LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Julia Kopf, Achim Zeileis, Carolin Strobl

Anchor selection strategies for DIF analysis: Review, assessment, and new approaches

Technical Report Number 150, 2013
Department of Statistics
University of Munich

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

Anchor selection strategies for DIF analysis: Review, assessment, and new approaches

Julia Kopf
LMU München

Achim Zeileis
Universität Innsbruck

Carolin Strobl
UZH Zürich

Abstract

Differential item functioning (DIF) indicates the violation of the invariance assumption for instance in models based on item response theory (IRT). For item-wise DIF analysis using IRT, a common metric for the item parameters of the groups that are to be compared (e.g. for the reference and the focal group) is necessary. In the Rasch model, therefore, the same linear restriction is imposed in both groups. Items in the restriction are termed the *anchor items*. Ideally, these items are DIF-free to avoid artificially augmented false alarm rates. However, the question how DIF-free anchor items are selected appropriately is still a major challenge. Furthermore, various authors point out the lack of new anchor selection strategies and the lack of a comprehensive study especially for dichotomous IRT models. This article reviews existing anchor selection strategies that do not require any knowledge prior to DIF analysis, offers a straightforward notation and proposes three new anchor selection strategies. An extensive simulation study is conducted to compare the performance of the anchor selection strategies. The results show that an appropriate anchor selection is crucial for suitable item-wise DIF analysis. The newly suggested anchor selection strategies outperform the existing strategies and can reliably locate a suitable anchor when the sample sizes are large enough.

Keywords: Rasch model, differential item functioning (DIF), anchor selection, anchor class, uniform DIF, measurement invariance.

1. Introduction

Differential item functioning (DIF) is present if test-takers from different groups – such as male and female test-takers – display different probabilities of solving an item even if they have the same latent trait. In this case, the test results no longer represent the ability alone and the groups of test-takers cannot be compared in an objective, fair way.

Various methods have been suggested to analyze item-wise DIF (see Millsap and Everson 1993, for an overview). DIF tests based on item response theory (IRT) such as the item-wise Wald test (see, e.g., Glas and Verhelst 1995) rely on the comparison of the estimated item parameters of the underlying IRT model. For this purpose, *anchor methods* are employed to place the estimated item parameters onto a common scale.

Previous studies showed that a careful consideration of the anchor method is crucial for suitable DIF analysis: If the anchor contains DIF items, which is referred to as *contamination* (see, e.g., Finch 2005; Woods 2009; Wang, Shih, and Sun 2012), the construction of a common scale for the item parameters may fail and seriously increased false alarm rates can result (see, e.g., Wang and Yeh 2003; Wang 2004; Wang and Su 2004; Finch 2005; Stark, Chernyshenko, and Drasgow 2006; Woods 2009; Kopf, Zeileis, and Strobl 2013). Thus, items truly free of DIF may appear to have DIF and jeopardize the results of the DIF analysis as well as the associated investigation of the causes of DIF (Jodoin and Gierl 2001). One alternative to reduce the risk of a contaminated anchor is to employ a short anchor that should be easier to find from the set of DIF-free items. However, the statistical power to detect DIF increases with the length of the (DIF-free) anchor (Thissen, Steinberg, and Wainer 1988; Wang and Yeh 2003; Wang 2004; Shih and Wang 2009; Woods 2009; Kopf *et al.* 2013).

In the literature, one can find both methods that do and methods that do not require an explicit anchor selection. While at first sight it may seem that methods that do not require an anchor selection strategy have an advantage, it has been shown that there are situations where these methods are not suitable for DIF detection. The *all-other anchor method*, for example, uses all items except for the currently studied item as anchor (see, e.g., Cohen, Kim, and Wollack 1996; Kim and Cohen 1998) and, thus, requires no anchor selection strategy. However, the method was shown to be inadvisable for DIF detection when the test contains DIF items that favor one group (Wang and Yeh 2003; Wang 2004). Excluding DIF items from the anchor by using iterative steps does not solve the problem when the test contains many DIF items (Wang *et al.* 2012). In practice, there is usually no prior knowledge about the exact composition of the DIF effects and, thus, it is advisable to use an anchor method that relies on an explicit anchor selection strategy such as an anchor of the constant length of four items (used, e.g., by Thissen *et al.* 1988; Wang 2004; Shih and Wang 2009). An anchor selection strategy then guides the decision which particular items are used as anchor items.

Several anchor selection strategies have already been proposed, some of which rely on prior knowledge of a set of DIF-free items or on the advice of content experts, while others are based on preliminary item analysis (for an overview see Woods 2009). Here, only those strategies that do not require any information prior to data analysis, such as the knowledge of certain DIF-free items, will be reviewed and presented in a straightforward notation in Section 3.2. The reason for excluding strategies that require prior knowledge about DIF-free items from this review is that in practical testing situations sets of truly DIF-free items are most likely unknown (as opposed to simulation analysis, where the true DIF pattern is known) and even the judgment of content experts is unreliable (for a literature overview where this approach

fails see Frederickx, Tuerlinckx, De Boeck, and Magis 2010). New suggestions of anchor selection strategies are often only compared to few alternative strategies or in situations of only a limited range of the sample size and "have not been exhaustively compared for the dichotomous case" (González-Betanzos and Abad 2012, p. 2). Therefore, in this article, we systematically evaluate the performance of the existing anchor selection strategies for DIF analysis in the Rasch model by conducting an extensive simulation study.

Furthermore, we assess the appropriateness of the anchor selection strategies to find a suitable short anchor (of four anchor items) and also their ability to select a suitable longer anchor, which "is a challenging question for researchers and practitioners" (Wang *et al.* 2012, p. 19). For practical research, recommendations how anchor items can be found appropriately are still required (Lopez Rivas, Stark, and Chernyshenko 2009, p. 252). We also provide guidelines how to choose anchor items when no prior knowledge of DIF-free items is at hand.

In addition to the existing strategies, new developments of anchor selection strategies have also been encouraged (Wang *et al.* 2012, p. 19). Therefore, we also suggest three new anchor selection strategies. The new anchor selection strategies are implemented and the results show an improvement of the classification accuracy in the analysis of DIF.

The article is organized as follows. The technical aspects of the anchoring process in the Rasch model are introduced in the next section. Details of the anchor classes and of the existing as well as of the newly suggested anchor selection strategies are given in Section 3. The simulation design is addressed in Section 4 and the results are discussed in Section 5. A concluding summary and practical recommendations are presented in Section 6.

2. Model and notation

In this section, the model and notation are introduced along with some technical statistical details about the anchoring process: (1) how parameter estimates under certain restrictions can be obtained and (2) how the associated item-wise parameter differences between a focal and reference group can be assessed given a selection of anchor items. In our discussion we focus on the Rasch model but the underlying ideas can also be applied to other related IRT models.

Based on the resulting item-wise tests, the subsequent sections will then discuss how the tests can be combined employing a wide range of classes of anchors and different strategies for selecting the anchor items.

2.1. Model estimation and scale indeterminacy

To fix notation, we employ the Rasch model with item parameter vector $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$ (where k denotes the number of items in the test). It is usually estimated using the conditional maximum likelihood (CML) approach, due to its unique statistical properties, its widespread application (Wang 2004) and the fact that it does not rely on the person parameters (Molenaar 1995).

To overcome the *scale indeterminacy* (Fischer 1995) of the item parameters β, one linear restriction is typically imposed on them. Hence, only $k - 1$ parameters can be freely estimated and the remaining one parameter is determined by the restriction. Commonly-used approaches restrict a set $\mathcal{A} \subseteq \{1, \ldots, k\}$ of one or more or all item parameters to sum to

zero $\sum_{\ell \in \mathcal{A}} \beta_\ell = 0$ (Eggen and Verhelst 2006). These types of restrictions can be equivalently represented by a vector a indicating the items that should sum to zero, i.e., $\sum_{\ell=1}^{k} a_\ell \beta_\ell = 0$. For example, if the first and second item parameter should sum to zero, the set is $\mathcal{A} = \{1, 2\}$ with associated indicator vector $a = (1, 1, 0, 0, 0, \ldots)^\top$.

Conveniently, the item parameter estimates $\hat{\beta}$ under any such restriction can be easily obtained from any other set of parameter estimates $\tilde{\beta}$ fulfilling another restriction. Hence, without loss of generality we employ the restriction $\tilde{\beta}_1 = 0$ for the initial CML parameter estimates and also obtain the corresponding covariance matrix $\widehat{\mathrm{Var}}(\tilde{\beta})$ which consequently has zero entries in the first row and in the first column. To obtain any other restriction of the sum type above, the item parameter estimates $\hat{\beta}$ and corresponding covariance matrix estimate can be obtained as follows:

$$\hat{\beta} = A\tilde{\beta} \tag{1}$$

$$\widehat{\mathrm{Var}}(\hat{\beta}) = A\widehat{\mathrm{Var}}(\tilde{\beta})A^\top, \tag{2}$$

where $A = I_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot a^\top$ is the contrast matrix corresponding to the indicator vector a with I_k denoting the identity matrix and $1_k = (1, 1, \ldots, 1)^\top$ a vector of one entries of length k. To emphasize that the parameter estimates $\hat{\beta}$ depend on the set of restricted item parameters, we sometimes employ the notation $\hat{\beta}(\mathcal{A})$ in the following (although the dependence on $\mathcal{A}$ is mostly suppressed).

2.2. Item-wise parameter differences

In DIF analysis using IRT models, groups are to be compared regarding their item parameters. We focus here on the situation of item-wise comparisons between two groups (reference and focal). In order to establish a common scale for the item parameters the same linear restriction

$$\sum_{\ell \in \mathcal{A}} \hat{\beta}_\ell^g = 0 \qquad (g \in \{\mathrm{ref}, \mathrm{foc}\}) \tag{3}$$

has to be imposed on the item parameters in both groups (Glas and Verhelst 1995). Thus, $\mathcal{A}$ is the set of *anchor items* employed to align the scales between the two groups g.

More specifically, to assess differences between the two groups for the j-th item parameter β_j ($j = 1, \ldots, k$), the following steps are carried out:

1. Obtain the intial CML estimates $\tilde{\beta}^g$ in both groups g (i.e., using the restriction $\tilde{\beta}_1^g = 0$).
2. Based on the same set of anchor items $\mathcal{A}$, compute $\hat{\beta}^g = \hat{\beta}^g(\mathcal{A})$ and corresponding $\widehat{\mathrm{Var}}(\hat{\beta}^g)$ using Equations 1 and 2 so that Equation 3 holds in both groups g.
3. Carry out an item-wise Wald test (see, e.g., Glas and Verhelst 1995) for the j-th item with test statistic $t_j = t_j(\mathcal{A})$ given by

$$t_j = \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}})}} = \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{ref}})_{j,j} + \widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{foc}})_{j,j}}}. \tag{4}$$

Either the test statistic t_j or the associated p-value p_j can then be employed as a DIF index

because under the null hypothesis of no DIF the item parameters from both groups should be equal: $\beta_j^{\text{ref}} = \beta_j^{\text{foc}}$.

Note that this item-wise Wald test is applied to the CML estimates (as in Glas and Verhelst 1995) and not the joint maximum likelihood (JML) estimates (as in Lord 1980). The inconsistency of the JML estimates leads to highly inflated false alarm rates (McLaughlin and Drasgow 1987; Lim and Drasgow 1990). In case other IRT models are regarded, the recent work of Woods, Cai, and Wang (2012) showed that an improved version of the Wald test, termed Wald-1 (see Paek and Han 2013, and the references therein), also displayed well-controlled false alarm rates if the anchor items were DIF-free. Since the Wald-1 test also requires anchor items, it can in principle be combined with the anchor methods discussed here as well.

3. Anchor methods

Under this null hypothesis of equality between *all* item parameters, in principle any set of items could be chosen for the anchor $\mathcal{A}$. However, under the alternative that some of the k item parameters are actually affected by DIF, the results of the analysis strongly depend on the choice of the anchor items, as previous studies illustrated. If the anchor contains at least one DIF item, it is referred to as *contaminated* (see, e.g., Finch 2005; Woods 2009; Wang *et al.* 2012). The scales may then be artificially shifted apart and the false alarm rates of the DIF tests may be seriously inflated (see, e.g., Wang and Yeh 2003, Wang 2004, Wang and Su 2004, Finch 2005, Stark *et al.* 2006, Woods 2009). Instructive examples that illustrate the artificial scale shift are provided by Wang (2004) and Kopf *et al.* (2013).

For distinguishing between the different approaches, we employ a framework for anchor methods previously used in Kopf *et al.* (2013) where the *anchor class* determines characteristics of the anchor methods, such as a predefined anchor length, and the *anchor selection strategy* guides the decision which items are used as anchor items. The combination of an anchor class together with an anchor selection strategy is then termed an *anchor method*. Different anchor classes are now briefly reviewed.

3.1. Anchor classes

The *constant anchor class* consists of an anchor with a predefined, constant length. Usually, it is claimed that a constant anchor of four items assures sufficient power (cf. e.g., Shih and Wang 2009; Wang *et al.* 2012). An anchor selection strategy is needed to guide the decision which items are used as anchor items. The *all-other anchor class* uses all items except for the currently studied item as anchor and the *equal-mean difficulty anchor class* uses all items as anchor (see, e.g., Wang 2004, and the references therein). These latter two anchor classes do not require an additional anchor selection strategy. Furthermore, iterative anchor classes build the anchor in an iterative manner. The *iterative backward class* (used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002) starts with all other items as anchor and excludes DIF items from the anchor, whereas the *iterative forward anchor class* starts with a single anchor item and then, iteratively, includes items in the anchor (Kopf *et al.* 2013). The latter anchor class also requires an explicit anchor selection strategy.

Wang (2004), Wang and Yeh (2003) and González-Betanzos and Abad (2012) compared the *all-other* and the *equal-mean difficulty anchor class* to different versions of the *constant anchor*

class regarding various IRT models. All methods from the *constant anchor class* were built using prior knowledge about the set of DIF-free items to locate the anchor items. Methods from the *constant anchor class* yielded well-controlled false alarm rates, whereas methods from the *all-other* and the *equal-mean difficulty anchor class* displayed seriously inflated false alarm rates when the direction of DIF was unbalanced (i.e. the DIF effects did not cancel out between groups and one group was favored in the test) and it is doubtful whether the situation of balanced DIF (i.e. no group has an advantage in the test) is met in practice (Wang and Yeh 2003; Wang *et al.* 2012). This is of utmost importance for practical testing situations, since items truly free of DIF display artificial DIF and may be eliminated by mistake.

Thus, all three studies showed that the direction of DIF has a major impact on the results of the DIF analysis for the *all-other* and the *equal-mean difficulty anchor class* as opposed to the *constant anchor class* based on DIF-free anchor items. Thus, the *constant anchor class* is in principle able to yield appropriate results for the DIF analysis even if DIF is unbalanced. However, since Wang and Yeh (2003), Wang (2004) and González-Betanzos and Abad (2012) used prior knowledge of the set of DIF-free items to select the constant anchor items, no information is yet available on how well anchor selection strategies without prior knowledge perform and "[f]urther research is needed to investigate how to locate anchor items correctly and efficiently" (Wang and Yeh 2003, p. 496).

Another anchor class was recently suggested by Kopf *et al.* (2013). Instead of a predefined anchor length, the *iterative forward anchor class* builds the anchor in a step-by-step procedure. First, one anchor item is used for the initial DIF test. As long as the current anchor length is shorter than the number of items currently not displaying statistically significant DIF (termed the presumed DIF-free items in the following), one item is added to the current anchor and DIF analysis is conducted using the new current anchor. The sequence which item is first included and which items are added to the anchor is determined by an anchor selection strategy. In a simulation study, the *iterative forward anchor class* and the *constant anchor class* were combined with two different anchor selection strategies and compared to the *all-other class* and the *iterative backward anchor class*. The *iterative forward anchor class* was found to be superior since it yielded high hit rates and, simultaneously, low false alarm rates for sufficiently large sample sizes in any studied condition of balanced or unbalanced DIF if the *number of significant threshold anchor selection strategy* (see Section 3.2) was employed (Kopf *et al.* 2013).

To assess the appropriateness of the anchor selection strategies in this article, we combine them with the *constant four anchor class* and the *iterative forward anchor class*. The reason for this is that both classes require an anchor selection strategy and it is claimed that they assure sufficient power when the anchor selection works adequately. Furthermore, both classes are structurally different. The *constant four anchor class* always includes four anchor items, and, thus, leads to a short anchor, whereas the *iterative forward class* allows for a longer anchor that is built in an iterative way. The *all-other*, the *equal-mean difficulty* and the *iterative backward class* displayed seriously inflated false alarm rates when the direction of DIF is unbalanced (Wang and Yeh 2003; Wang 2004; González-Betanzos and Abad 2012; Kopf *et al.* 2013), and are, thus, not considered as anchor classes in this article.

3.2. Anchor selection strategies

Anchor selection strategies determine a ranking order of candidate anchor items. We focus

on those strategies that are based on preliminary item analysis since these strategies are most common in practice. This approach has been referred to as the DIF-free-then-DIF strategy by Wang *et al.* (2012) because auxiliary DIF tests are conducted to locate (ideally DIF-free) anchor items before the final DIF tests are carried out.

Auxiliary DIF tests

For each item $j = 1, \ldots, k$, auxiliary DIF tests are conducted using step 1 to 3 in Section 2.2. Typically, there are two alternative ways to conduct auxiliary DIF tests:

(I) The auxiliary DIF tests are conducted using all-other items $\{1, \ldots, k\} \backslash j$ as anchor. This yields to one observed test statistic $t_j(\{1, \ldots, k\} \backslash j)$ for every currently studied item j.

(II) The auxiliary DIF tests are conducted using every other item $\ell \neq j$ as constant single anchor. This results in $(k-1)$ test statistics per item $t_j(\{\ell\})$ with the corresponding p-values $p_j(\{\ell\})$. Anchor selection strategies decide how all tests are aggregated to obtain the ranking order of candidate anchor items. Note that the test statistics and p-values display the following symmetry properties $|t_j(\{\ell\})| = |t_\ell(\{j\})|$ and $p_j(\{\ell\}) = p_\ell(\{j\})$ since the constant scale shift of one single anchor item is reflected in the test statistic of the item currently investigated and vice versa. Even though the p-values represent a monotone decreasing transformation of the absolute test statistics, the aggregations of both measures may yield different ranking orders.

Rank-based approach

Anchor selection strategies use the information from the auxiliary DIF tests of type (I) or of type (II) to define a criterion c_j for each item j that ideally reflects how strong the item is affected by DIF. All anchor selection strategies that are regarded in this article follow a rank-based approach that was first suggested together with auxiliary tests of type (I) by Woods (2009). The ranking order of candidate anchor items is defined by the ranks of the criterion values $\text{rank}(c_j)$. The item displaying the lowest rank is the first candidate anchor item, whereas the item corresponding to the highest rank is the last candidate anchor item.

The ranking order resulting from the anchor selection strategies is used within the anchor classes to conduct the final DIF analysis. For the *constant four anchor class*, the items with the lowest four ranks are selected as the final anchor set $\mathcal{A}_{\text{final}}$. For the *iterative forward anchor class*, items are selected into the anchor as long as the anchor is shorter than the number of currently presumed DIF-free items. In this anchor class, anchor items are selected in a step-by-step procedure following the ranking order that results from the anchor selection. When the stopping criterion is reached, the final anchor set $\mathcal{A}_{\text{final}}$ is found.

Final DIF analysis

The final DIF tests are carried out using the anchor set $\mathcal{A}_{\text{final}}$. 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. To overcome the problem that the classification of an item as a DIF or a DIF-free item is intended for each of the k items, we classify the first final anchor item with the lowest rank to be DIF-free – a decision that may be false if even the item with the lowest rank does indeed have DIF, but in this case this would be noticeable in the final test results. Note that this decision is by no means as drastic as testing only those items for DIF that have not been

selected as anchor, as was done e.g. by Woods (2009), or as choosing the anchor items only from the set of items that are known to be DIF-free in a simulation, as was done by Wang and Yeh (2003) and Wang (2004) but cannot be done in any real study where the true DIF and DIF-free items are unknown.

In the following, first, the strategies that are built on auxiliary DIF tests of type (I) are reviewed. Second, selection strategies that rely on auxiliary DIF tests of type (II) are discussed. Third, three new strategies are suggested at the end of the section that also rely on auxiliary DIF tests of type (II). Note that the DIF tests mentioned in the next paragraphs are only used as preliminary steps to assess the criterion values that determine the ranking order of candidate anchor items.

All-other selection

The *all-other selection strategy* (AO-selection) was proposed by Woods (2009) as what she called the rank-based strategy. For this strategy a predefined number of anchor items is chosen according to the lowest ranks of the absolute DIF test statistics resulting from the auxiliary DIF tests of type (I):

$$c_j^{\mathrm{AO}} = |t_j(\{1, \ldots, k\} \backslash j)|. \tag{5}$$

(Note that, originally, Woods (2009) suggested to use the ratios of the test statistics and the degrees of freedom, that may vary across items if the items display a different number of response categories. However, this is not discussed here since the responses are always dichotomous in the Rasch model that we focus on here.)

The *constant anchor method* of 20% of the items based on the AO-selection was found to be superior compared to the *all-other anchor method* in the majority of the simulated settings and compared to the *constant single anchor method* based on the AO-selection (Woods 2009). Nevertheless, the author claimed that "[a] study comparing the strategy proposed here to the various other suggestions for empirically selecting anchors is needed" (Woods 2009, p. 53).

All-other purified selection

Recently, Wang *et al.* (2012) suggested a modification (here referred to as AOP-selection for *all-other purified selection*) of the *all-other anchor selection strategy* proposed by Woods (2009) by adding a scale purification procedure. First, auxiliary DIF tests of type (I) are carried out. Similar to the iterative procedures (used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002), those items displaying DIF are excluded from the set of anchor items and DIF tests are conducted using the new anchor set. These steps are repeated until two successive steps reach the same results. In the next step, DIF tests are conducted using the purified anchor $\mathcal{A}_{\mathrm{purified}}$. Here, the first anchor item obtains no DIF test statistic, since only $k-1$ test statistics are available, and is, thus, omitted in the ranking of candidate anchor items. The criterion values of the remaining $k-1$ items are defined by

$$c_j^{\mathrm{AOP}} = |t_j(\mathcal{A}_{\mathrm{purified}})|. \tag{6}$$

In a simulation study, Wang *et al.* (2012) found the modified AOP-selection to be superior to the AO-selection since both methods displayed comparable results when DIF was balanced but the AOP-selection yielded more often a DIF-free anchor set when DIF was unbalanced. Still, there were conditions where the proportions of replications yielding a DIF-free anchor

set were far away from 100%, e.g. 13% for the AO- and 17% for the AOP-selection when the sample size was small (i.e. 250 observations in each group in their most difficult scenarios).

Number of significant threshold selection

An anchor selection strategy that is a simplified version of the proposition of Wang (2004) is called *number of significant threshold* (NST) selection strategy here. Now, auxiliary DIF tests of type (II) are carried out and the number of significant DIF tests defines the criterion values

$$c_j^{\text{NST}} = \sum_{\ell \in \{1,\ldots,k\}\setminus j} \mathbb{1}\{p_j(\{\ell\}) \leq \alpha\} \tag{7}$$

that is written as the number of p-values that do not exceed the threshold α, e.g. $\alpha = 0.05$. $\mathbb{1}$ denotes the indicator function. Thus, the item displaying the least number of significant DIF tests is chosen as the first anchor item. If more than one item displays the same number of significant results, one of the corresponding items is selected randomly.

Originally, Wang (2004) suggested the next candidate (NC) modification: The item that was selected by the NST-selection strategy functions as the current single anchor item and DIF tests are again carried out (see Wang 2004, p. 249). The next candidate is then included in the anchor if it displays "the least magnitude" (Wang 2004, p. 250) of (non-significant) DIF and the steps are repeated until either the pre-defined anchor length is reached or the candidate item displays significant DIF. Since Kopf *et al.* (2013) found the NST-selection superior to the original NC-strategy, only the former is investigated in this article.

Mean test statistic selection

To reach an ideally pure set of anchor items, Shih and Wang (2009) introduced the following anchor selection procedure: Every item is assigned the mean absolute DIF test statistic from the auxiliary DIF tests of type (II)

$$c_j^{\text{MT}} = \frac{1}{k-1} \sum_{\ell \in \{1,\ldots,k\}\setminus j} |t_j(\{\ell\})|. \tag{8}$$

We abbreviate this method MT-selection (for *mean test statistic selection*). Shih and Wang (2009) found high rates of correctly locating one or four DIF-free anchor items when the sample size was high (i.e. 1500 observations in each group in their most difficult scenarios).

Mean p-value selection

In addition to the existing approaches described above, we propose three new anchor selection strategies. First, we suggest an idea similar to the MT-strategy of Shih and Wang (2009) (see equation 8) that we abbreviate MP-strategy (for *mean p-value selection*). Instead of the lowest mean absolute DIF test statistic, items are here chosen that display the highest mean p-value from the auxiliary tests of type (II) and, thus, (for easier comparability with the previous methods) the criterion is defined by negative mean p-values

$$c_j^{\text{MP}} = -\frac{1}{k-1} \sum_{\ell \in \{1,\ldots,k\}\setminus j} p_j(\{\ell\}). \tag{9}$$

The next two suggestions were inspired by the threshold approach of the NST-selection (see equation 7) where those items are chosen as anchor items that display the least number of

significant DIF test results. Kopf *et al.* (2013) showed that this strategy was superior to the AO-selection when the DIF direction was unbalanced. The major drawback using the NST-selection was that it was strongly affected by the sample size. The reason for this is that the selection is based on the decisions of statistical significance tests which are strongly influenced by the sample size. Therefore, the next two newly suggested anchor selection strategies rely on a different criterion and both methods assume – similar to the MT- and the MP-selection – that the majority of items is DIF-free, an assumption that is often found in the construction of anchor or DIF methods (see, e.g., Shih and Wang 2009; Magis and De Boeck 2011).

Mean test statistic threshold selection

Our second suggestion is the following: For every item the absolute mean of the test statistics resulting from the auxiliary tests of type (II) is calculated and the resulting values are ordered. The threshold for the MTT-selection (for *mean test statistic threshold*) is the $(\lceil 0.5 \cdot k \rceil)$-th ordered value, which is indicated by the index in parenthesis, for an even number of items or the next larger whole number in case of an odd number of items (indicated by the ceiling function $\lceil\,\rceil$). The number of absolute test statistics exceeding this threshold determines the criterion value:

$$c_j^{\mathrm{MTT}} = \sum_{\ell \in \{1,\ldots,k\}\setminus j} \mathbb{1}\left\{ |t_j(\{\ell\})| > \left(\left| \frac{1}{k-1} \sum_{\ell \in \{1,\ldots,k\}\setminus j} t_j(\{\ell\}) \right| \right)_{(\lceil 0.5 \cdot k \rceil)} \right\}. \tag{10}$$

The items corresponding to the lowest number of test statistics above the threshold are chosen as anchor items. Here, we follow an argumentation similar to the argumentation of Shih and Wang (2009, p. 193). When the anchor item is DIF-free, which is assumed to be the case for the majority of the items, the DIF tests work appropriately. On the other hand, if a DIF item functions as the anchor, those items with the same direction of DIF display less DIF (or even no DIF in the most indistinct situation when the magnitude of DIF is approximately the same for the respective items), those items with the opposite direction of DIF display on average their original magnitude of DIF plus the artificial magnitude of DIF of the anchor item and the items truly free of DIF display on average the artificial DIF magnitude of the anchor item.

Thus, those DIF tests where the anchor is truly DIF-free should display the least absolute mean test statistics. Since the majority of items – i.e. at least 50% of all k items – is assumed to be DIF-free, the $(\lceil 0.5 \cdot k \rceil)$-th mean test statistic should correspond to a DIF-free item. In order to use the information of every single test statistic as opposed to the mere mean values, we use the indicator function to provide the information whether the single test statistics exceed the $(\lceil 0.5 \cdot k \rceil)$-th ordered absolute mean test statistic. Furthermore, in case of unbalanced DIF, the absolute mean test statistics may be very similar, when the DIF proportion is close to 0.5. The binary decisions are assumed to yield more accurate classifications of the truly DIF-free items. The selection strategy is designed for all directions of DIF and intended for all sample sizes. In contrast to the MT-selection proposed by Shih and Wang (2009), we use the absolute mean test statistics instead of the mean absolute test statistics. The reason for this is that all item parameters vary slightly between reference and focal group due to sampling fluctuation. These differences are expected to cancel out when the absolute values are taken after the mean statistic and, hence, should yield a better threshold.

Mean p-value threshold selection

In our third suggestion, similar to the MTT-selection in equation 10, the threshold of the MPT-selection (for *mean p-value threshold*) relies again on auxiliary DIF tests of type (II). Now, the ($\lceil 0.5 \cdot k \rceil$)-th ordered (from large to small) value of the mean of the resulting p-values $p_j(\{\ell\})$ is used as the threshold. The criterion value is defined by the number of tests per item that yield p-values exceeding the threshold p-value

$$c_j^{\text{MPT}} = - \sum_{\ell \in \{1,\ldots,k\} \setminus j} \mathbb{1} \left\{ p_j(\{\ell\}) > \left(\frac{1}{k-1} \sum_{\ell \in \{1,\ldots,k\} \setminus j} p_j(\{\ell\}) \right)_{(\lceil 0.5 \cdot k \rceil)} \right\}. \tag{11}$$

In summary, the newly suggested methods are developed for balanced and unbalanced DIF situations and should, thus, outperform not only the AO-selection that initiates with the potentially biased DIF test results using the all-other method, but also the AOP-selection that may not be able to exclude all DIF items from the anchor set when the proportion of DIF items is high (Wang *et al.* 2012). In comparison with the NST-selection, which uses the binary decisions of the significance tests (Woods 2009), the newly suggested methods should be less affected by sample size. While the MT- and the MP-selection use mere mean values, the MPT- and the MTT-selection use all individual test results and are, therefore, expected to better distinguish between DIF and DIF-free anchor items. By employing a threshold, the new methods should select those items as anchor that display little artificial DIF which can be caused by contamination (see, e.g., Finch 2005; Woods 2009) or by random sampling fluctuation (Kopf *et al.* 2013).

4. Simulation study

In order to evaluate the performance of the newly suggested anchor selection strategies, we conducted an extensive simulation study in the free R system for statistical computing (R Development Core Team 2011). Parts of the simulation design were inspired by the settings used by Wang *et al.* (2012). Each setting from the simulation study is replicated 1000 times to ensure reliable results.

4.1. Data generating processes

One replication corresponds to a data set that contains the information of the test including the item responses, the group membership and the ability variable.

- Test characteristics

  Here, we consider a test length of $k = 40$ items.

- IRT model

  The responses follow the Rasch model

$$P\left(U_{ij} = 1 \,|\, \theta_i, \beta_j\right) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{12}$$

with the difficulty parameters β = (-2.522, -1.902, -1.351, -1.092, -0.234, -0.317, 0.037, 0.268, -0.571, 0.317, 0.295, 0.778, 1.514, 1.744, 1.951, -1.152, -0.526, 1.104, 0.961, 1.314, -2.198, -1.621, -0.761, -1.179, -0.610, -0.291, 0.067, 0.706, -2.713, 0.213, 0.116, 0.273, 0.840, 0.745, 1.485, -1.208, 0.189, 0.345, 0.962, 1.592)$^\top$ used by Wang *et al.* (2012). The first 45% of the items are simulated as the DIF items (see Section DIF proportion and DIF magnitude below).

- Ability distribution

  In the following simulation study, ability differences are simulated since this case is often found to be more challenging for the methods than a situation where no ability differences are present (see, e.g., Penfield 2001). The ability parameters θ_i follows a standard normal distribution for the reference group $\theta^{\text{ref}} \sim N(0, 1)$ and a normal distribution with a lower mean for the focal group $\theta^{\text{foc}} \sim N(-1, 1)$ similar to Wang *et al.* (2012).

- DIF proportion and DIF magnitude

  The **proportion** of simulated DIF items is set to 45% (this proportion was found e.g. in the study of Allalouf, Hambleton, and Sireci 1999). We also considered lower proportions $\%DIF \in \{0, 0.1, 0.2, 0.3, 0.4\}$ which led to smaller differences between the methods but qualitatively similar results.
  For those items j affected by DIF, the **magnitude** of DIF as simulated is set to the constant value of $\Delta_{\text{DIF}} = \beta_{\text{ref}} - \beta_{\text{foc}} = 0.6$. This magnitude has previously been used in DIF simulation studies (Swaminathan and Rogers 1990; Finch 2005; Wang *et al.* 2012) and reflects a moderate effect size measured by Raju's area (Raju 1988; Jodoin and Gierl 2001).

4.2. Manipulated variables

In addition to the anchor selections investigated by Wang *et al.* (2012), namely the AO- and the AOP-selection, five other anchor selection strategies and also the perfect selection of DIF-free items that serves as a benchmark method are included (for a summary see Table 1).

- Sample size

  The **sample size** is defined by the following pairs of reference and focal group sizes: $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250, 250), (500, 250), (500, 500), (750, 500), (750, 750), \ldots, (1500, 1500)\}$.

- DIF direction

  The sign of Δ_{DIF} is set consistent with the intended direction of DIF. The **direction** of DIF is either balanced or unbalanced. In case of balanced DIF, the DIF items either favor the focal or the reference group, and on average, no group has an advantage in the test. In case of unbalanced DIF, all items favor the reference group.

- Anchor methods

  Anchor classes: All anchor selection strategies are combined with two anchor classes, the *constant four anchor class* (abbreviated constant4) and the *iterative forward class* (abbreviated forward).

  Anchor selections: Eight different anchor selection strategies (for a brief summary see Table 1) are compared across the simulated settings: The **AO**-, **AOP**[1]-, **NST**- and **MT**-selection as well as the newly suggested **MP**-, **MTT**- and **MPT**-selection and the **perfect**-selection that serves as the benchmark condition: The perfect selection for the *four anchor class* includes four randomly chosen DIF-free items. For the *iterative forward anchor class*, a random ranking order that includes the DIF-free items first, followed by the DIF items is handed to the procedure. The remaining steps of the iterative procedure are carried out as usual. Thus, for the *'perfect' forward method*, it may happen that DIF items occur in the anchor because the length of the iteratively selected anchor may exceed the length of the sequence of DIF-free items, which is not the case for the *perfect four anchor method*.

  Anchor methods: 16 anchor methods result from the combination of the eight anchor selection strategies with the two anchor classes. Their names (constant4-AO, constant4-AOP, constant4-NST, constant4-MT, constant4-MP, constant4-MTT, constant4-MPT, constant4-perfect, forward-AO, forward-AOP, forward-NST, forward-MT, forward-MP, forward-MTT, forward-MPT, forward-perfect) include the anchor class (constant4 or forward) together with the abbreviation of the anchor selection.

4.3. Outcome variables

In order to evaluate whether the anchor selection strategies locate anchor items that allow to correctly classify DIF and DIF-free items, the following outcome variables are recorded in each of the 1000 replications of one simulated setting:

- 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 in the final DIF test. The estimated *false alarm rate* for each simulated setting is computed as the mean over all 1000 replications and, hence, represents the type one error rate of the final DIF test.

- Hit rate

  The *hit rate* for a single replication is computed as the proportion of DIF items that are (correctly) diagnosed with DIF in the final DIF test. Analogously, the estimated *hit rate* is again computed as the mean over all 1000 replications and, thus, corresponds to the statistical power of the final DIF test.

[1] In case all items were excluded from the anchor in the initial step (which happened in only 2 out of $33,000$ replications), here, one single anchor item was chosen using the AO-strategy. When the constant four anchor class was combined with the AO-strategy, four anchor items were selected according to the lowest ranks of the resulting DIF test statistics using the AO-selected anchor item. Similarly, when the iterative forward anchor class was investigated, the ranking order was then built from the resulting DIF test statistics using the AO-selected anchor item and the iterative procedure was conducted normally.

Selection	Description
AO	The items are ranked according to the lowest absolute test statistics $\lvert t_j(\{1,\ldots,k\}\backslash j)\rvert$.
AOP	Beginning with all other items as anchor, DIF items are iteratively excluded from the anchor until the purified anchor set $\mathcal{A}_{\text{purified}}$ is reached; the items are ranked according to the lowest absolute test statistics $\lvert t_j(\mathcal{A}_{\text{purified}})\rvert$.
NST	The items are ranked according to the lowest number of significant test statistics $t_j(\{\ell\})$.
MT	The items are ranked according to the lowest mean absolute test statistics $\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\backslash j}\lvert t_j(\{\ell\})\rvert$.
MP	The items are ranked according to the largest mean p-values $\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\backslash j} p_j(\{\ell\})$.
MTT	The items are ranked according to the smallest number of test statistics $t_j(\{\ell\})$ exceeding the $(\lceil 0.5\cdot k\rceil)$-th ordered absolute mean test statistic $\left\lvert\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\backslash j} t_j(\{\ell\})\right\rvert$.
MPT	The items are ranked according to the largest number of p-values $p_j(\{\ell\})$ exceeding the $(\lceil 0.5\cdot k\rceil)$-th ordered mean p-value $\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\backslash j} p_j(\{\ell\})$.
perfect	The perfect ranking consists of randomly permuted DIF-free items followed by randomly permuted DIF items.

Table 1: A short summary of the anchor selection strategies that are investigated in this article.

5. Results

In the following, we restrict the presentation to the extreme condition where 45% of the items display DIF. The reason for this is that large proportions of DIF items may indeed occur in practical testing situations (examples can be found in Shih and Wang 2009, p. 186) and the anchor selection strategies should be compared in a situation where the classification of DIF and DIF-free items is rather challenging.

5.1. Anchor selection for the constant four anchor class

In this section, the anchor selection strategies combined with the *constant anchor class* are regarded. Thus, four anchor items were selected by the respective strategy and the results of the final DIF tests are discussed. Figure 1 contains the results of the false alarm rate (top row) and the hit rate (bottom row) in case of 45% DIF items that did not systematically favor one group (balanced DIF pattern, left column) or that systematically favored the reference group (unbalanced DIF pattern, right column).

In the balanced condition, almost all anchor methods displayed false alarm rates holding the 5% level in the observed range of the sample size. The only exception was the method relying on the NST-selection with the maximum observed false alarm rate of 0.09 that occurred at the sample size of 750 observations in each group. The corresponding false alarm rate for the NST-selection displayed an inversely u-shaped pattern that was also found and discussed in

detail in the study of Kopf *et al.* (2013). The perfect selection was near the significance level, while the remaining methods (except for the constant4-NST method) stayed below 0.05.

The method relying on the NST-strategy also displayed a lower hit rate compared to the other anchor methods. All remaining anchor methods displayed a hit rate that increased with the sample size. Surprisingly, the perfect anchor did not display a substantially higher hit rate. In summary, four anchor items were selected appropriately in the balanced condition by all anchor selections except for the NST-selection strategy in regions of medium sample sizes.

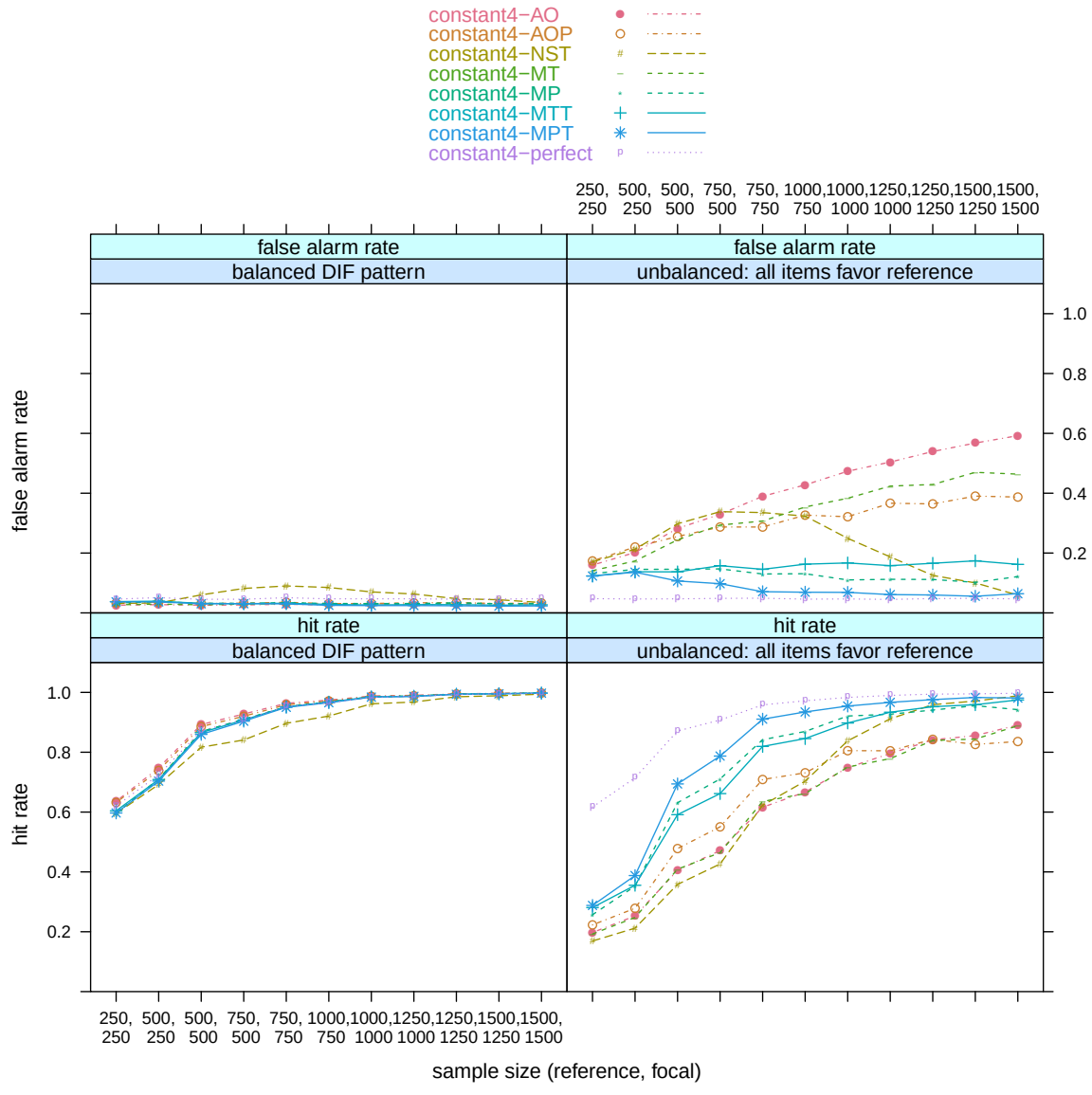

Figure 1: Balanced condition: 45% DIF items with no systematic advantage for one group; unbalanced condition: 45% DIF items favoring the reference group; sample size varied from $(250, 250)$ up to $(1500, 1500)$; top row: false alarm rates; bottom row: hit rates.

In the unbalanced condition, the anchor selections strongly differed regarding their ability to select four anchor items that yielded appropriate final DIF tests. Both selections based on auxiliary DIF tests of type (I) using all other items as anchor were not appropriate for the unbalanced DIF condition: The AO-selection yielded a strongly augmented false alarm rate that even increased with the sample size. The purified AOP-selection did not solve the problem, even though the false alarm rate was less augmented. Similarly, the MT-selection that was based on auxiliary DIF tests of type (II) with every other item as single anchor displayed an increasing false alarm rate. The NST-selection, again, yielded an inversely u-shaped pattern with a false alarm rate that exceeded the significance level when the sample sizes were below 1500 observations in each group. In regions of small to medium sample sizes, the newly suggested methods (the MP-, the MTT- and the MPT-selection) outperformed the existing strategies. The MP-selection that chose the anchor items according to the largest mean p-values (as opposed to the mean absolute test statistics in the MT-selection) yielded an almost constant mean false alarm rate around 0.12 (as opposed to the higher and increasing false alarm rate of the MT-selection). The MTT-selection yielded an almost constant false alarm rate around 0.16. The MPT-selection reached the lowest false alarm rate that was decreasing and well-controlled in regions of medium and large sample sizes. The perfect selection was, again, near the significance level.

The hit rates in the unbalanced condition also strongly differed across the anchor selections. The highest hit rate occurred for the perfect selection followed by the newly suggested anchor selections: The MPT-selection that corresponded to the lowest false alarm rate also showed the highest hit rate and was, thus, the best performing method to select four anchor items empirically. The MP-selection displayed the second best result in the majority of the simulated settings, followed by the MTT-selection. The existing anchor selections (the AO-, AOP-, NST- and MT-selection) displayed far lower hit rates and are, thus, not recommended for the DIF-free-then-DIF strategy. Only when 1500 observations were available in each group, the NST-selection also yielded a low false alarm rate in combination with a high hit rate.

In summary, the MPT-selection outperformed the other suggestions in selecting four anchor items by yielding a low false alarm rate while simultaneously achieving a high hit rate. The newly suggested MP-selection yielded clearly better results than the MT-selection even though both methods were structurally very similar and the MPT-selection outperformed the MTT-selection. Thus, an anchor selection based on p-values instead of mean test statistics is advisable for selecting an anchor of constant length four. As expected, the methods based on threshold comparisons (MPT- and MTT-selection) improved the final DIF test results compared to the corresponding strategies based on mere mean values (MP- and MT-selection).

5.2. Anchor selection for the iterative forward anchor class

In the next section, we investigate the combination of the anchor selection strategies with the *iterative forward anchor class* that was designed to specify a longer anchor (Kopf *et al.* 2013). Similar to Section 5.1, Figure 2 includes the results for the false alarm rate (top row) and the hit rate (bottom row) in case of 45% DIF items that did not systematically favor one group (balanced DIF pattern, left column) or that systematically favored the reference group (unbalanced DIF pattern, right column).

In case of balanced DIF, there was neither a visible difference in the false alarm rates nor in the hit rates for any of the investigated methods. Again, all selections yielded test results

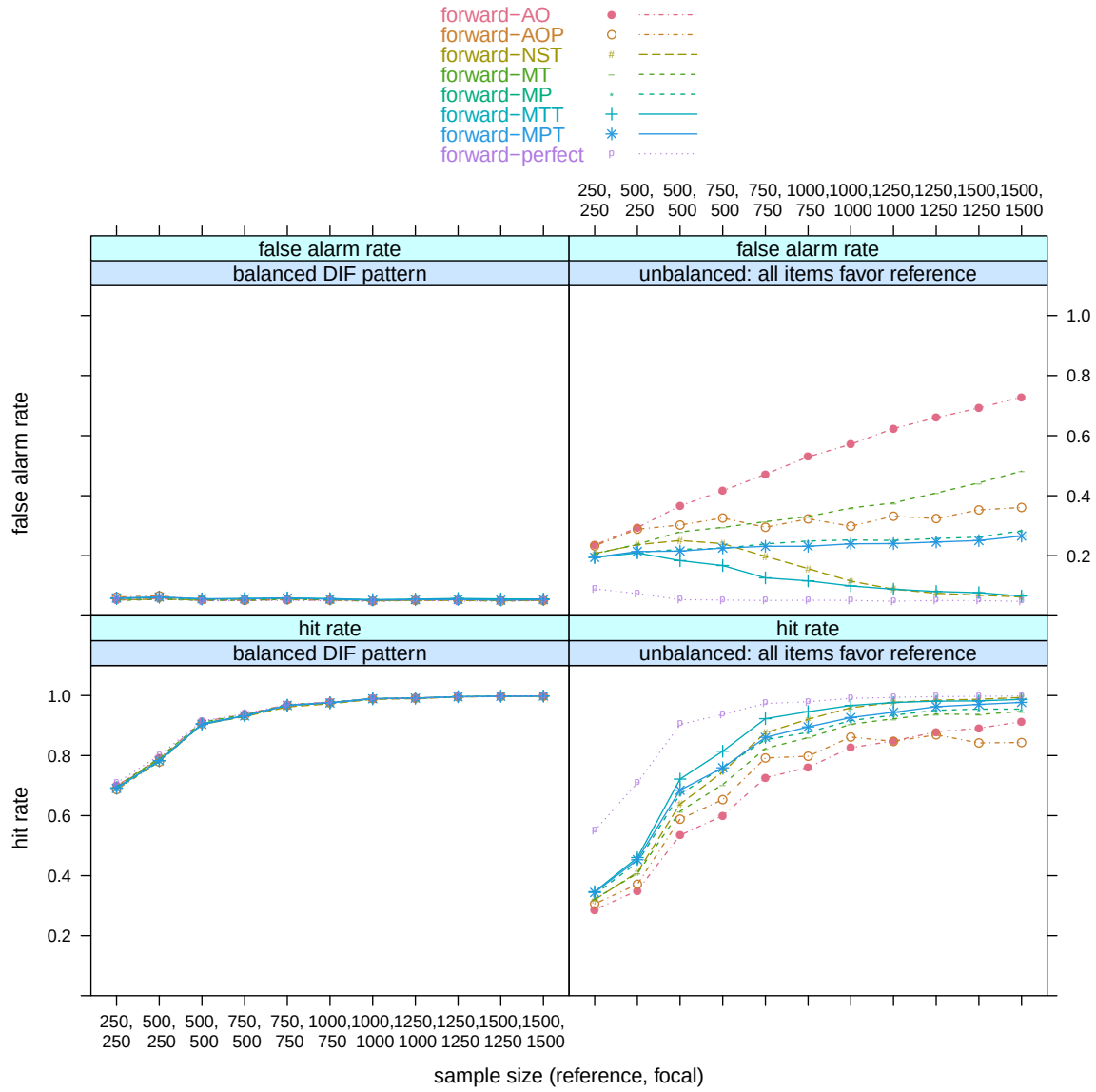

Figure 2: Balanced condition: 45% DIF items with no systematic advantage for one group; unbalanced condition: 45% DIF items favoring the reference group; sample size varied from $(250, 250)$ up to $(1500, 1500)$; top row: false alarm rates; bottom row: hit rates.

similar to the iterative method based on the perfect anchor. Hence, all selection strategies were advisable and the *iterative forward anchor class* was robust against the anchor selection strategy employed in this case.

In contrast to this, the results of the final DIF tests varied notably with the anchor selection strategies when DIF was unbalanced. The largest false alarm rates occurred for the AO-, AOP- and also the MT-method that were again not suited to locate an anchor in the unbalanced

condition. Their false alarm rates even increased with the sample size. The NST-selection was more appropriate in selecting the longer iterative anchor since it achieved a lower false alarm rate in regions of medium to large sample sizes. The MP- and the MPT-selection now yielded test results with very similar false alarm rates (with a slight advantage for the threshold method) that also clearly exceeded the significance level. The lowest false alarm rate among the empirical selection strategies occurred for the forward-MTT method. But there is still room for improvement as the perfect anchor selection still displays test results with a lower false alarm rate, especially in regions of small or medium sample sizes.

The hit rates of the unbalanced condition again increased with the sample size. Except for the perfect forward method, the newly suggested forward-MTT method reached the highest hit rate in the vast majority of the simulated settings. In regions of small sample sizes, it was followed by the two other newly suggested methods (the forward-MP and the forward-MPT method), whereas in regions of medium to large sample sizes, the forward-NST method reached the second best hit rate. The remaining AO-, AOP- and MT-strategy displayed DIF test results reaching unsatisfying hit rates.

In summary, the newly suggested MTT-selection outperformed the other empirical selection strategies by yielding test results with a low false alarm rate and a high hit rate in any regarded condition. Compared to the selection of an anchor of constant length four, where the MPT-selection based on p-values reached the best final DIF test results, for the longer, iteratively selected anchor the MTT-selection that is built on mean test statistics is advisable. A detailed explanation for this finding will be given in the next section. Again, the methods based on threshold comparisons (MPT- and MTT-selection) outperformed the corresponding strategies based on mere mean values (MP- and MT-selection).

5.3. Comparison of the mean test statistic and mean p-value threshold selection

To explain the fact that the MPT-selection yielded better results when it was combined with the *constant four anchor class*, whereas the MTT-selection performed better combined with the *iterative forward anchor class*, the ranking order of candidate anchor items is now regarded in detail for one balanced and one unbalanced setting (again with 45% DIF items and 1000 observations in each group). Figure 3 contains the proportions of DIF items in the ranking order of candidate anchor items. In the regarded setting, 22 items were DIF-free and, ideally, the 22 lowest ranks (from left to the vertical line) should display low proportions of DIF items.

In the balanced condition (Figure 3, top panel), the first items of the sequence of anchor candidates – i.e. the items to the left of the vertical line – displayed low proportions of DIF items over the simulation runs for both the MPT-selection (black bars) and the MTT-selection (gray bars). In contrast to this, the items that were assigned the highest ranks – i.e. the items to the right of the vertical line – displayed large proportions of DIF items. Thus, both anchor selection strategies yielded appropriate ranking orders that clearly separated DIF and DIF-free items: The first candidates displayed low proportions of DIF items, whereas the last candidates displayed large proportions of DIF items as intended for all ranks above 22.

In the unbalanced condition (Figure 3, bottom panel), the separation of candidates with low proportions of DIF items for the first ranks and high proportions for the last ranks was harder for both methods. Now the first anchor candidates displayed higher proportions of DIF items. Generally, the MTT-selection (gray) yielded lower DIF proportions for items up to the vertical

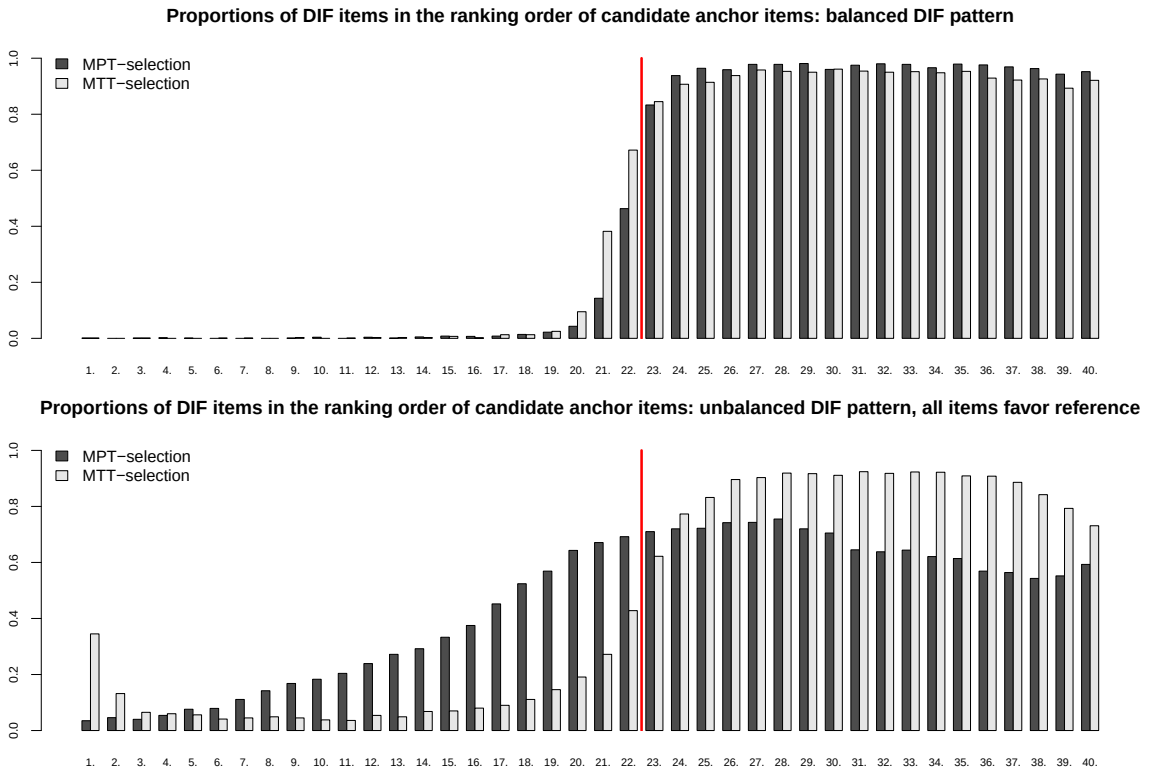

Figure 3: Top row: proportion of DIF items in the ranking order of anchor candidates in the balanced condition: 45% DIF items with no systematic advantage for one group; bottom row: proportion of DIF items in the ranking order of anchor candidates in the unbalanced condition: 45% DIF items favoring the reference group; sample size was set to 1000 observations in each group.

line compared to the MPT-selection (black) and was, thus, better suited to locate a longer anchor. However, when an anchor of constant length four was intended, only the first four candidates were included in the anchor. The first four ranks selected by the MPT-selection displayed lower proportions of DIF items items compared to the MTT-selection (see very left of Figure 3, bottom panel). Thus, the MPT-selection was better suited to locate four anchor items.

Now the question is addressed which of the methods – the constant4-MPT or the forward-MTT method – can be considered as overall superior. Therefore, we review the results from Section 5.1 and 5.2 together with information about the variation of the false alarm and the hit rate (not shown).

In the balanced condition, both methods led to low false alarm rates that fluctuated less for the constant4-MPT method, which should thus be preferred with respect to the false alarm rate. In contrast to this, the forward-MTT method achieved a higher and simultaneously less fluctuating hit rate and was, hence, superior regarding the hit rate. Only when the sample size was large, the constant4-MPT method might have a slight advantage by yielding a lower

and less fluctuating false alarm rate and a comparably high hit rate.

In the unbalanced condition, on one hand, the constant4-MPT method led to a lower false alarm rate, especially when the sample size was small. On the other hand, the false alarm rate of the forward-MTT method displayed far less fluctuation. It may be preferred in regions of large sample sizes since the results were more reliable. Regarding the hit rate, the forward-MTT method was superior since the hit rate was not only slightly higher but also, again, less fluctuating.

In summary, the first anchor candidates were more likely found from the set of DIF-free items by the MPT-selection, whereas the MTT-selection was better suited for longer anchors. However, first results show that neither the constant4-MPT nor the forward-MTT method was clearly superior in all settings regarding strictly smaller and less fluctuating false alarm rates and higher and less fluctuating hit rates.

6. Discussion and practical recommendations

In this article, we introduced three new anchor selection strategies and compared them to existing methods that do not rely on any prior knowledge of DIF-free items. Moreover, we introduced a straightforward notation of the anchor selection strategies to facilitate the implementation and the usage of the newly suggested anchor selection strategies. An extensive simulation study was conducted to evaluate the performance of the anchor selection strategies in combination with the *constant four anchor class* and the *iterative forward anchor class*. The two anchor classes are structurally different, since the *constant four anchor class* always uses a short anchor of constant length four, whereas the *iterative forward class* determines the anchor length in an iterative way and usually yields a longer anchor.

Our analysis showed that the results of the DIF tests evaluated by means of the false alarm and the hit rate strongly depended on the anchor selection strategies employed. This highlights the importance of a suitable anchor selection strategy that allows the researcher to correctly classify DIF and DIF-free items and to study the underlying causes of DIF (Jodoin and Gierl 2001). Consistent with previous results (see, e.g., Wang and Yeh 2003; Wang 2004; González-Betanzos and Abad 2012; Kopf *et al.* 2013), seriously inflated false alarm rates occurred if the anchor selection did not work appropriately, especially when DIF was unbalanced. This was the case for several existing anchor selection strategies. Anchor selections based on the *all-other anchor method* (the AO- and the AOP-selection) are inadvisable, since the tests were biased in the unbalanced DIF condition and even additional purification steps included in the AOP-selection were not able to completely reduce the bias. Hence, we advise against constructing new anchor selection strategies that use all other items as anchor. Unsatisfactory results were also found for the MT-selection that is based on mean absolute test statistics resulting from DIF tests for every item using every other item as single anchor. In the vast majority of the simulated settings, the newly suggested anchor selection strategies based on a threshold criterion clearly outperformed the existing suggestions.

Our results showed that the appropriateness of the anchor selection not only depended on the sample size, the proportion of DIF items and the direction of DIF, but also on the intended anchor length.

In case of the selection of a short anchor of constant length four, the MPT-selection outperformed all other investigated empirical anchor selection strategies by yielding a low false

alarm rate and simultaneously reaching a high hit rate in all regarded conditions. Thus, we recommend to use the MPT-selection if a short constant anchor length is intended.

When the selection strategies were combined with the *iterative forward anchor class*, the newly suggested MTT-selection reached the best results in the majority of the simulated settings and is, thus, recommended for DIF analysis when the *iterative forward anchor class* is used, as well as in general when a longer anchor length is intended.

Nevertheless, the benchmark method of the perfect anchor selection still reached lower false alarm rates and higher hit rates in regions of small to medium sample sizes when DIF was simulated unbalanced. Hence, new developments for anchor selection strategies that ideally follow the threshold approach are needed to further improve the classification of DIF and DIF-free items when the sample sizes are small. When the sample sizes are large, the newly suggested constant4-MPT and the forward-MTT method reached satisfying results in our simulation study. Future research may investigate the performance of these methods when other IRT models or other DIF tests are used and may evaluate modifications of the iterative anchor method.

Acknowledgements

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

Allalouf A, Hambleton RK, Sireci SG (1999). "Identifying the Causes of DIF in Translated Verbal Items." *Journal of Educational Measurement*, **36**(3), 185–198.

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.

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.

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 Developements, 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.

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

Kopf J, Zeileis A, Strobl C (2013). "Anchor Methods for DIF Detection: A Comparison of the Iterative Forward, Backward, Constant and All-Other Anchor Class." *Technical Report 141*, Department of Statistics, LMU Munich.

Lim RG, Drasgow F (1990). "Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning." *Journal of Applied Psychology*, **75**(2), 164 – 174.

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.

Magis D, De Boeck P (2011). "Identification of Differential Item Functioning in Multiple-Group Settings: A Multivariate Outlier Detection Approach." *Multivariate Behavioral Research*, **46**(5), 733–755.

McLaughlin ME, Drasgow F (1987). "Lord's Chi-Square Test of Item Bias With Estimated and With Known Person Parameters." *Applied Psychological Measurement*, **11**(2), 161–173.

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.

Paek I, Han KT (2013). "IRTPRO 2.1 for Windows (Item Response Theory for Patient-Reported Outcomes)." *Applied Psychological Measurement*, **37**(3), 242–252.

Penfield RD (2001). "Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures." *Applied Measurement in Education*, **14**(3), 235 – 259.

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

Raju N (1988). "The Area Between Two Item Characteristic Curves." *Psychometrika*, **53**(4), 495–502.

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.

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

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.

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.

Woods CM, Cai L, Wang M (2012). "The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT." *Educational and Psychological Measurement*. Online first.

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
Psychologische Methodenlehre, Evaluation und Statistik
Department of Psychology
Universität Zürich (University of Zurich)
Binzmühlestrasse 14
CH-8050 Zurich, Switzerland
E-mail: Carolin.Strobl@psychologie.uzh.ch