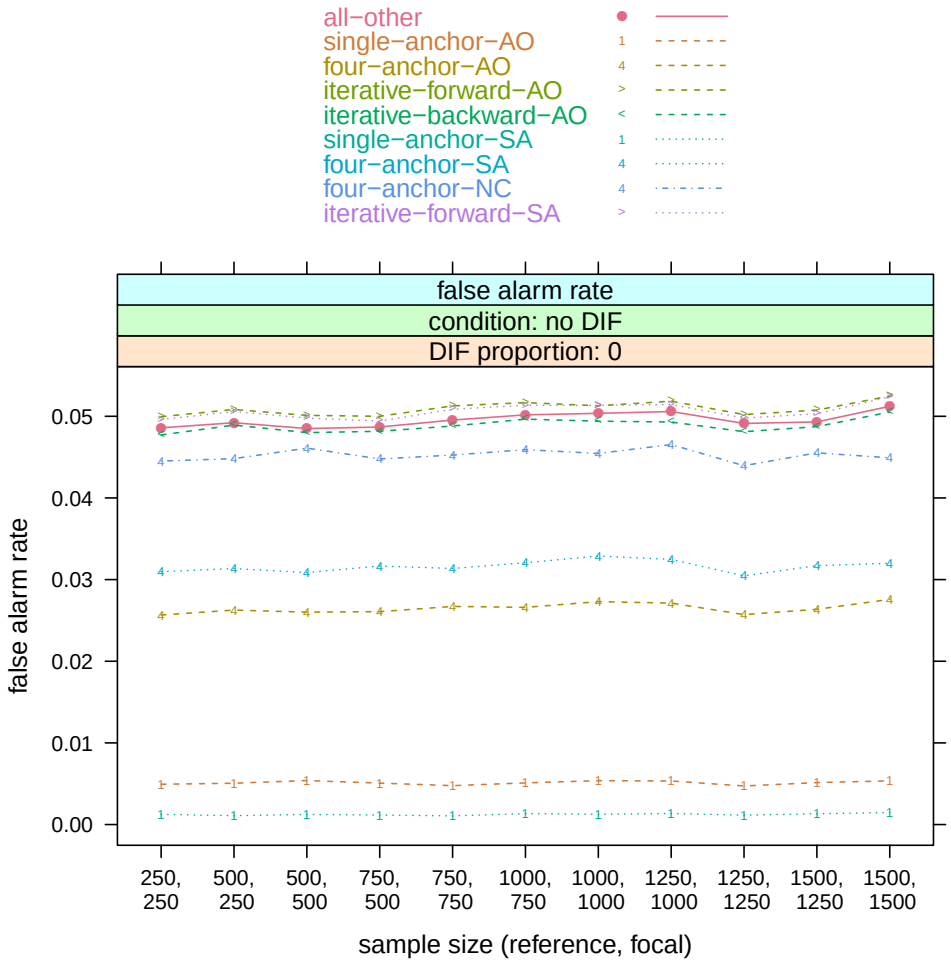

Figure 2: False alarm rates under the null hypothesis of no DIF.

4.2. Balanced DIF: No advantage for one group

In the balanced condition, a certain proportion of DIF items (15%, 30% or 45%) is present. Each DIF item favors either the reference or the focal group, but the single advantages cancel out.

False alarm rates

Figure 3 (top row) contains the false alarm rates for the balanced condition, reported also for equal sample sizes together with the standard errors in Table 3 in the Appendix. Most methods display well-controlled false alarm rates – similar to the null condition – with the following exceptions: The constant four-anchor-SA and the constant four-anchor-NC method show a false alarm rate that first increases but then decreases again with growing sample size. The same inverse u-shaped pattern occurs in case of unbalanced DIF and will be explained in more detail in Section 6.

Both constant single-anchor methods (single-anchor-AO and -SA) as well as the four-anchor-AO method, again, remain below the significance level. Hence, DIF tests based on the single-anchor-AO, the single-anchor-SA and the four-anchor-AO method are over-conservative.

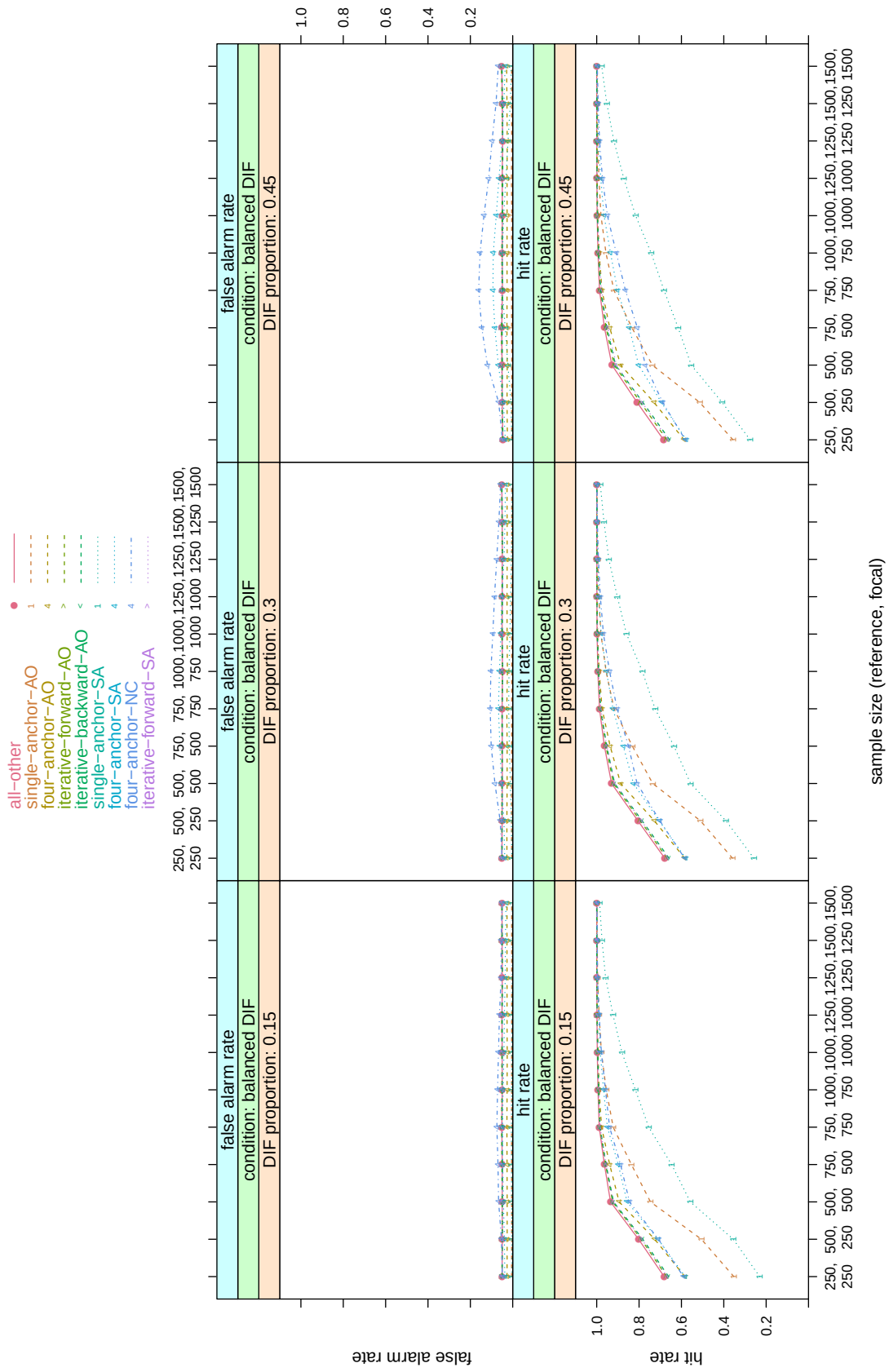

Figure 3: Balanced condition: 15%, 30% and 45% DIF items with no systematic advantage for one group; sample size varies from $(250, 250)$ up to $(1500, 1500)$; top row: false alarm rates; bottom row: hit rates in the balanced condition.

Hit rates

For DIF test results, hit rates specify how likely true DIF is detected. Figure 3 (bottom row) depicts the hit rates in the balanced condition that increase monotonically with the sample size (for standard errors see also Table 4 in the Appendix). The hit rates that rise slowly correspond to the constant single-anchor methods, but also to the constant four-anchor methods. The methods from the constant anchor class that are combined with the AO-selection (single-anchor-AO, four-anchor-AO) achieve higher hit rates than those combined with the SA-selection (single-anchor-SA, four-anchor-SA) or its modification (four-anchor-NC). In terms of hit rates, all iterative procedures (iterative-forward-AO, iterative-forward-SA and iterative-backward-AO) as well as the all-other method show rapidly increasing hit rates that converge to one for sample sizes above 750 in each group.

Summary

In the balanced condition, the AO-selection strategy outperforms the SA-selection by yielding higher hit rates as expected. The difference is large for methods from the constant anchor class, but negligible for methods from the iterative forward anchor class.

All anchor methods show a well-controlled false alarm rate, except the constant four-anchor-SA and the four-anchor-NC method. All iterative methods (from the forward and backward class) and the all-other method display the most rapidly rising hit rates. The newly suggested iterative-forward-AO and iterative-forward-SA method enable a high rate of correctly classified DIF items and simultaneously maintain the significance level in the balanced condition.

4.3. Unbalanced DIF: Advantage for the focal group

In the unbalanced condition, all items simulated with different item parameters favor the focal group. False alarm rates for the unbalanced condition are shown in Figure 4 (top row) and in Table 5 in the Appendix together with the standard errors (again only for settings with equal sample sizes in reference and focal group to save space).

False alarm rates

As opposed to the previous results, in this condition the majority of the anchor methods produce augmented false alarm rates: When the proportion of DIF items increases, the false alarm rates rise as well. Moreover, for most anchor methods, the false alarm rates increase with growing sample size. The settings from the unbalanced condition – especially with 30% and 45% DIF items – are now discussed in more detail in groups of anchor classes.

The all-other method yields the highest false alarm rate in the majority of the simulation settings. The reason for this is that the all-other method is always contaminated in situations where more than one item has DIF. On average, the mean item parameters of the focal group are lower than the mean item parameters of the reference group. These mean differences in the item parameters shift the scales of focal and reference group apart when the all-other method defines the restriction (similar to the instructive example in Section 2). These artificial differences become significant when the sample size increases and, thus, result in an augmented false alarm rate.

For methods from the constant anchor class, the selection strategy explains the false alarm rates: The strategy of selecting anchors based on the DIF tests with all-other items as anchor yields biased DIF test results that induce a high false alarm rate when the sample size is large (as illustrated and discussed in more detail regarding the impact of contamination in

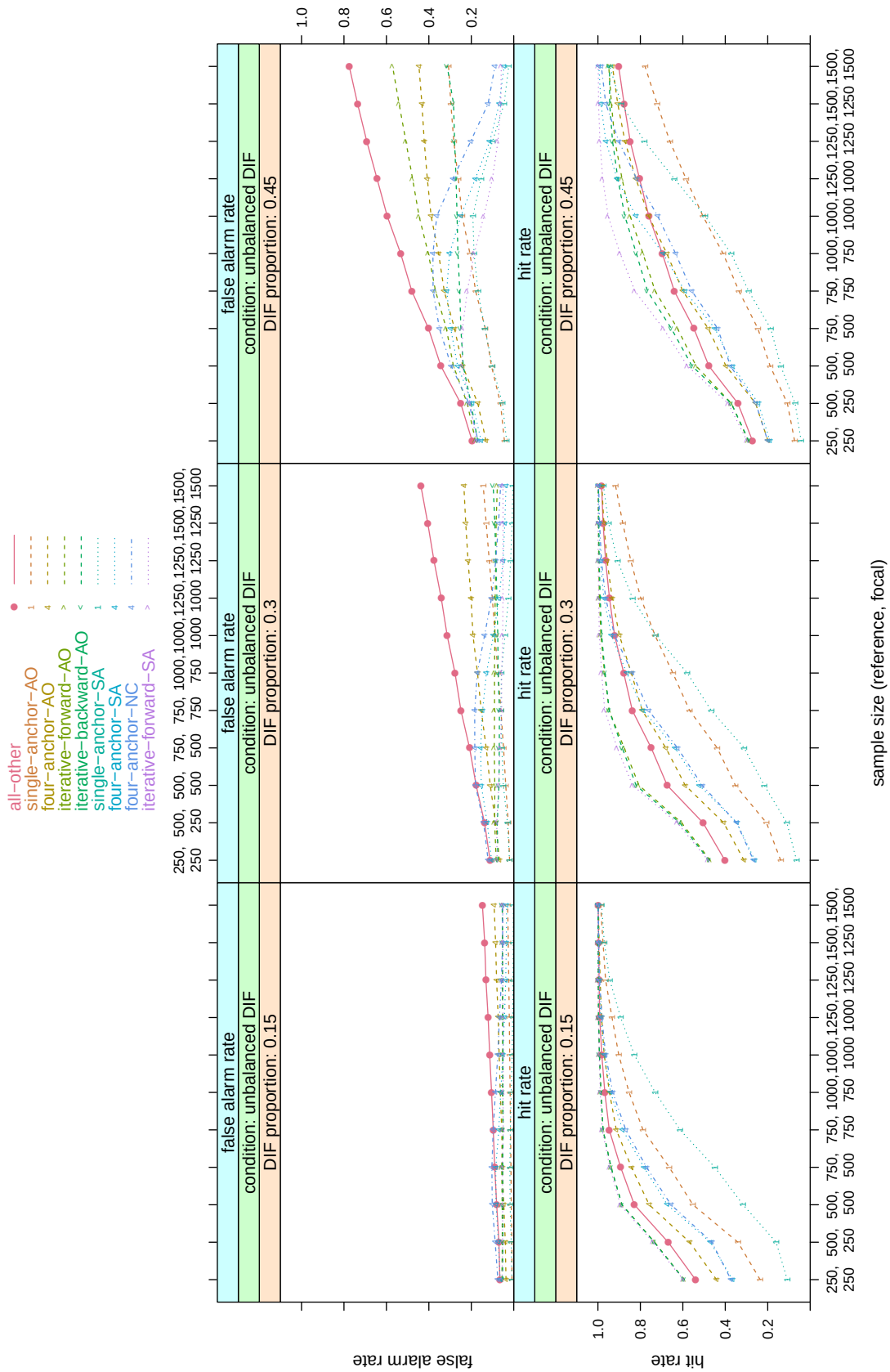

Figure 4: Unbalanced condition: 15%, 30% and 45% DIF items favoring the focal group; sample size varies from $(250, 250)$ up to $(1500, 1500)$; top row: false alarm rates; bottom row: hit rates in the unbalanced condition.

Section 5).

Constant anchors selected by the single-anchor strategy produce lower false alarm rates in regions of medium or large sample sizes. Here, again, an inversely u-shaped form is visible. After a certain point, the false alarm rates decrease again (a detailed explanation will be given in Section 6). The constant single-anchor methods show lower false alarm rates than the corresponding constant four-anchor methods. For all constant methods, the single-anchor-SA method has the lowest false alarm rate when the sample size is large. Only in the condition with 45% DIF items, it systematically exceeds the significance level at medium sample sizes.

The method from the iterative backward anchor class, which starts the initial step by using the all-other method, also leads to augmented false alarm rates (with a maximum observed rate of 0.31 for the highest sample size in case of 45% DIF) that rise when sample size increases.

Methods from the iterative forward class display heterogeneous false alarm rates. The iterative-forward-AO method leads to increased false alarm rates – similar to the constant methods with the AO-selection criterion – in the setting with 30% or 45% DIF (up to 0.57 for the highest sample size). The clearly best iterative method in terms of a low false alarm rate is the new iterative-forward-SA method (yielding a false alarm rate of 0.06 for the highest sample size in case of 45% DIF and an observed maximum of 0.24).

Hit rates

The hit rate in the unbalanced condition (cf. Figure 4 (bottom row) and Table 6 in the Appendix) in the settings of larger proportions of DIF items is different: Generally, the overall level of the hit rate is lower. The rise of the hit rates with the sample size is slow for the methods from the constant anchor class as well as for the all-other method. These methods also have lower hit rates compared to the methods from the iterative forward or backward class that are the only methods that display rapidly increasing and high hit rates. The new iterative forward-SA method provides the highest hit rate and a rapid rise of the hit rate with increasing sample size.

The SA-selection strategy in combination with methods from the constant anchor class is more suitable than the AO-selection strategy regarding the hit rates when the sample size is large. The four-anchor-SA method outperforms the modified constant four-anchor method (four-anchor-NC) in terms of higher hit rates (and lower false alarm rates). The iterative forward procedure with the SA-selection is equal or superior to the iterative-forward-AO method over the entire range of simulated sample sizes.

Summary

In the unbalanced condition, the SA-selection strategy is superior to the AO-selection strategy when the sample size and the DIF proportion are high as expected, since it not only allows a higher hit rate but it also corresponds to a lower false alarm rate.

In the condition of unbalanced DIF, the false alarm rates are no longer well-controlled. When the DIF proportion is high, only the single-anchor-SA and the iterative-forward-SA method have low false alarm rates in regions of large sample sizes. Both constant single-anchor methods yield low false alarm rates – but also low hit rates – when the sample size is small. All methods from the constant anchor class, especially in regions of small sample sizes, show poor hit rates. The highest hit rate – in all settings from the unbalanced condition – corresponds to the newly proposed iterative-forward-SA method.

5. The impact of anchor contamination

As discussed in Section 2 the contamination of the anchor may induce artificial DIF and thus induce a seriously inflated false alarm rate. Short anchors are preferred in the constant anchor class in order to minimize the risk of contamination that includes only the information whether all anchor items are DIF-free or not (e.g., Shih and Wang 2009). When new anchor methods are proposed, they are judged by their ability to correctly locate a completely DIF-free anchor in order to avoid anchor contamination (e.g., Wang *et al.* 2012). Since contamination is an important argument in the construction of new anchor methods and since it affects the results of DIF detection (see Section 2), we will take a deeper look at the simulation results focussing on the aspect of anchor contamination.

For one exemplary simulation setting, the extreme condition of unbalanced DIF where 45% of the items favor the focal group will now be discussed in more detail in groups of anchor classes to illustrate the impact of contamination.

Figure 5 (top row) depicts the proportion of replications where at least one item of the anchor is a DIF item (top-left) – this is referred to as *risk of contamination* – and the proportion of DIF items in the anchor when the anchor is contaminated (top-right) – this is referred to as *degree of contamination*. The false alarm rates including only the replications that resulted in a contaminated anchor are displayed in Figure 5 (bottom-left) next to the false alarm rates including only the replications that resulted in a pure anchor (bottom-right). If no pure replications resulted, the respective false alarm rate is omitted.

The results show the following: For the all-other method all items function as anchor items. Due to the fact that more than one item has DIF in our simulation design, the anchor is contaminated in 100% of the replications. The proportion of DIF items in the anchor is 45% as simulated. With increasing sample size, the power of detecting artificial DIF (DIF-free items that display DIF due to the anchor method chosen) increases and, thus, the false alarm rate rises.

Methods from the constant anchor class are investigated next. The risk of contaminated anchors decreases when the sample size increases. The anchor selection strategy determines how rapidly: When the SA-selection strategy chooses the anchor, the convergence to zero percent contamination is clearly visible in the simulated range of the sample size. The AO-selection strategy shows a decreasing tendency, but even when 1500 observations for each group are available, 70% of the anchors are still contaminated for the four-anchor-AO and 17% for the single-anchor-AO method.

If the constant single-anchors are contaminated, this means that the single anchor item has DIF and, inevitably, the false alarm rates explode when the sample size is large enough to detect significant artificial DIF (in case of 1500 observations in each group: single-anchor-AO: 0.70, single-anchor-SA: 0.54).

Surprisingly, there is a large gap between the degree of contamination for the constant four-anchor methods: When the AO-strategy or the SA-strategy are directly chosen and the sample size is large, on average about one out of four anchor items has DIF. In contrast to this, about three out of four anchor items have DIF when the four-anchor-NC method is chosen. In contaminated situations, consequently, the four-anchor-NC method corresponds to a larger false alarm rate (observed maximum: 0.75) than the four-anchor-AO (observed maximum: 0.54) or the four-anchor-SA method (observed maximum: 0.36). Therefore, the four-anchor-NC method corresponds to larger false alarm rates compared to the four-anchor-SA method over all unbalanced conditions with 45% DIF items (see again Figure 4, top row), even though it has a lower risk of contamination. Hence, the degree of contamination is

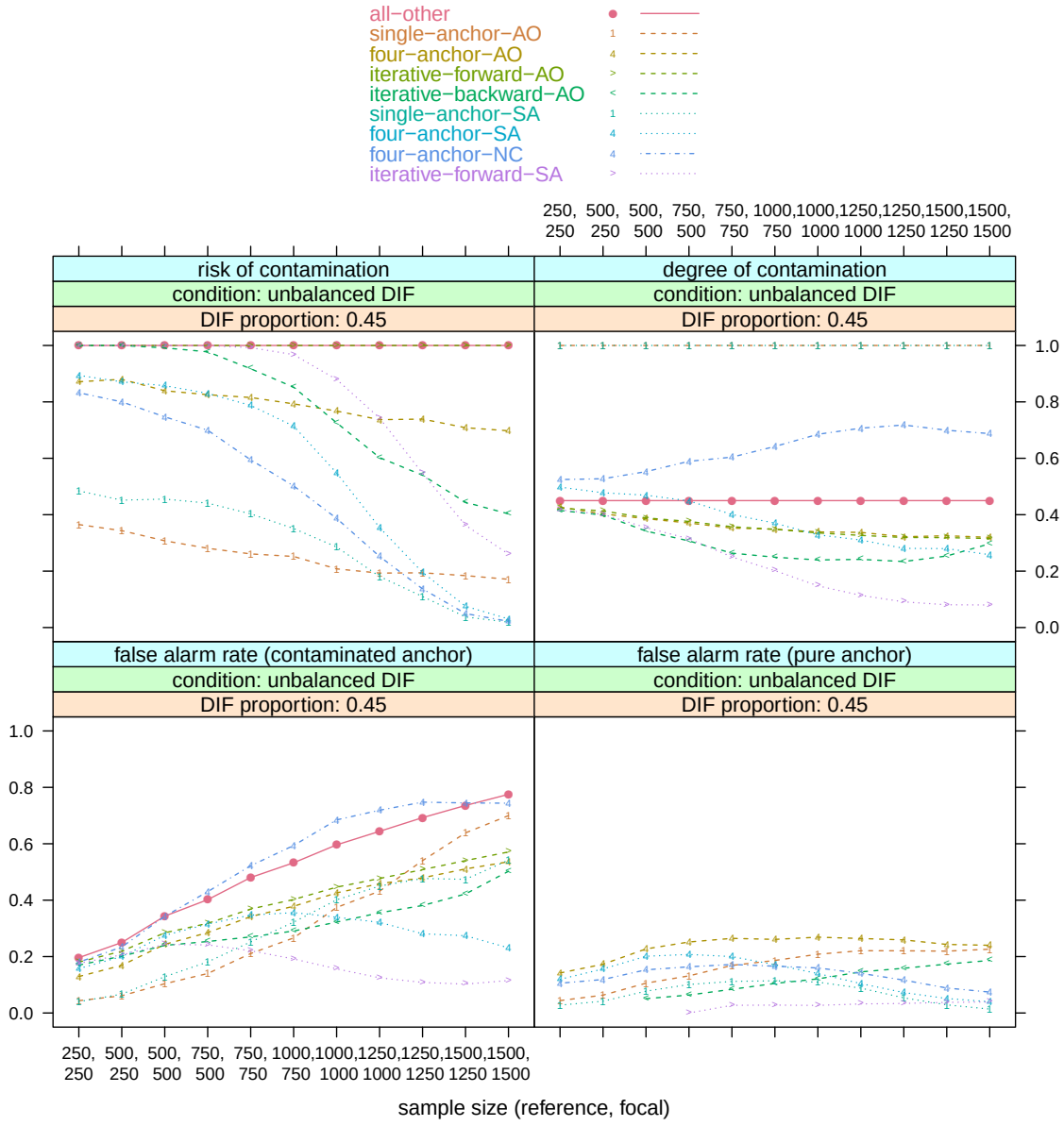

Figure 5: Condition of unbalanced DIF with 45% DIF items favoring the focal group; sample size ranges from $(250, 250)$ to $(1500, 1500)$; top-left: risk of contaminated anchors (at least one DIF item included in the anchor); top-right: degree of contamination (proportion of DIF items in contaminated anchors); bottom-left: false alarm rates when the anchor is contaminated; bottom-right: false alarm rates when the anchor is pure.

important for the results of the DIF assessment.

If the anchor is pure, the false alarm rates for the constant four-anchor method located by the SA-selection strategy (or the NC-selection) are lower. Note, however, that even if the anchor is pure, the false alarm rates of the constant anchor methods exceed the significance level. To clarify this fact, we will present an additional simulation study in the next section.

The methods from the iterative forward and backward anchor class are more often contaminated than the constant single-anchors (see Figure 5). This is not surprising, as the anchors include more items. Similar to the all-other method, all replications of the iterative-forward-AO method are contaminated. The iterative-forward-SA and the iterative-backward-AO method yield a risk of contamination that decreases with the sample size (minimum 0.26 and 0.40 in the simulated range). In case of contaminated anchors, the methods from the iterative forward and backward class also produce augmented false alarm rates. When the sample size in each group exceeds 750, the iterative-forward-SA method definitely has the lowest false alarm rate. When the sample size is 1000 in each group, the mean false alarm rate of the iterative-forward-SA method including all replications does not even exceed the mean false alarm rate of all three constant four-anchor methods including only the pure replications while simultaneously the hit rate is higher for the iterative-forward-SA method.

Our findings clarify that it is not the risk of contamination alone that explains the augmented false alarm rates. The best method in the unbalanced condition when the sample size is large is the iterative-forward-SA method even if it has a high risk of contamination. Nevertheless, the iterative-forward-SA method displays a low false alarm rate and also a high hit rate. Therefore, the consequences of contamination depend on the degree of contamination which is low for this method due to the SA-selection strategy that is suitable for unbalanced DIF. Research on DIF analysis and anchor selection procedures should, thus, not only concentrate on the risk of contamination, but also focus on the consequences, which strongly depend on the proportion of contaminated items in the anchor.

6. Characteristics of the anchor items inducing artificial DIF

In our simulation study, several anchor methods – especially the four-anchor-SA and the four-anchor-NC method – display inversely u-shaped false alarm rates that are yet to be explained. There are two mechanisms at work here: On one hand, the risk and the degree of contamination decrease with increasing sample size when the anchor selection strategy works appropriately and, thus, the extent of artificial DIF decreases. On the other hand, the power of detecting artificial DIF increases with growing sample size. One possible explanation for the inversely u-shaped pattern is the interaction between the decreasing extent of artificial DIF induced by anchor contamination and the increasing power of detecting statistically significant artificial DIF. In the beginning the false alarm rate increases due to the increasing power for detecting artificial DIF but at some point the false alarm rate decreases again as the risk of contamination decreases. This explanation is consistent with the findings from Section 5, where the anchor is contaminated: The four-anchor-SA method, for example, displays a degree of contamination that decreases with sample size (see again Figure 5, top-right) and an inversely u-shaped false alarm rate when the anchor is contaminated (see again Figure 5, bottom-left).

However, with this argument we cannot yet explain why the false alarm rates show the same pattern for pure (uncontaminated) replications (see again Figure 5, bottom-right). Here, the single-anchor-SA, the four-anchor-SA as well as the four-anchor-NC method display inversely

u-shaped false alarm rates even though the anchor is pure. Therefore, the presence of artificial DIF that is induced by contamination alone cannot explain this finding. To understand this finding, it is important to note that artificial DIF can also be caused by special characteristics of the anchor items that were located by an anchor selection strategy.

To clarify how artificial DIF is related to the observed patterns of the false alarm rates, we conducted an additional simulation study focussing again on the extreme condition of unbalanced DIF where 45% of the items favor the focal group. Here, we examined the difference in the sum of the estimated anchor item parameters between focal and reference group that we termed *scale shift* (because it measures how far both scales are shifted apart during the construction of the common scale and this shift can cause artificial DIF as illustrated in Figure 1, right) for all constant four-anchor methods (the four-anchor-SA and the four-anchor-NC method that displayed inversely u-shaped patterns as well as the four-anchor-AO method that displayed an increasing false alarm rate, see again Figure 4, top-right). To assess reliable estimates of the scale shift, we used all items that are DIF-free by design as anchor items to build the ideal common scale. The scale shift represents how far the scales are shifted apart and reflects the extent of artificial DIF. The scale shift may be caused by contamination, as discussed in the previous sections, or by special characteristics of the anchor items in particular when the selection strategies locate anchor items that show relatively high empirical differences in the estimated item parameters due to random sampling fluctuation even if the located anchor items were simulated to be DIF-free.

To determine whether anchor items found by a selection strategy display this characteristic, we included a benchmark method of four anchor items that are randomly selected from the set of all DIF-free items. The benchmark method, thus, represents the ideal four-anchor method that does not select items with high differences more often than others. The results, separated for contaminated and pure replications, are depicted in Figure 6.

In case of contaminated anchors (Figure 6, left), the four-anchor methods display positive scale shifts. Even though the scale shifts are almost constant over the sample size in regions of small to medium sample sizes for the four-anchor-AO and four-anchor-NC or even slightly decreasing for the four-anchor-SA method, the false alarm rates rise with growing sample size in the respective range of the sample sizes (Figure 5, bottom-left). We attribute this fact to the increasing power of detecting artificial DIF. This also explains the increasing false alarm rates of the four-anchor-AO and the four-anchor-NC methods: The scale shifts are almost constant over the simulated range of the sample size but the false alarm rates increase (Figure 5, bottom-left).

For the four-anchor-SA method the scale shift also decreases with increasing sample size in regions of medium or large sample sizes (Figure 6, left) and so does the false alarm rate in the respective range of the sample sizes (Figure 5, bottom-left).

Our second argument now becomes important in the case of pure anchors: The scale shift for the benchmark method of randomly chosen DIF-free anchor items (Figure 6, right) fluctuates around zero and displays no systematic shift in one direction. However, the scale shift of all remaining constant four-anchor methods is positive. This represents the fact that the supposedly pure items chosen by an anchor selection strategy display different characteristics than randomly chosen pure anchor items. From all items that are "pure" by definition (i.e. were drawn from distributions with no parameter difference) the anchor selection strategies select not the ones with the lowest empirical difference (due to random sampling), as one might hope, but those with a large empirical difference which again induces artificial DIF for the other items.

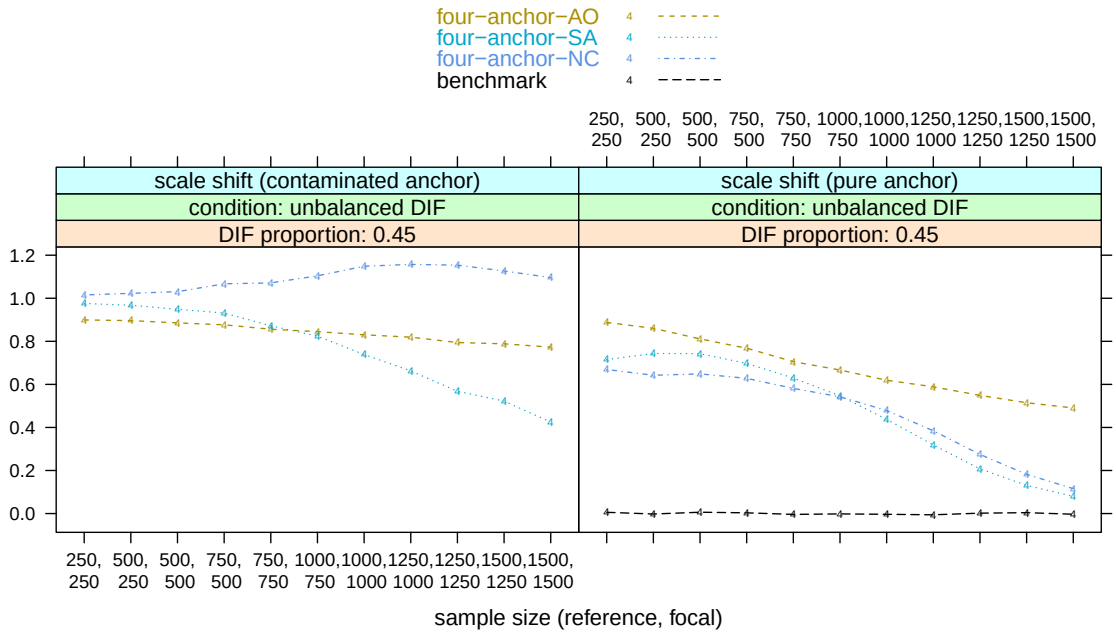

Figure 6: Condition of unbalanced DIF with 45% DIF items favoring the focal group; sample size ranges from $(250, 250)$ to $(1500, 1500)$; left: the scale shift when the anchor is contaminated; right: the scale shift in case of pure anchors.

As can be seen from Figure 6 (right), the scale shift for the four-anchor methods reduces with increasing sample size. In regions of large sample sizes, the scale shift is directly related to the false alarm rate: When the scale shift is high (as is the case for the four-anchor-AO method), the false alarm rate is high as well (Figure 5, bottom-right). When the scale shift decreases with growing sample size (e.g., for the four-anchor-SA method), the corresponding false alarm rate decreases as well. In regions of smaller sample sizes, the scale shift of all four-anchor methods is high, but the false alarm rates are low at the beginning and then increase with growing sample size. Here, again, the interaction between the extent of artificial DIF – now induced by large empirical differences in the anchor items – and the power of detecting artificial DIF is visible.

These findings explain why the u-shaped patterns occur for the four-anchor-SA and the four-anchor-NC method: These methods are able to reduce the scale shift with increasing sample size because the scale shift in pure settings reduces and the risk of contamination reduces as well (i.e. the number of pure settings increases). Taking the increasing power of detecting artificial DIF with growing sample size into account, an inversely u-shaped pattern results for the false alarm rates. In contrast to this, the four-anchor-AO method always displays a relatively high scale shift (that only reduces slightly when the anchor is pure). The power of detecting artificial DIF increases with growing sample size and, therefore, the false alarm rate shows an increase and no considerable decrease.

In summary, the interaction between a decreasing extent of artificial DIF and an increasing statistical power to detect artificial DIF with growing sample size results in an inversely u-shaped false alarm rate. The risk and degree of contamination alone cannot explain the presence of artificial DIF. The anchor items selected by certain anchor selection strategies differ systematically from randomly chosen pure anchor items even if the located anchor

items are by definition DIF-free. Counterintuitively instead of items with small differences, these methods tend to select exactly those items with large differences. Therefore, the anchor items found by the SA-, the NC- or the AO-selection strategy display a positive scale shift in the additional simulation study and, thus, shift the scales apart and induce artificial DIF. This implies that not only the risk and the degree of contamination but also the scale shift in by definition pure replications should be regarded when anchor methods are developed and investigated in simulation studies. Otherwise, augmented false alarm rates might occur even if the anchor is pure.

7. Summary and discussion

To conclude, the results from the simulation study are briefly reviewed. Thereafter, practical guidelines on how to choose the anchor for DIF analysis in the Rasch model are given. Finally, future research questions are addressed.

Conclusions from the simulation study

The assessment of differential item functioning for the Rasch model was investigated. The results of the Wald test were compared by means of hit and false alarm rates under three main conditions: The null hypothesis of no DIF, the balanced condition where no group had a systematic advantage and the unbalanced condition where all items favored the focal group.

Under the null hypothesis, all methods from the iterative forward and backward class as well as the all-other method exhaust the significance level, while methods from the constant anchor class remain below that level.

When DIF is balanced, the all-other method and also methods from the iterative forward and backward class yield high hit rates while simultaneously exhausting the significance level. As expected, the all-other selection strategy outperforms the single-anchor selection strategy.

In case of unbalanced DIF, the SA-selection procedure is superior to the AO-selection strategy when the sample size is large. In this case, the newly suggested iterative-forward-SA method yields the highest hit rate and a low false alarm rate and is, thus, the best performing anchor method.

The constant four-anchor class was not only combined with the AO-selection and the SA-selection strategy, but also modified in a way Wang (2004) suggested. Even though the four-anchor-NC method leads to a low risk of contamination (see Section 5), it yields higher false alarm rates and lower hit rates compared to the four-anchor-SA method that was directly build with the SA-selection proposed by Wang (2004). Thus, the four-anchor-SA method is superior.

Based on these results, when no reliable prior knowledge about the DIF situation exists, as will be the case in most real data analysis settings (as opposed to simulation analysis where the true DIF pattern is known), we thus recommend to use the iterative-forward-SA method. When the sample size is large enough, the false alarm rates are low in any condition even if the anchor is contaminated. Hit rates rapidly grow with the sample size and converge to one. The forward item-wise selection of anchor items outperforms the iterative-backward-AO, iterative-forward-AO, the all-other as well as anchor methods from the constant anchor class.

There are several reasons that explain the superior performance of the iterative-forward-SA method. Firstly, the method has a head start compared to the methods that rely on DIF tests using the all-other items as anchor (e.g., the classical iterative procedures, such as

the iterative-backward-AO). While the latter start with a criterion that is severely biased when DIF is unbalanced, the iterative-forward-SA method does not require that DIF effects almost cancel out (for a detailed discussion of the assumption for the all-other method see Wang 2004). Secondly, the SA-selection strategy combined with the iterative forward anchor class also performs well in case of balanced DIF. While the AO-selection strategy performed better than the SA-selection strategy when it is combined with the methods from the constant anchor class, the advantage in combination with the iterative forward class appears negligible. Thirdly, our study showed that the consequences of contamination depend on the proportion of contaminated items rather than on the risk of contamination itself. Therefore, the iterative-forward-SA method yielded better results in DIF analysis even though the anchor is long and, thus, often contaminated. The risk of contamination decreases with increasing sample size and, beyond that, the proportion of DIF items in contaminated situations decreases. Fourthly, the iterative forward anchor class adds items to the anchor as long as the number of anchor items is smaller than the set of presumed DIF-free items. If the sample size is large enough, this leads to the desirable property, that it produces a longer anchor when the proportion of DIF items is low and a shorter anchor if the proportion of DIF items is high, similar to the iterative backward method. It may appear as a drawback that the iterative forward anchor class uses a short anchor in the initial steps, beginning with only one anchor item. The resulting DIF tests may lack statistical power due to fact that the anchor is short. However, this does not affect the performance of the new iterative forward anchor methods since the test results are only used for the decision whether the anchor should include one more anchor item and, thus, a small statistical power of the DIF tests in the first iterations automatically yields to a longer anchor that is expected to increase the power of the DIF test itself. Another astounding finding of our simulations was that anchor items located by an anchor selection strategy display different characteristics compared to randomly chosen DIF-free items and may be exactly those items that again induce artificial DIF. Including more anchor items (than, e.g., four anchor items) reduces the artificial scale shift that is induced by anchor items with empirical group differences and, thus, can also occur when the anchor is (by definition) pure.

Practical Recommendations

Our simulation study highlights the importance of the anchor selection for the correct classification of DIF and DIF-free items. A careful consideration of the anchor method used is necessary to avoid high misclassification rates and doubtful test results.

In case of balanced DIF, the all-other method is slightly better than the iterative-forward-SA strategy. However, due to the fact that the all-other method results in seriously inflated false alarm rates when the situation is unbalanced – and that it is doubtful whether the situation of balanced DIF is ever met in practice (Wang and Yeh 2003; Wang *et al.* 2012) – the usage of this anchor method is inadvisable. Thus, the newly suggested iterative-forward-SA strategy is recommended. When the sample size is large enough, the false alarm rates are low in any condition even if the anchor is contaminated and the hit rates grow rapidly.

The adequacy of the selection strategies – by single-anchor (SA) or by all-other (AO) – depends on the DIF situation. In the balanced condition, the AO-selection strategy performed suitable, whereas in the unbalanced condition the SA-selection strategy is more appropriate. But when the iterative-forward class is used, the advance of the AO-selection strategy is marginal. Therefore, we recommend the newly suggested iterative-forward-SA method over the iterative-forward-AO method.

Future research

The simulation study presented here was limited to DIF analysis using the Wald test in the Rasch model. Thus, future research may investigate the usefulness of the iterative-forward-SA method for other IRT models and combine it with other methods that test for DIF.

Furthermore, the iterative forward anchor class with the SA-selection may be compared with modifications of the anchor selection strategy. For example, Shih and Wang (2009) suggest to use the items corresponding to the lowest rank of the mean absolute DIF statistics similar to the rank-based strategy of Woods (2009). Then items are anchor candidates if they display the lowest mean DIF test statistic when every item is tested for DIF using every other item as constant single-anchor. This modification may be less affected by sample size.

Wang *et al.* (2012) established an improvement of the AO-selection strategy by incorporating additional iterations. Firstly, every item is tested for DIF using the all-other method. Then, iteratively, DIF items are excluded from the anchor candidates and a new DIF analysis using the current anchor is conducted until two steps reach the same results. Finally, the anchor items are selected from the remaining candidates using the rank-based strategy. Future research could compare the improved AO-selection to the SA-selection strategy.

Moreover, the DIF test results may also be improved by the construction of new anchor selection strategies. Ideally, the anchor items are DIF-free and induce no artificial scale shift. Furthermore, the impact of the degree of contamination is important for the appropriateness of the results in DIF detection. Therefore, improving the anchor selection strategies with the aim to locate anchors with a small degree of contamination remains an important task.

Acknowledgement

Julia Kopf is supported by the German Federal Ministry of Education and Research (BMBF) within the project “Heterogeneity in IRT-Models” (grant ID 01JG1060). The authors would like to thank Thomas Augustin for his expert advice.

References

Andrich D, Hagquist C (2012). "Real and Artificial Differential Item Functioning." *Journal of Educational and Behavioral Statistics*, **37**(3), 387–416.

Bolt DM, Hare RD, Vitale JE, Newman JP (2004). "A Multigroup Item Response Theory Analysis of the Psychopathy Checklist – Revised." *Psychological Assessment*, **16**(2), 155–168.

Candell GL, Drasgow F (1988). "An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory." *Applied Psychological Measurement*, **12**(3), 253–260.

Cheung GW, Rensvold RB (1999). "Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method." *Journal of Management*, **25**(1), 1–27.

Cohen AS, Kim SH, Wollack JA (1996). "An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning." *Applied Psychological Measurement*, **20**(1), 15–26.

Drasgow F (1987). "Study of the Measurement Bias of Two Standardized Psychological Tests." *Journal of Applied Psychology*, **72**(1), 19–29.

Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K (2006). "Identification of Differential Item Functioning Using Item Response Theory and the Likelihood-based Model Comparison Approach. Application to the Mini-Mental State Examination." *Medical Care*, **44**(22), 134–142.

Eggen T, Verhelst N (2006). "Loss of Information in Estimating Item Parameters in Incomplete Designs." *Psychometrika*, **71**(2), 303–322.

Finch H (2005). "The MIMIC Model As a Method for Detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio." *Applied Psychological Measurement*, **29**(4), 278–295.

Fischer GH (1995). "Derivations of the Rasch Model." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 2. Springer, New York.

Frederickx S, Tuerlinckx F, De Boeck P, Magis D (2010). "RIM: A Random Item Mixture Model to Detect Differential Item Functioning." *Journal of Educational Measurement*, **47**(4), 432–457.

Glas CAW, Verhelst ND (1995). "Testing the Rasch Model." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 5. Springer, New York.

González-Betanzos F, Abad FJ (2012). "The Effects of Purification and the Evaluation of Differential Item Functioning with the Likelihood Ratio Test." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **8**(4), 134–145.

Hidalgo-Montesinos MD, Lopez-Pina JA (2002). "Two-Stage Equating in Differential Item Functioning Detection under the Graded Response Model with the Raju Area Measures and the Lord Statistic." *Educational and Psychological Measurement*, **62**(1), 32–44.

Holland PW, Thayer DT (1988). "Differential Item Performance and the Mantel-Haenszel Procedure." In H Wainer, HI Braun (eds.), *Test Validity*, chapter 9. Lawrence Erlbaum, Hillsdale, New Jersey.

Jodoin MG, Gierl MJ (2001). "Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection." *Applied Measurement in Education*, **14**(4), 329–349.

Kim SH, Cohen AS (1995). "A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning." *Applied Measurement in Education*, **8**(4), 291–312.

Kim SH, Cohen AS (1998). "Detection of Differential Item Functioning under the Graded Response Model with the Likelihood Ratio Test." *Applied Psychological Measurement*, **22**(4), 345–355.

Lopez Rivas GE, Stark S, Chernyshenko OS (2009). "The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test." *Applied Psychological Measurement*, **33**(4), 251–265.

Lord F (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, New Jersey.

Mellenbergh GJ (1982). "Contingency Table Models for Assessing Item Bias." *Journal of Educational Statistics*, **7**(2), 105–118.

Millsap RE, Everson HT (1993). "Methodology Review: Statistical Approaches for Assessing Measurement Bias." *Applied Psychological Measurement*, **17**(4), 297–334.

Molenaar IW (1995). "Estimation of Item Parameters." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 3. Springer, New York.

Park DG, Lautenschlager GJ (1990). "Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification." *Applied Psychological Measurement*, **14**(2), 163–173.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shealy R, Stout W (1993). "A Model-based Standardization Approach That Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF." *Psychometrika*, **58**(2), 159–194.

Shih CL, Wang WC (2009). "Differential Item Functioning Detection Using the Multiple Indicators, Multiple Causes Method with a Pure Short Anchor." *Applied Psychological Measurement*, **33**(3), 184–199.

Stark S, Chernyshenko OS, Drasgow F (2006). "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology*, **91**(6), 1292–1306.

Strobl C, Kopf J, Zeileis A (2010). "Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben." In S Trepte, M Verbeet (eds.), *Wissenswelten des 21. Jahrhunderts – Erkenntnisse aus dem Studentenpisa-Test des SPIEGEL*. VS Verlag, Wiesbaden.

Swaminathan H, Rogers HJ (1990). "Detecting Differential Item Functioning Using Logistic Regression Procedures." *Journal of Educational Measurement*, **27**(4), 361–370.

Thissen D (2001). *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. Unpublished manuscript, University of North Carolina, Chapel Hill.

Thissen D, Steinberg L, Wainer H (1988). "Use of Item Response Theory in the Study of Group Differences in Trace Lines." In H Wainer, HI Braun (eds.), *Test Validity*, chapter 10. Lawrence Erlbaum, Hillsdale, New Jersey.

Trepte S, Verbeet M (eds.) (2010). *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL Studentenpisa-Test*. VS Verlag, Wiesbaden.

Wang WC (2004). "Effects of Anchor Item Methods on the Detection of Differential Item Functioning within the Family of Rasch Models." *Journal of Experimental Education*, **72**(3), 221–261.

Wang WC, Shih CL, Sun GW (2012). "The DIF-Free-Then-DIF Strategy for the Assessment of Differential Item Functioning." *Educational and Psychological Measurement*, **72**(4), 687–708.

Wang WC, Su YH (2004). "Effects of Average Signed Area Between Two Item Characteristic Curves and Test Purification Procedures on the DIF Detection via the Mantel-Haenszel Method." *Applied Measurement in Education*, **17**(2), 113–144.

Wang WC, Yeh YL (2003). "Effects of Anchor Item Methods on Differential Item Functioning Detection with the Likelihood Ratio Test." *Applied Psychological Measurement*, **27**(6), 479–498.

Woods CM (2009). "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement*, **33**(1), 42–57.

Appendix

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
no DIF									
250,250	0.05	0.00	0.03	0.05	0.05	0.00	0.03	0.04	0.05
	(0.03)	(0.01)	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	(0.04)	(0.03)
500, 500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.03)	(0.01)	(0.03)	(0.04)	(0.03)
750, 750	0.05	0.00	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.03)	(0.01)	(0.03)	(0.04)	(0.04)
1000, 1000	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.03)	(0.01)	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	(0.04)	(0.03)
1250, 1250	0.05	0.00	0.03	0.05	0.05	0.00	0.03	0.04	0.05
	(0.03)	(0.01)	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	(0.04)	(0.03)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.04	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.03)	(0.01)	(0.03)	(0.04)	(0.04)

Table 2: False alarm rates and standard errors under the null hypothesis (no DIF) with equal sample sizes in reference and focal group.

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250,250	0.05	0.01	0.03	0.05	0.05	0.00	0.04	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)
500, 500	0.05	0.00	0.03	0.05	0.05	0.01	0.04	0.07	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.06)	(0.04)
750, 750	0.05	0.01	0.03	0.05	0.05	0.01	0.05	0.07	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.06)	(0.04)
1000, 1000	0.05	0.01	0.03	0.05	0.05	0.00	0.04	0.07	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.06)	(0.04)
1250, 1250	0.05	0.01	0.03	0.05	0.05	0.00	0.04	0.06	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.05)	(0.04)
1500, 1500	0.05	0.00	0.03	0.05	0.05	0.00	0.03	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)
0.30									
250,250	0.05	0.00	0.03	0.05	0.05	0.00	0.04	0.05	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.05)	(0.04)
500, 500	0.05	0.01	0.03	0.05	0.05	0.01	0.05	0.08	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.05)	(0.07)	(0.04)
750, 750	0.05	0.00	0.03	0.05	0.05	0.02	0.07	0.11	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.08)	(0.04)
1000, 1000	0.05	0.00	0.03	0.05	0.05	0.01	0.06	0.09	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.05)	(0.08)	(0.04)
1250, 1250	0.05	0.00	0.03	0.05	0.05	0.01	0.04	0.08	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.04)	(0.07)	(0.04)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.00	0.03	0.06	0.05
	(0.04)	(0.01)	(0.03)	(0.04)	(0.04)	(0.01)	(0.04)	(0.05)	(0.04)
0.45									
250,250	0.05	0.00	0.03	0.05	0.05	0.00	0.04	0.05	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.01)	(0.04)	(0.06)	(0.05)
500, 500	0.05	0.00	0.03	0.05	0.05	0.02	0.07	0.12	0.05
	(0.05)	(0.01)	(0.03)	(0.05)	(0.05)	(0.04)	(0.06)	(0.11)	(0.05)
750, 750	0.05	0.01	0.03	0.05	0.05	0.05	0.09	0.16	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.04)	(0.06)	(0.07)	(0.13)	(0.05)
1000, 1000	0.05	0.00	0.03	0.05	0.05	0.04	0.08	0.14	0.05
	(0.05)	(0.02)	(0.03)	(0.05)	(0.05)	(0.06)	(0.07)	(0.12)	(0.05)
1250, 1250	0.05	0.00	0.03	0.05	0.05	0.02	0.05	0.10	0.05
	(0.05)	(0.01)	(0.03)	(0.05)	(0.04)	(0.05)	(0.05)	(0.09)	(0.05)
1500, 1500	0.05	0.01	0.03	0.05	0.05	0.01	0.04	0.07	0.05
	(0.05)	(0.02)	(0.04)	(0.05)	(0.05)	(0.02)	(0.04)	(0.07)	(0.05)

Table 3: False alarm rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.

hit rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250,250	0.68	0.35	0.58	0.67	0.66	0.23	0.58	0.59	0.67
	(0.19)	(0.20)	(0.20)	(0.19)	(0.19)	(0.15)	(0.20)	(0.20)	(0.19)
500, 500	0.93	0.75	0.89	0.92	0.92	0.56	0.86	0.85	0.92
	(0.10)	(0.18)	(0.12)	(0.11)	(0.11)	(0.15)	(0.13)	(0.14)	(0.11)
750, 750	0.99	0.92	0.98	0.99	0.98	0.75	0.95	0.94	0.98
	(0.04)	(0.11)	(0.06)	(0.05)	(0.05)	(0.15)	(0.09)	(0.10)	(0.05)
1000, 1000	1.00	0.98	0.99	1.00	1.00	0.88	0.98	0.98	1.00
	(0.02)	(0.06)	(0.03)	(0.03)	(0.03)	(0.13)	(0.05)	(0.06)	(0.03)
1250, 1250	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.08)	(0.02)	(0.03)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.05)	(0.01)	(0.02)	(0.00)
0.30									
250,250	0.68	0.36	0.58	0.67	0.66	0.26	0.58	0.58	0.66
	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)	(0.11)	(0.14)	(0.13)	(0.13)
500, 500	0.93	0.74	0.89	0.92	0.92	0.56	0.83	0.81	0.92
	(0.08)	(0.13)	(0.09)	(0.08)	(0.08)	(0.10)	(0.11)	(0.11)	(0.08)
750, 750	0.99	0.92	0.97	0.98	0.98	0.72	0.92	0.90	0.98
	(0.03)	(0.08)	(0.05)	(0.04)	(0.04)	(0.11)	(0.08)	(0.09)	(0.04)
1000, 1000	1.00	0.98	1.00	1.00	1.00	0.86	0.98	0.97	1.00
	(0.01)	(0.04)	(0.02)	(0.01)	(0.01)	(0.10)	(0.04)	(0.05)	(0.02)
1250, 1250	1.00	0.99	1.00	1.00	1.00	0.94	0.99	0.99	1.00
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.08)	(0.02)	(0.03)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.04)	(0.01)	(0.01)	(0.00)
0.45									
250,250	0.68	0.36	0.59	0.67	0.66	0.28	0.58	0.58	0.66
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.09)	(0.11)	(0.11)	(0.11)
500, 500	0.93	0.74	0.89	0.92	0.91	0.55	0.80	0.77	0.91
	(0.06)	(0.10)	(0.07)	(0.07)	(0.07)	(0.07)	(0.10)	(0.11)	(0.07)
750, 750	0.99	0.92	0.97	0.98	0.98	0.68	0.90	0.86	0.98
	(0.03)	(0.06)	(0.04)	(0.03)	(0.03)	(0.09)	(0.09)	(0.10)	(0.04)
1000, 1000	1.00	0.98	1.00	1.00	1.00	0.82	0.97	0.95	1.00
	(0.01)	(0.03)	(0.02)	(0.01)	(0.01)	(0.10)	(0.05)	(0.07)	(0.02)
1250, 1250	1.00	1.00	1.00	1.00	1.00	0.92	0.99	0.99	1.00
	(0.00)	(0.02)	(0.01)	(0.01)	(0.01)	(0.09)	(0.02)	(0.04)	(0.01)
1500, 1500	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.04)	(0.01)	(0.02)	(0.00)

Table 4: Hit rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.

false alarm rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250,250	0.07	0.01	0.04	0.06	0.05	0.01	0.06	0.07	0.06
	(0.04)	(0.02)	(0.03)	(0.04)	(0.04)	(0.02)	(0.04)	(0.06)	(0.04)
500, 500	0.08	0.01	0.05	0.05	0.05	0.02	0.08	0.10	0.05
	(0.04)	(0.02)	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.07)	(0.04)
750, 750	0.10	0.02	0.06	0.05	0.05	0.02	0.07	0.10	0.05
	(0.04)	(0.02)	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.07)	(0.04)
1000, 1000	0.11	0.02	0.07	0.05	0.05	0.01	0.05	0.08	0.05
	(0.05)	(0.02)	(0.04)	(0.04)	(0.04)	(0.02)	(0.04)	(0.06)	(0.04)
1250, 1250	0.13	0.02	0.08	0.05	0.05	0.00	0.04	0.06	0.05
	(0.05)	(0.03)	(0.04)	(0.04)	(0.04)	(0.01)	(0.04)	(0.05)	(0.04)
1500, 1500	0.15	0.03	0.09	0.05	0.05	0.00	0.03	0.05	0.05
	(0.05)	(0.03)	(0.05)	(0.04)	(0.04)	(0.01)	(0.03)	(0.05)	(0.04)
0.30									
250,250	0.11	0.02	0.07	0.08	0.07	0.02	0.10	0.12	0.08
	(0.05)	(0.03)	(0.04)	(0.06)	(0.06)	(0.03)	(0.06)	(0.08)	(0.06)
500, 500	0.18	0.04	0.11	0.09	0.07	0.05	0.15	0.18	0.07
	(0.06)	(0.03)	(0.05)	(0.06)	(0.06)	(0.05)	(0.07)	(0.12)	(0.05)
750, 750	0.25	0.06	0.15	0.09	0.07	0.06	0.15	0.18	0.06
	(0.07)	(0.05)	(0.07)	(0.06)	(0.05)	(0.06)	(0.08)	(0.14)	(0.05)
1000, 1000	0.31	0.09	0.19	0.09	0.08	0.04	0.10	0.14	0.05
	(0.07)	(0.05)	(0.08)	(0.06)	(0.05)	(0.06)	(0.07)	(0.11)	(0.04)
1250, 1250	0.38	0.12	0.22	0.08	0.08	0.01	0.05	0.08	0.05
	(0.07)	(0.06)	(0.09)	(0.06)	(0.06)	(0.03)	(0.05)	(0.08)	(0.04)
1500, 1500	0.44	0.14	0.24	0.08	0.10	0.00	0.04	0.06	0.05
	(0.07)	(0.08)	(0.10)	(0.06)	(0.06)	(0.02)	(0.04)	(0.06)	(0.04)
0.45									
250,250	0.20	0.04	0.13	0.18	0.17	0.03	0.15	0.17	0.18
	(0.07)	(0.04)	(0.07)	(0.09)	(0.10)	(0.04)	(0.08)	(0.11)	(0.10)
500, 500	0.34	0.11	0.24	0.28	0.24	0.10	0.26	0.29	0.24
	(0.08)	(0.07)	(0.09)	(0.12)	(0.15)	(0.09)	(0.11)	(0.18)	(0.14)
750, 750	0.48	0.18	0.33	0.37	0.25	0.17	0.32	0.38	0.22
	(0.09)	(0.10)	(0.12)	(0.13)	(0.18)	(0.13)	(0.14)	(0.27)	(0.16)
1000, 1000	0.60	0.24	0.39	0.45	0.27	0.19	0.25	0.36	0.14
	(0.09)	(0.15)	(0.16)	(0.15)	(0.20)	(0.19)	(0.16)	(0.32)	(0.15)
1250, 1250	0.69	0.28	0.42	0.51	0.28	0.10	0.11	0.20	0.07
	(0.08)	(0.20)	(0.18)	(0.16)	(0.20)	(0.17)	(0.13)	(0.26)	(0.09)
1500, 1500	0.78	0.31	0.45	0.57	0.31	0.02	0.05	0.09	0.06
	(0.08)	(0.25)	(0.21)	(0.17)	(0.23)	(0.09)	(0.06)	(0.13)	(0.06)

Table 5: False alarm rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.

hit rate	all-other	single-anchor-AO	four-anchor-AO	iterative-forward-AO	iterative-backw.-AO	single-anchor-SA	four-anchor-SA	four-anchor-NC	iterative-forward-SA
0.15									
250,250	0.54	0.24	0.44	0.59	0.59	0.11	0.37	0.37	0.60
	(0.20)	(0.17)	(0.20)	(0.22)	(0.22)	(0.13)	(0.19)	(0.21)	(0.22)
500, 500	0.83	0.55	0.76	0.89	0.89	0.32	0.67	0.66	0.89
	(0.15)	(0.20)	(0.17)	(0.14)	(0.14)	(0.22)	(0.19)	(0.21)	(0.14)
750, 750	0.95	0.79	0.91	0.98	0.98	0.61	0.88	0.87	0.98
	(0.09)	(0.16)	(0.11)	(0.06)	(0.06)	(0.25)	(0.14)	(0.16)	(0.06)
1000, 1000	0.98	0.90	0.97	0.99	0.99	0.83	0.97	0.96	0.99
	(0.05)	(0.12)	(0.07)	(0.03)	(0.03)	(0.20)	(0.07)	(0.08)	(0.03)
1250, 1250	0.99	0.96	0.99	1.00	1.00	0.94	1.00	0.99	1.00
	(0.03)	(0.07)	(0.04)	(0.01)	(0.02)	(0.11)	(0.03)	(0.04)	(0.01)
1500, 1500	1.00	0.99	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	(0.01)	(0.05)	(0.02)	(0.01)	(0.01)	(0.06)	(0.01)	(0.02)	(0.01)
0.30									
250,250	0.40	0.14	0.31	0.47	0.48	0.06	0.26	0.27	0.48
	(0.13)	(0.10)	(0.13)	(0.17)	(0.18)	(0.08)	(0.13)	(0.15)	(0.18)
500, 500	0.67	0.35	0.59	0.81	0.82	0.22	0.52	0.51	0.84
	(0.12)	(0.13)	(0.14)	(0.13)	(0.14)	(0.15)	(0.16)	(0.20)	(0.13)
750, 750	0.84	0.57	0.79	0.95	0.95	0.47	0.79	0.76	0.97
	(0.10)	(0.14)	(0.12)	(0.07)	(0.07)	(0.21)	(0.14)	(0.19)	(0.06)
1000, 1000	0.92	0.73	0.90	0.99	0.98	0.73	0.94	0.92	0.99
	(0.08)	(0.13)	(0.09)	(0.04)	(0.04)	(0.21)	(0.09)	(0.11)	(0.02)
1250, 1250	0.96	0.85	0.96	1.00	0.99	0.91	0.99	0.98	1.00
	(0.05)	(0.11)	(0.06)	(0.02)	(0.02)	(0.13)	(0.03)	(0.05)	(0.01)
1500, 1500	0.98	0.92	0.98	1.00	1.00	0.97	1.00	1.00	1.00
	(0.04)	(0.08)	(0.04)	(0.01)	(0.02)	(0.06)	(0.01)	(0.02)	(0.00)
0.45									
250,250	0.27	0.07	0.20	0.29	0.29	0.04	0.19	0.20	0.30
	(0.09)	(0.06)	(0.09)	(0.13)	(0.14)	(0.05)	(0.09)	(0.12)	(0.14)
500, 500	0.48	0.19	0.40	0.54	0.56	0.14	0.37	0.38	0.58
	(0.10)	(0.10)	(0.13)	(0.16)	(0.20)	(0.11)	(0.14)	(0.20)	(0.18)
750, 750	0.64	0.34	0.60	0.73	0.77	0.29	0.59	0.56	0.83
	(0.10)	(0.15)	(0.15)	(0.14)	(0.18)	(0.20)	(0.18)	(0.28)	(0.15)
1000, 1000	0.76	0.51	0.76	0.85	0.87	0.49	0.82	0.72	0.95
	(0.09)	(0.20)	(0.15)	(0.11)	(0.14)	(0.28)	(0.17)	(0.32)	(0.08)
1250, 1250	0.85	0.66	0.87	0.92	0.93	0.78	0.96	0.90	0.99
	(0.07)	(0.23)	(0.13)	(0.08)	(0.11)	(0.26)	(0.08)	(0.23)	(0.03)
1500, 1500	0.90	0.78	0.93	0.95	0.95	0.94	1.00	0.98	1.00
	(0.06)	(0.23)	(0.10)	(0.07)	(0.11)	(0.13)	(0.02)	(0.10)	(0.01)

Table 6: Hit rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.

Affiliation:

Julia Kopf
Graduate Researcher
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstraße 33
DE-80539 München, Germany
E-mail: Julia.Kopf@stat.uni-muenchen.de

Prof. Dr. Achim Zeileis
Department of Statistics
Universität Innsbruck
Universitätsstraße 15
AT-6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org

Prof. Dr. Carolin Strobl
Department of Psychology
Universität Zürich
Binzmühlestrasse 14
CH-8050 Zürich, Switzerland
E-mail: Carolin.Strobl@psychologie.uzh.ch