Article

A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection

Applied Psychological Measurement
2015, Vol. 39(2) 83–103

DOI: 10.1177/0146621614544195
apm.sagepub.com
SAGE

Julia Kopf[1], Achim Zeileis[2], and Carolin Strobl[3]

Abstract

In differential item functioning (DIF) analysis, 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 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, the new *iterative forward* anchor class is proposed. Several anchor classes are implemented with different anchor selection strategies and are compared in an extensive simulation study. The results show that the new anchor class combined with the single-anchor selection strategy is superior in situations where no prior knowledge about the direction of DIF is available.

Keywords

item response theory (IRT), Rasch model, anchor methods, anchor selection, contamination, differential item functioning (DIF), item bias

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 article, the focus is on *uniform* DIF where one group has a higher probability of solving an item (given the latent trait) over the

[1]Ludwig-Maximilians-Universität München, Germany
[2]Universität Innsbruck, Austria
[3]Universität Zürich, Switzerland

Corresponding Author:
Julia Kopf, Ludwig-Maximilians-Universität München, Ludwigstraße 33, München 80539, Germany.
Email: julia.kopf@stat.uni-muenchen.de

entire latent continuum and the group differences in the logit remain constant (Mellenbergh, 1982; Swaminathan & Rogers, 1990).

A variety of testing procedures for DIF on the item-level is available (for an overview, see, e.g., Millsap & Everson, 1993). These testing procedures can be divided into IRT-based methods that rely on the estimation of an IRT model and non-IRT methods, following a classification used, for example, by Magis, Raîche, Béland, and Gérard (2011). They list Lord's chi-square test, Raju's area method and the likelihood ratio test as the most commonly known IRT-based methods, and the Mantel–Haenszel method, the Simultaneous Item Bias Test (SIBTEST) method, and the logistic regression procedure as the most widely used non-IRT methods. In the analysis of DIF using IRT, item parameters are to be compared across groups. Mostly, research focuses on the comparison of two predefined groups, the reference and the focal group. Thus, a common scale for the item parameters of 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 & Verhelst, 1995). The items included in the restriction are termed *anchor items*.

An anchor method determines how many items are used as anchor items and how they are located. 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 the ''Anchor Process for the Rasch Model'' section). Even though the importance of the anchor method is undeniable, Lopez Rivas, Stark, and Chernyshenko (2009) claim that ''at this point, little evidence is available to guide applied researchers through the process of choosing anchor items'' (p. 252). 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, the authors introduce a new conceptual framework that distinguishes between the *anchor class* and the *anchor selection strategy*. First, *anchor classes* that describe the pre-specification of the anchor characteristics are reviewed and a new anchor class named the iterative forward anchor class is introduced. Second, the *anchor selection strategy* determines which items are chosen as anchor items. The complete procedure to choose the anchor is then called an *anchor method*. To derive guidelines which anchor method is appropriate for DIF detection in the Rasch model, the authors conduct an extensive simulation study. In the subsequent study, the authors compare the all-other, the constant, the iterative backward, and the newly suggested iterative forward anchor class for the first time. Furthermore, the subsequent study is to the authors' knowledge the first to systematically contrast different anchor selection strategies that are combined with the anchor classes. We discuss the all-other (AO) selection strategy (introduced as rank-based strategy by Woods, 2009) and the single-anchor (SA) selection strategy (based on a suggestion by Wang, 2004). Finally, practical recommendations are given to facilitate the anchor process for DIF analysis in the Rasch model. In the next section, necessary technical details are explained. The conceptual framework is introduced in detail in the ''A Conceptual Framework for Anchor Methods'' section. The simulation study is presented in the ''Simulation Study'' section and the results are discussed in the ''Results'' section. The problem of contamination and its impact are addressed in the ''Impact of Anchor Contamination'' section. Characteristics of the selected anchor items are discussed in the ''Characteristics of the Anchor Items Inducing Artificial DIF'' section. A concluding summary and practical recommendations are given in the ''Summary and Discussion'' section.

The Anchor Process for the Rasch Model

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).

Scale Indeterminacy

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,

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

with constants d_ℓ holding $\sum_{\ell=1}^{k} d_\ell \neq 0$ is placed on the item parameter estimates $\tilde{\beta}_\ell$ (Eggen & 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 estimated item parameter $\tilde{\beta}_\ell = 0$ or restricting all estimated item parameters to sum zero $\sum_{\ell=1}^{k} \tilde{\beta}_\ell = 0$ (Eggen & Verhelst, 2006). Without loss of generality, here the item parameter vector $\tilde{\beta}$ is estimated 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. In the following, different restrictions for which the sum of the estimated item parameters of a selection of items is set to zero are discussed. 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 $A = \mathrm{I}_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot \mathrm{a}^\top$, 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, . . .)$^\top$ including items 1 and 3). In addition, 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 to build a common scale for the item parameters of two groups is that the same restriction is employed in both groups (Glas & 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 estimated 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 to zero in each group ($\sum_{\ell \neq j} \hat{\beta}_\ell^g = 0$) for the so-called all-other anchor method. The item parameters and covariance matrices, estimated separately in each group, are transformed to the respective anchor method by means of Equations 2 and 3, so that all items are then shifted on the scale by $-\frac{1}{|\mathcal{A}|}\sum_{\ell \in \mathcal{A}} \tilde{\beta}_\ell^g$.

Item-Wise Wald Test

As a statistical test for DIF, the focus is on the item-wise Wald test here (see, e.g., Glas & Verhelst, 1995), but the underlying ideas in the next section can also be applied to other tests for DIF. Note that this item-wise Wald test is applied to the CML estimates (as in Glas & 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 (see, e.g., McLaughlin & Drasgow, 1987). The recent work of Woods, Cai, and Wang (2013) showed that an improved version of the Wald test, termed Wald-1 (see Paek & Han, 2013, and the references therein), also displayed well-controlled false alarm rates in their simulated settings if the anchor items were DIF-free. As the Wald-1 test also requires anchor items, it can in principle be combined with the anchor methods discussed here as well.

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 $H_0 : \beta_j^{\text{ref}} = \beta_j^{\text{foc}}$, where β_j^{ref} and β_j^{foc} denote the item difficulties for reference and focal group for item j and $\hat{\beta}_j^{\text{ref}}$ and $\hat{\beta}_j^{\text{foc}}$ the corresponding estimated item parameters using the anchor $\mathcal{A}^j$, has the following form:

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

Note that the estimated and anchored item parameters $\hat{\beta}^g = \hat{\beta}^g(\mathcal{A}^j)$, which can be calculated using Equation 2, depend on the anchor and, hence, so does the test statistic $T_j = T_j(\mathcal{A}^j)$. A detailed empirical example is provided in an online supplement. 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.

From a theoretical perspective and from the instructive example in the online supplement, 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. Empirical findings also show that, ideally, the anchor items should be DIF-free. Unfortunately, as prior to DIF analysis, it cannot be known which items are DIF-free, a somewhat circular problem is faced, as pointed out by Shih and Wang (2009). If DIF items are included in the anchor, this *contamination* may lead to seriously inflated false alarm rates in DIF detection (see, e.g., Finch, 2005; Wang, 2004; Wang & Su, 2004; Wang & Yeh, 2003; 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 & 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 (Shih & Wang, 2009; Thissen, Steinberg, & Wainer, 1988; Wang, 2004; Wang & Yeh, 2003; Woods, 2009).

A Conceptual Framework for Anchor Methods

In the following, the authors introduce a conceptual framework in which a variety of previously suggested anchor methods can be embedded. The new conceptual framework distinguishes between the *anchor class* and the *anchor selection strategy*.

Anchor Classes

In this conceptual framework, *anchor classes* describe characteristics of the anchor that answer the following questions: Is the anchor length predefined? 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?

The equal-mean and the all-other anchor class. 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 (used, e.g., by Cohen, Kim, & Wollack, 1996) the sum of all item difficulties—except the item currently tested for DIF—is restricted to be zero and the anchor set $\mathcal{A}^j = \{1, \ldots, k\} \setminus j$ depends on the studied item $j = 1, \ldots, k$. Both anchor classes have a predefined 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) and, hence, only the all-other method is included in the following simulation study.

The constant anchor class. The *constant* anchor class (used, e.g., by Shih & Wang, 2009; Thissen et al., 1988; Wang, 2004) includes a predefined number of the items (e.g., one or four items according to Thissen et al., 1988) or a certain proportion of the items (e.g., 10% or 20% according to Woods, 2009) as anchor. The term *constant* reflects the constant set of anchor items with a predefined, constant anchor length. In the subsequent simulation study, the authors 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 & Wang, 2009; Wang, Shih, & Sun, 2012). The constant anchor class needs to be combined with an explicit anchor selection strategy. For the constant single-anchor class, the first item of the ranking order of candidate anchor items is used as anchor, whereas for the constant four-anchor class, the first four items of the ranking order of candidate anchor items are used as anchor.

The iterative backward anchor class. The *iterative backward* anchor class (used, e.g., by Candell & Drasgow, 1988; Drasgow, 1987; Hidalgo-Montesinos & Lopez-Pina, 2002) 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 relinking procedure where one parameter estimation step suffices to conduct DIF analysis. First, 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,[1] the scales are re-linked using the new current anchor, the DIF analysis is carried out for all items except for the first anchor candidate (see the ''Anchor Methods'' section) and the steps are repeated until two steps reach the same results (e.g., Drasgow, 1987; Candell & Drasgow, 1988; Hidalgo-Montesinos & Lopez-Pina, 2002). This iterative procedure is referred to here as the *iterative backward* anchor class, as 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, what is referred to as

unbalanced DIF, DIF tests using all other items as anchor result in inflated false alarm rates. 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.

The iterative forward anchor class. Inspired by this result, the authors 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, the authors suggest to build the iterative anchor in a step-by-step forward procedure. Starting with the first candidate anchor item—determined by the anchor selection strategy—as single anchor item, the scales are linked and DIF is estimated. Then, iteratively, one item—located again by means of the respective anchor selection strategy—is added to the current anchor and DIF analysis is conducted using the new current anchor. These steps are repeated as long as the current anchor length is shorter than the number of nonsignificant test results in the current DIF tests (in short the number of currently presumed DIF-free items). 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.

Anchor Selection Strategies

The anchor selection strategies discussed here are based on preliminary item analyses. This means that—before the final DIF test is done—preliminary DIF tests are conducted to locate (ideally) DIF-free anchor items. The (nonstatistical) 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, & Magis, 2010).

The AO anchor selection. In the subsequent simulation study, the authors 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 is termed AO anchor selection strategy here. Initially, every item is tested for DIF using all other items as anchor. The ranking order of candidate anchor items is defined according to the lowest ranks of the resulting (absolute) DIF test statistics.

The next candidate (NC) and the SA anchor selection. Originally, Wang (2004) suggested an anchor method that is referred to as the NC method here. It includes both an anchor selection and an anchor class and is, thus, discussed in detail in the next section. Moreover, the authors simplify the suggestion of Wang (2004) for the anchor selection and call it the SA-selection strategy. It is, to the authors' knowledge, for the first time systematically compared with the AO-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 ranking order of candidate anchor items is defined according to 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.

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 and summarized in Table 1. All anchor methods that rely on an anchor selection consist of two steps: First, the anchor selection is carried out to determine a ranking order of candidate anchor items and the procedure defined by the anchor class is carried out to determine the final anchor. Second, 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). The final anchor $\mathcal{A}$ is independent of which item is studied. As $k-1$ 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, the first item selected as anchor item is declared DIF-free in the final DIF test, a decision that may be false if even the item with the lowest rank does indeed have DIF, but in this case, this would result in a lower hit rate in the final test results. All remaining items are tested for DIF using the final anchor $\mathcal{A}$. The all-other anchor method does not require an additional anchor selection and k tests result using the anchor $\mathcal{A}^j = \{1, \ldots, k\} \setminus j$. The constant anchor class consisting of one anchor item or four anchor items can be combined with the AO-selection strategy (*single-anchor-AO, four-anchor-AO*) and also with the SA-selection strategy (*single-anchor-SA, four-anchor-SA*).

Furthermore, the authors implemented the original suggestion of Wang (2004) that is referred to as the four-anchor-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). In this step, one DIF test 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 is defined here 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). 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.[2] These steps are repeated until either the next candidate anchor item displays DIF or the maximum anchor length (of four items in our implementation of the *four-anchor-NC* method) is reached. 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., Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). Note that the iterative backward class is not combined with the SA-selection as 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 AO-selection strategy (*iterative-forward-AO*) and with the SA-selection strategy (*iterative-forward-SA*).

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. Details about the background and motivation of the subsequent simulation study are provided in the online supplement. A total of 2,000 data sets (i.e., replications) are generated from each of 77 different simulation settings. For every data set, the item-wise Wald test (see the ''Anchor Process for the Rasch Model'' section)—based on one out of nine investigated anchor methods—is conducted at the significance level of .05 in the free R system for statistical computing (R Core Team, 2013). 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 Wang et al. (2012), Woods (2009), and Wang (2004).

Table 1. Classification and Nomenclature of the Investigated Anchor Methods.

Anchor class	Anchor selection	Combination	Initial step and anchor selection strategy
All-other	None	All-other Cf. for example, 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, Shih, and Sun (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 For example, 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.

Note. DIF = differential item functioning; AO = all-other selection; SA = single-anchor selection; NC = next candidate selection.

Data-Generating Process

Each data set 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. The authors also considered different test lengths of 20, 60, or 80 items (results not shown). In all cases, the results were qualitatively similar albeit the differences between the iterative forward and constant four-anchor class are somewhat smaller for 20 items (due to more similar anchor lengths) and larger for 60 and 80.

Person and item parameters. In the following simulation study, the authors have included ability differences as this case is often found more challenging for the methods than a situation where no ability differences are present (see, e.g., Penfield, 2001). The person parameters are generated from a normal ability distribution with a higher mean for the reference group $\theta^{\text{ref}} \sim N(0, 1)$ than for the focal group $\theta^{\text{foc}} \sim N(-1, 1)$ similar to Wang et al. (2012). For the item parameters, the authors chose the values that were already used by Wang et al. (2012).[3]

DIF items. In case of DIF, the first 15%, 30%, or 45% of the items (see the ''Directions and proportions of DIF'' section) are chosen 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 $+.6$ or $-.6$ (consistent with the intended direction of DIF). These differences have been used in previous DIF simulation studies (Finch, 2005; Swaminathan & Rogers, 1990; Wang et al., 2012) and reflect a moderate effect size measured by Raju's area (Jodoin & Gierl, 2001; Raju, 1988).

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 inserting the corresponding item and person parameters in the Rasch model Equation 6. The binary responses are then drawn from a binomial distribution with the resulting probabilities:

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

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 $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250; 250), (500; 250), (500; 500), (750; 500), (750; 750), \ldots, (1{,}500; 1{,}500)\}$. 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 focal group is generated such that every DIF item favors the reference group. In addition to the sample size, also the proportion of DIF is manipulated including the following percentages $p \in \{15\%; 30\%; 45\%\}$. The sample sizes, the DIF percentages and the DIF conditions (balanced and unbalanced) were fully crossed.

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 2,000 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 deviation over all 2,000 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 simulated DIF item (*risk of contamination*) is computed over all replications of one setting. The average proportion of simulated DIF items as compared with 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.

Results

Null Hypothesis: No DIF

In the first condition, all items were truly DIF-free. Therefore, only the false alarm rates (proportions of DIF-free items that were diagnosed with DIF) were computed and are displayed in Figure C.1 in the online supplement. The standard errors are reported in Table C.1 in the online supplement for equal sample sizes.

False alarm rates. All anchor methods held the 5% level. Although methods from the all-other, the iterative backward (iterative-backward-AO) and the iterative forward class (iterative-forward-SA, iterative-forward-AO) together with the constant four-anchor-NC method were near the significance level, most methods from the constant anchor class (single-anchor-AO and single-anchor-SA; four-anchor-AO and four-anchor-SA) remained below that level. Hence, DIF tests with an anchor method from the constant anchor class combined with the AO- and the SA-selection—especially the constant single-anchor methods, but also the constant four-anchors—were over-conservative.

Balanced DIF: No Advantage for One Group

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

False alarm rates. Figure 1 (top row) contains the false alarm rates for the balanced condition, reported also for equal sample sizes together with the standard errors in Table C.2 in the online supplement. Most methods displayed well-controlled false alarm rates—similar to the null

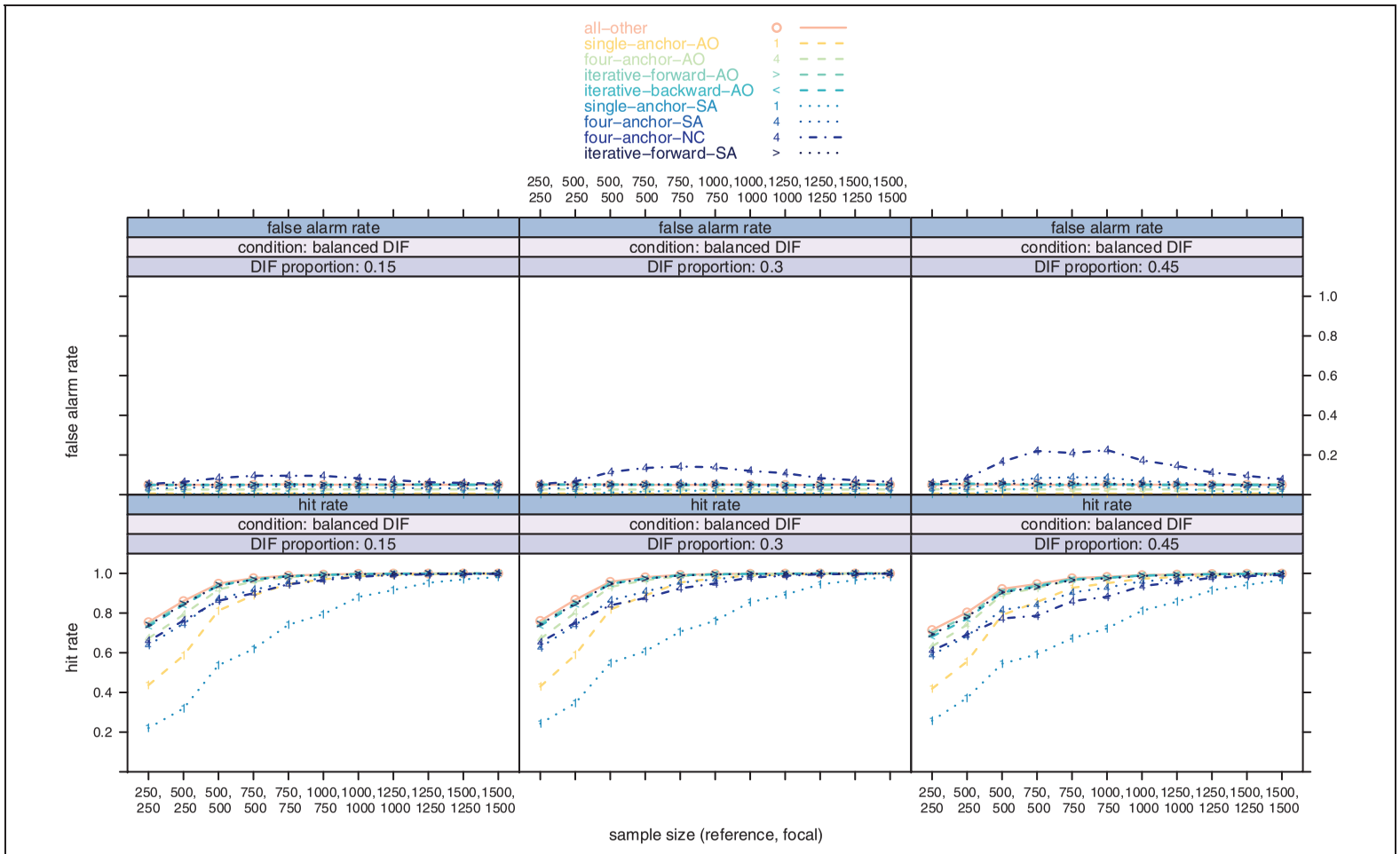

Figure 1. Balanced condition: 15%, 30%, and 45% DIF items with no systematic advantage for one group; sample size varies from (250; 250) up to (1,500; 1,500); top row: false alarm rates; bottom row: hit rates in the balanced condition.

Note. DIF = differential item functioning; AO = all-other selection; SA = single-anchor selection; NC = next candidate selection.

condition—with the following exceptions: The constant four-anchor-NC method and the four-anchor-SA method showed a false alarm rate that first increased but then decreased again with growing sample size in case of 45% DIF. The same inverse u-shaped pattern occurred in case of unbalanced DIF and is discussed in more detail in the ''Characteristics of the Anchor Items Inducing Artificial DIF'' section. Both constant single anchor methods (single-anchor-AO and single-anchor-SA) as well as the four-anchor-AO method, again, remained below the significance level. Hence, DIF tests based on the single-anchor-AO, the single-anchor-SA and the four-anchor-AO method were over-conservative.

Hit rates. Figure 1 (bottom row) depicts the hit rates (that specify how likely true DIF is detected) in the balanced condition, which increased monotonically with the sample size (for standard errors, see also Table C.3 in the online supplement). The hit rates with the slowest increase were from the constant single-anchor methods, but also from the constant four-anchor methods. The methods from the constant anchor class that were combined with the AO-selection (single-anchor-AO, four-anchor-AO) achieved higher hit rates than those combined with the SA-selection (single-anchor-SA, four-anchor-SA) or the NC-selection (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 showed rapidly increasing hit rates that converged to one for sample sizes above 750 in each group.

Unbalanced DIF: Advantage for the Reference Group

In the unbalanced condition, all items simulated with different item parameters favored the reference group. False alarm rates for the unbalanced condition are shown in Figure 2 (top row) and in Table C.4 in the online supplement together with the standard errors.

False alarm rates. As opposed to the previous results, in this condition, the majority of the anchor methods produced inflated false alarm rates: When the proportion of DIF items increased, the false alarm rates rose as well. Moreover, for most anchor methods, the false alarm rates increased 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 yielded 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 reference group were lower than the mean item parameters of the focal group. These mean differences in the item parameters shifted the scales of focal and reference group apart when the all-other method defined the restriction (similar to the instructive example in the online supplement). These artificial differences became significant when the sample size increased and, thus, resulted in an inflated 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 yielded biased DIF test results that induced a high false alarm rate when the sample size was large (as illustrated and discussed in more detail regarding the impact of contamination in the ''Impact of Anchor Contamination'' section). Constant anchors selected by the SA-strategy produced lower false alarm rates in regions of medium or large sample sizes. Here, again, an inverse u-shaped form is visible. After a certain point, the false alarm rates decreased again (a detailed explanation given in the ''Characteristics of the Anchor Items Inducing Artificial DIF'' section). The constant single-anchor methods showed lower false alarm rates than the corresponding constant four-anchor methods. For all constant methods, the single-anchor-SA method had the lowest false alarm rate when the sample size was large. The method from the iterative backward anchor class, which started the initial step

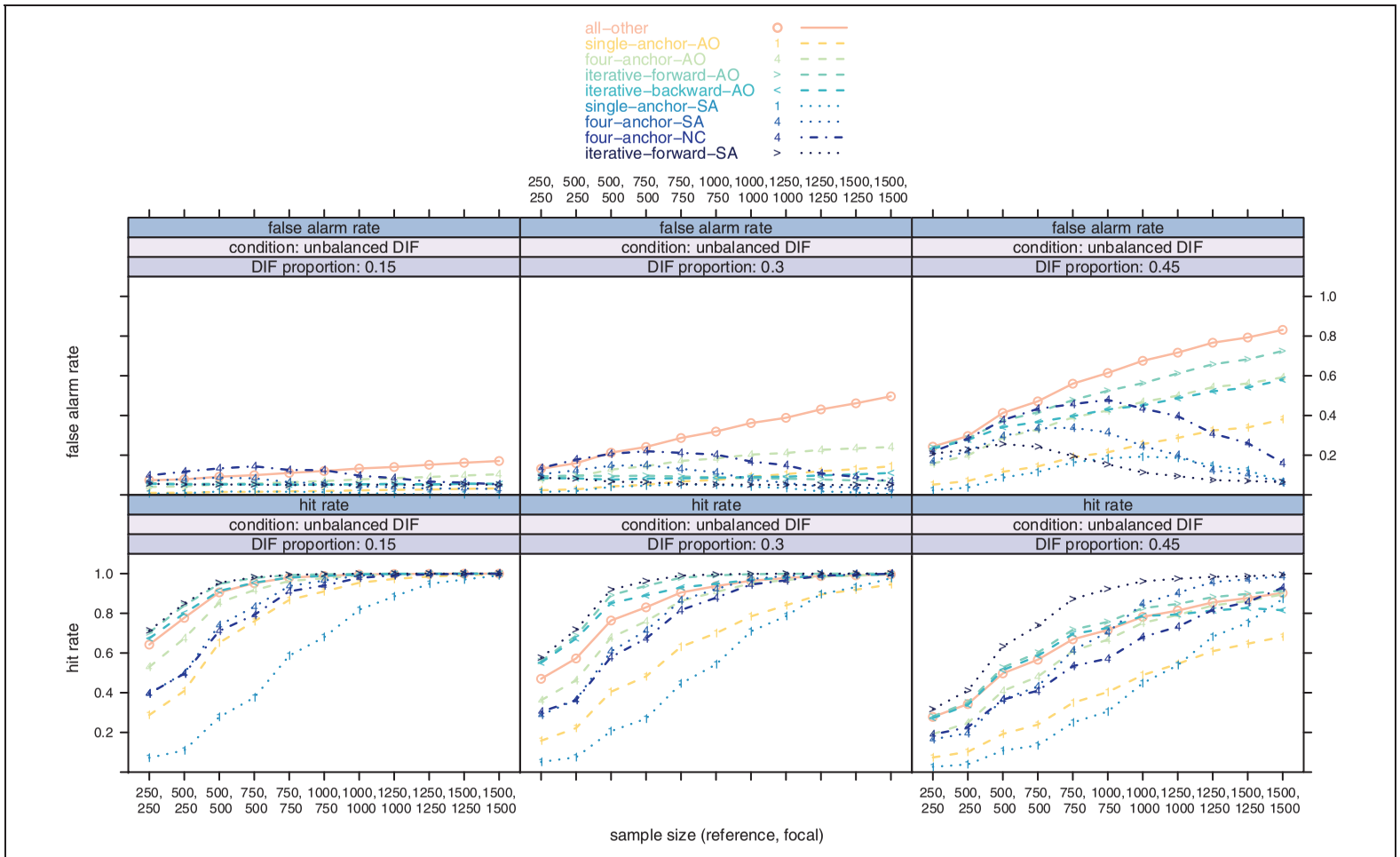

Figure 2. Unbalanced condition: 15%, 30%, and 45% DIF items favoring the reference group; sample size varies from (250; 250) up to (1,500; 1,500); top row: false alarm rates; bottom row: hit rates in the unbalanced condition.

Note. DIF = differential item functioning; AO = all-other selection; SA = single-anchor selection; NC = next candidate selection.

by using the all-other method, also led to inflated false alarm rates that rose when sample size increased. Methods from the iterative forward class displayed heterogeneous false alarm rates. The iterative-forward-AO method led to increased false alarm rates—similar to the constant methods with the AO-selection criterion—in the setting with 30% or 45% DIF. The clearly best iterative method in terms of a low false alarm rate was the new iterative-forward-SA method.

Hit rates. The hit rate in the unbalanced condition (cf. Figure 2, bottom row, and Table C.5 in the online supplement) in the settings of larger proportions of DIF items was different: Generally, the overall level of the hit rate was lower. Methods from the constant anchor class showed the slowest increase with the sample size. These methods also had lower hit rates compared with the methods from the iterative forward or backward class that were the only methods that displayed rapidly increasing and high hit rates. The all-other method was between the constant anchor methods and the iterative anchor methods. The new iterative-forward-SA method provided the highest hit rate and a rapid rise of the hit rate with increasing sample size. In case of 45% DIF, it displayed a much higher hit rate compared with all remaining methods in the majority of the simulated settings. The SA-selection strategy in combination with methods from the constant anchor class was more suitable than the AO-selection strategy regarding the hit rates when the sample size was large. The simplified four-anchor-SA method outperformed the originally suggested 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 was equal or superior to the iterative-forward-AO method over the entire range of simulated sample sizes. When accounting for both, the false alarm rate and the hit rate, the newly suggested iterative-forward-SA method is the only reasonable choice among the investigated methods in the simulated settings.

The Impact of Anchor Contamination

As discussed in the ''Anchor Process for the Rasch Model'' section and in the online supplement, the contamination of the anchor may induce artificial DIF and, thus lead to a seriously inflated false alarm rate. New anchor methods are often judged by their ability to correctly locate a completely DIF-free (i.e., pure, uncontaminated) anchor (e.g., Wang et al., 2012). Thus, the authors take a brief look at the simulation results focusing on the aspect of anchor contamination for one exemplary setting of 45% unbalanced DIF items in this section and provide a more detailed discussion in the online supplement. Figure 3 (top row) depicts the proportion of replications where at least one item of the anchor was a simulated DIF item (top-left)—this is referred to as *risk of contamination*—and the proportion of simulated DIF items in the anchor when the anchor was contaminated (top-right)—this is referred to as *degree of contamination* together with the false alarm rates (bottom row), including only the replications that resulted in a contaminated anchor (bottom-left) next to those including only the replications that resulted in a pure (i.e., DIF-free) anchor (bottom-right). If none of these pure replications resulted, the respective false alarm rate is omitted.

The results showed the following: All methods that rely on tests with all other items as anchor (namely, the all-other, single-anchor-AO, four-anchor-AO, iterative-forward-AO, iterative-backward-AO) displayed risks and also degrees of contamination that did not or only slightly decrease with the sample size. The overall risk and degree level depended on the anchor length. Short anchors, e.g., displayed a lower risk of contamination compared with longer anchors. The corresponding false alarm rates with a contaminated anchor increased, as—with increasing sample size—the power of detecting artificial DIF (DIF-free items that displayed DIF due to the chosen anchor method) increased. Those methods that are built using the

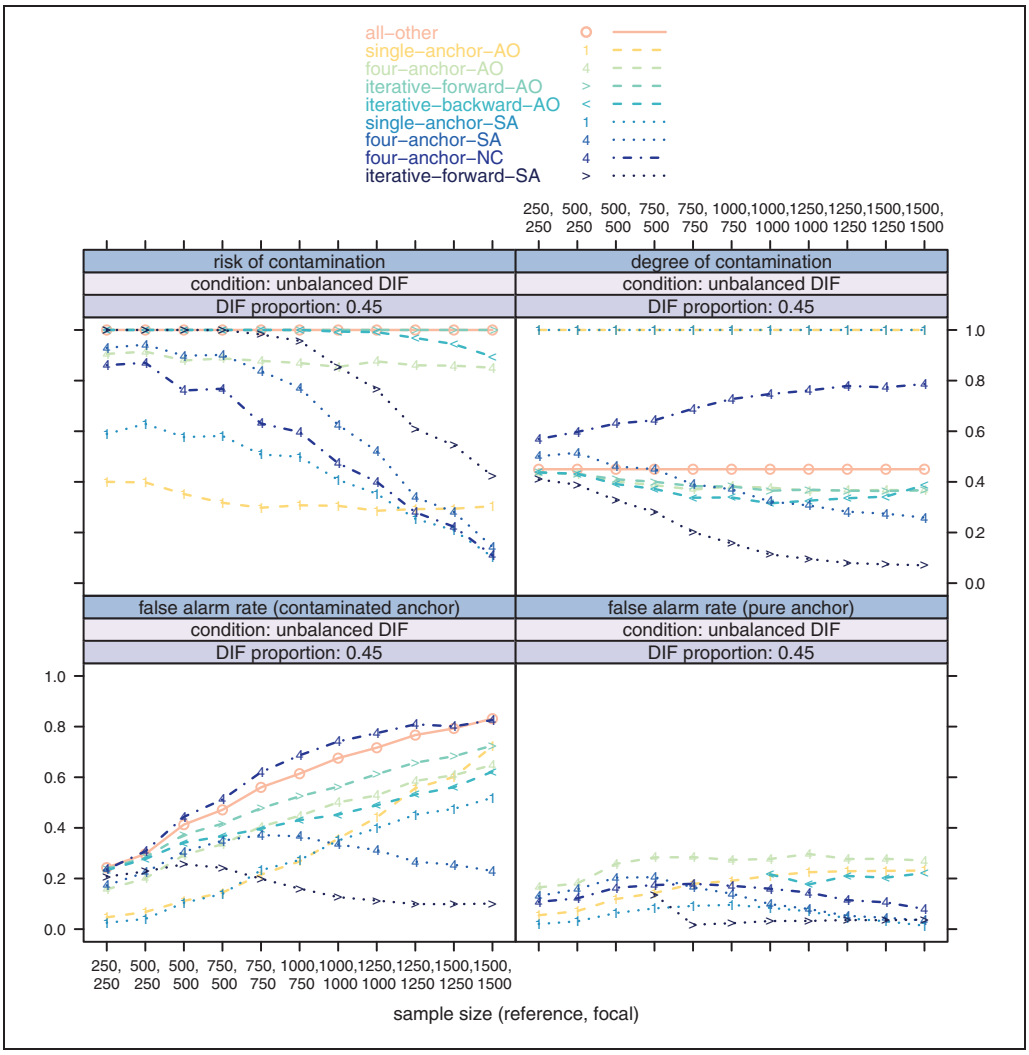

Figure 3. Condition of unbalanced DIF with 45% DIF items favoring the reference group; sample size ranges from (250; 250) to (1,500; 1,500); top-left: risk of contaminated anchors (at least one DIF item included in the anchor); top-right: degree of contamination (proportion of DIF items in contaminated anchors); bottom-left: false alarm rates when the anchor is contaminated; bottom-right: false alarm rates when the anchor is pure (not contaminated).

Note. DIF = differential item functioning; AO = all-other selection; SA = single-anchor selection; NC = next candidate selection.

SA-selection (namely, the single-anchor-SA, four-anchor-SA, iterative-forward-SA) showed risks and degrees that decreased with the sample size (except for the degree of the single-anchor). Their false alarm rates in contaminated replications were also lower when the sample size was high. An interesting finding here is the result for the four-anchor-NC method: It displayed a rapidly decreasing risk of contamination, but also a very high degree of contamination. As a consequence, the false alarm rate in contaminated replications was very high and even increased in the sample size. This explains the weak overall performance (see again Figure 2, top row right). This result makes clear that it is not the risk of contamination alone that

determines the performance of the anchor method. The iterative-forward-SA method (that performed best—in terms of a low false alarm rate together with a high hit rate—in this condition, see again Figure 2, right) displayed a higher risk of contamination but a lower degree of contamination compared with the four-anchor-NC method. The false alarm rate of the iterative-forward-SA method was low, independent of whether the anchor was contaminated or not (see Figure 3, bottom row). Thus, we conclude that research on anchor methods should not only concentrate on the risk of contamination but also focus on the consequences, which strongly depend on the degree of contamination, that is, the proportion of DIF items in the contaminated anchor. The second astounding finding, which is addressed in the next section, was that we observed false alarm rates exceeding the significance level, even in the case when only pure replications without anchor contamination were regarded (see Figure 3, bottom row right).

Characteristics of the Anchor Items Inducing Artificial DIF

In the simulation study, several anchor methods displayed inverse 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 inverse 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 the ''Impact of Anchor Contamination'' section, when the anchor was contaminated, and the authors provide a more detailed discussion of the contaminated replications in the online supplement. However, with this argument, the authors cannot yet explain why the false alarm rates showed a similar pattern for pure (uncontaminated) replications (see again Figure 3, bottom-right), where the single-anchor-SA, the four-anchor-SA as well as the four-anchor-NC method displayed inverse u-shaped false alarm rates. Therefore, the presence of artificial DIF induced by contamination alone cannot explain this finding. To understand this phenomenon, 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, the authors conducted an additional simulation study focusing again on the extreme condition of 45% unbalanced DIF items. Here, the authors examined the difference in the sum of the estimated anchor item parameters between focal and reference group that the authors termed *scale shift* (because it measures how far both scales of the item parameters are shifted apart during the construction of the common scale) for all constant four-anchor methods. To assess reliable estimates of the scale shift, the authors used all items that were DIF-free by design as anchor items to build the ideal common scale. The scale shift reflects the extent of artificial DIF and 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, the authors included a benchmark method of four-anchor items that were 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 4. The second

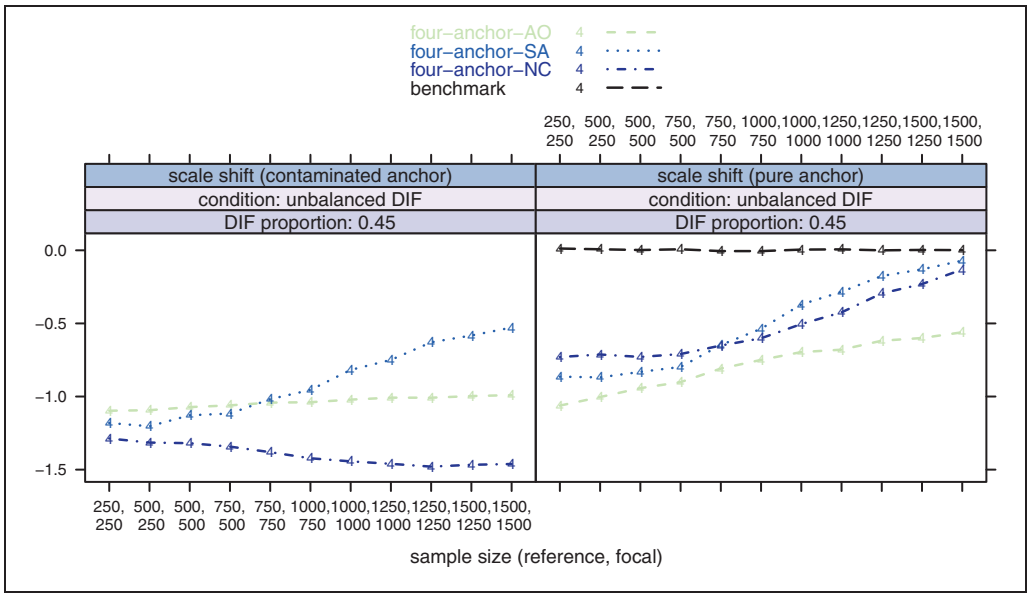

Figure 4. Condition of unbalanced DIF with 45% DIF items favoring the reference group; sample size ranges from (250; 250) to (1,500; 1,500); left: the scale shift when the anchor is contaminated; right: the scale shift in case of pure anchors.

Note. DIF = differential item functioning; AO = all-other selection; SA = single-anchor selection; NC = next candidate selection.

argument now becomes important in the case of pure anchors, which are discussed in more detail in this section: The scale shift for the benchmark method of randomly chosen DIF-free anchor items (Figure 4, right) fluctuated around zero and displayed no systematic shift in one direction. However, the scale shift of all remaining constant four-anchor methods was negative. This represents the fact that the supposedly pure items chosen by an anchor selection strategy displayed different characteristics than did randomly chosen pure anchor items. From all items that were ''pure'' by definition (i.e., were drawn from distributions with no parameter difference) the anchor selection strategies selected not the ones with the lowest empirical difference (due to random sampling), as one might hope, but those with a large empirical difference which induced artificial DIF for the other items. As can be seen from Figure 4 (right), the absolute scale shift for the four-anchor methods reduced with increasing sample size. In regions of large sample sizes, the absolute scale shift was directly related to the false alarm rate: When the absolute scale shift was high (as was the case for the four-anchor-AO method), the false alarm rate was high as well (Figure 3, bottom-right). In regions of smaller sample sizes, the scale shift of all four-anchor methods was high, but the false alarm rates were low at the beginning and then increased with growing sample size. When the scale shift decreased with growing sample size (e.g., for the four-anchor-SA method), the corresponding false alarm rate decreased as well and resulted in an inversely u-shaped pattern (see again Figure 3, bottom-right). Here, the interaction between the extent of artificial DIF—now induced by large empirical differences in the pure anchor items—and the power of detecting artificial DIF was visible that explained the false alarm rates.

Summary and Discussion

The assessment of DIF for the Rasch model based on the Wald test was investigated by means of hit and false alarm rates. Under the null hypothesis, all methods from the iterative forward and backward class as well as the all-other method held the significance level, while methods from the constant anchor class remained below that level. When DIF was balanced, the all-other method and also methods from the iterative forward and backward class yielded high hit rates while simultaneously exhausting the significance level. As expected, the AO-selection strategy outperformed the SA-selection strategy. In case of unbalanced DIF, the SA-selection procedure was superior to the AO-selection strategy when the sample size was large. The constant four-anchor class was not only combined with the AO-selection and the SA-selection strategy but also with the original NC-selection. Even though the four-anchor-NC method led to a low risk of contamination (see the ''Impact of Anchor Contamination'' section), it was outperformed by the four-anchor-SA method, which yielded lower false alarm rates and higher hit rates. In this unbalanced case, the newly suggested iterative-forward-SA method yielded the highest hit rate and a low false alarm rate and was, thus, the best performing anchor method. Based on these results, a careful consideration of the employed anchor method is necessary to avoid high misclassification rates and doubtful test results. Note, however, that the Rasch model, which was used for analyzing the data, was also the truly underlying data generating process. This assumption should be critically assessed in practical applications and future research should further investigate the separability of DIF and model misspecification. 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), the authors thus recommend to use the iterative-forward-SA method. When the sample size was large enough (above 1,000 observations in each group in the simulated settings), the false alarm rates were low in any condition even if the anchor was contaminated. Hit rates rapidly grew with the sample size and converged to one. The iterative-forward-SA method outperformed the iterative-backward-AO, iterative-forward-AO, the all-other as well as anchor methods from the constant anchor class by yielding a lower false alarm rate together with a higher hit rate. There are several reasons that explain the superior performance of the iterative-forward-SA method. First, the method has a head start compared with 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). The latter start with a criterion that is severely biased when DIF is unbalanced, whereas the iterative-forward-SA method does not require that DIF effects almost cancel out (for a discussion, see Wang, 2004). Second, the SA-selection strategy combined with the iterative forward anchor class also performed well in case of balanced DIF. Although the AO-selection strategy performed better than the SA-selection strategy when it was combined with the methods from the constant anchor class, the advantage in combination with the iterative forward class appeared negligible. Third, the 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 was long and, thus, often contaminated. The risk of contamination decreased with increasing sample size, and beyond that, the proportion of DIF items in the contaminated anchor (the degree of contamination) decreased. Fourth, 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.[4] Another astounding finding of the simulations presented here was that anchor items located by an anchor selection

strategy displayed different characteristics compared with 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. The reason for this is that a longer anchor, that contains some items that induce artificial DIF but also several items that do not, shifts the scales of the item parameters less strongly than a shorter anchor, where the proportion of items inducing artificial DIF is higher. The simulation study presented here was limited to DIF analysis in the Rasch model using the Wald test. Thus, future research (the interested reader is referred to the online supplement) may investigate the usefulness of the iterative-forward-SA method for other IRT models and combine it with other DIF detection methods.

Acknowledgment

The authors would like to thank Thomas Augustin for his expert advice and three anonymous reviewers for their very helpful and constructive feedback.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Julia Kopf is supported by the German Federal Ministry of Education and Research (BMBF) within the project ''Heterogeneity in IRT-Models'' (Grant ID 01JG1060).

Notes

1. In case all items were excluded from the anchor (which happened in only 7 out of 154,000 replications), one single-anchor item was chosen randomly in our simulation study.
2. 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 prespecified number of items. However, as in the subsequent simulation, it turned out that always four anchor items were selected for the final anchor, here the anchor class is classified as constant. Note that a significance level of .05 was used, but, of course, it would also be possible to choose, a higher level such as .30 as suggested by Wang (2004).
3. In addition to these item parameter values $\beta = (-2.522, -1.902, -1.351, -1.092, -0.234, -0.317, 0.037, 0.268, -0.571, 0.317, 0.295, 0.778, 1.514, 1.744, 1.951, -1.152, -0.526, 1.104, 0.961, 1.314, -2.198, -1.621, -0.761, -1.179, -0.610, -0.291, 0.067, 0.706, -2.713, 0.213, 0.116, 0.273, 0.840, 0.745, 1.485, -1.208, 0.189, 0.345, 0.962, 1.592)$, the main results with various other item parameter settings were replicated (results not shown). Therefore, the authors are confident that the different behavior of the anchor methods is not limited to the settings investigated here.
4. 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 located by the respective anchor selection strategy. 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 as the test results are only used for the decision whether the anchor should include one more anchor item. Thus, a small statistical power of the DIF tests in the first iterations automatically leads to a longer anchor that is expected to increase the power of the actual DIF test in the final step.

References

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

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

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19-29.

Edelen, M. O., Thissen, D., Teresi, J. A., 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*, 134-142.

Eggen, T., & Verhelst, N. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, *71*, 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*, 278-295.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models—Foundations, recent developments, and applications* (Chapter 2) (pp. 15-38). New York, NY: Springer.

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*, 432-457.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models—Foundations, recent developments, and applications* (Chapter 5) (pp. 69-96). New York, NY: Springer.

Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (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*, 32-44.

Jodoin, M. G., & Gierl, M. J. (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*, 329-349.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*, 251-265.

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

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*, 365-386.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161-173.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105-118.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models—Foundations, recent developments, and applications* (Chapter 3) (pp. 39-52). New York, NY: Springer.

Paek, I., & Han, K. T. (2013). IRTPRO 2.1 for Windows (item response theory for patient-reported outcomes). *Applied Psychological Measurement*, *37*, 242-252.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*, 235-259.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*, 184-199.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 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 & H. I. Braun (Eds.), *Test validity* (Chapter 10) (pp. 147-170). Hillsdale, NJ: Erlbaum.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221-261.

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*, 687-708.

Wang, W. C., & Su, Y. H. (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*, 113-144.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547.