universität innsbruck

Research Platform
Empirical and Experimental Economics


# Anchor Point Selection - Scale Alignment Based on an Inequality Criterion

**Carolin Strobl, Julia Kopf, Lucas Kohler, Timo von Oertzen, Achim Zeileis**

Working Papers in Economics and Statistics

2018-03

University of Innsbruck
https://www.uibk.ac.at/eeecon/

**University of Innsbruck**
**Working Papers in Economics and Statistics**

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
research platform “Empirical and Experimental Economics”
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 71022
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/eeecon/wopec/

For a list of recent papers see the backpages of this paper.

# Anchor Point Selection – Scale Alignment Based on an Inequality Criterion

**Carolin Strobl**
Universität Zürich

**Julia Kopf**
Universität Zürich

**Lucas Kohler**
Universität Zürich

**Timo von Oertzen**
Universität der Bundeswehr München

**Achim Zeileis**
Universität Innsbruck

**Abstract**

For detecting differential item functioning (DIF) between two or more groups of test takers in the Rasch model, their item parameters need to be placed on the same scale. Typically this is done by means of choosing a set of so-called anchor items based on statistical tests or heuristics. Here we suggest an alternative strategy: By means of an inequality criterion from economics, the Gini Index, the item parameters are shifted to an optimal position where the item parameter estimates of the groups best overlap. Several toy examples, extensive simulation studies and two empirical application examples are presented to illustrate the properties of the Gini Index as an anchor point selection criterion and compare its properties to those of the criterion used in the alignment approach of Asparouhov and Muthén. In particular, we show that – in addition to the globally optimal position for the anchor point – the criterion plot contains valuable additional information and may help discover unaccounted DIF-inducing multidimensionality. We further provide mathematical results that enable an efficient sparse grid optimization and make it feasible to extend the approach, e.g. to multiple group scenarios.

*Keywords*: Differential item functioning (DIF), item bias, anchor items, item clusters.

## 1. Introduction

One of the major advantages of probabilistic test theory is that its assumptions are empirically testable. With regard to test fairness, a crucial step in test validation is to identify items that exhibit differential item functioning (DIF) for different groups of test takers. DIF items can lead to unfair test decisions and threaten the validity of the test (cf., e.g., Cohen, Kim & Wollack 1996; Magis & De Boeck 2011) as well as its acceptance from the side of the test takers and

This is a preprint of an article accepted for publication in *Applied Psychological Measurement*.

policy makers. Once DIF items are identified, they can be improved or excluded from the final test form (cf., e.g., Westers & Kelderman 1992). But, in order to identify them, first the item parameters of the groups need to be placed on the same scale in a way that allows to compare the individual item parameters between the groups. This is usually done by choosing a set of so-called anchor items.

A large body of literature has been discussing and investigating different strategies for selecting these anchor items for DIF testing, particularly for the Rasch model (see, e.g., Teresi & Jones 2016, for a recent and broad overview on anchoring and DIF testing techniques). Questions that are being addressed in this literature – but have not all been answered satisfactorily yet – include the choice of the number of items (also termed anchor length) as well as different strategies to select those items. We will only go into detail in the following about a few exemplary anchoring methods, namely those that we will later use as comparison methods in our simulation studies. These methods have been selected to represent different approaches for anchor item selection, that are either widely used or have shown high performance in previous studies.

### 1.1. Exemplary anchor item selection methods

The first example we want to treat in a little more detail here is the anchor method suggested by Woods (2009), that is classified as the "constant all other" method by the taxonomy of Kopf, Zeileis & Strobl (2015b). It is an anchor of fixed length, that is selected based on the "all other" strategy: In the initial step, each item is tested for DIF using all other items together as the preliminary anchor. For this and the following method, we will fix the anchor length to four items in our simulation studies.[1] The four items corresponding to the lowest ranks of the absolute DIF statistics from the initial step are then chosen as the final set of anchor items. This method represents a commonly used and simple approach. However, by using all remaining items as the anchor in the initial step, this method (just like the similarly common "equal mean" approach, e.g., Magis & De Boeck 2011) assumes that DIF is balanced and cancels out over the items. If, however, DIF is not balanced, this strategy has been shown to exhibit a severely increased false alarm rate (Kopf, Zeileis & Strobl 2015a; Kopf et al. 2015b).

Our second example is the "constant four mean p-value threshold" (later abbreviated as "constant

[1]The literature on the anchor length shows that a too short anchor decreases the power of the following DIF tests, while a too long anchor increases the risk of a contaminated anchor (i.e., an anchor that includes DIF items), which can lead to artificial DIF (see also Andrich & Hagquist 2012). An anchor length of three to five, most often four, items has been suggested as a compromise (cf. Shih & Wang 2009; Wang, Shih & Sun 2012; Egberink, Meijer & Tendeiro 2015). Woods (2009) provides a more thorough discussion of anchor length choice, but this is not the focus of our study.

four MPT") anchor method suggested by Kopf et al. (2015a). Its selection of four anchor items is based on the number of p-values that exceed a threshold p-value determined from preliminary DIF tests for every item with every other item (one at a time) as a single anchor item (for a more detailed description see Kopf et al. 2015a). This method, together with the one described next, has been shown to be one of the two top performing methods in the extensive comparison study of Kopf et al. (2015a) and thus serves as a strong competitor here.

Another example is the "iterative forward mean test statistic threshold" (later abbreviated as "iterative forward") anchor method suggested by Kopf et al. (2015b). This method iteratively selects an anchor of variable length in a step-by-step procedure. The order in which new items are included in the anchor is determined by the mean test statistic threshold criterion. The rationale behind this crierion is that those DIF tests where the anchor is truly DIF-free should display the least absolute mean test statistics. Note, however, that the definition of the threshold depends on the assumption that the majority of the items are DIF-free. More details are provided in Kopf et al. (2015a).

### 1.2. Scale indeterminacy and anchoring for DIF detection

Going back one step in our reasoning, the fact that an anchor has to be chosen in the first place is due to the scale indeterminacy of the Rasch model (see, e.g., Fischer & Molenaar 1995). Anchoring solves this indeterminacy in a way that allows the item parameters of the groups to be compared in order to detect DIF. This is usually achieved by placing the same restriction on the item parameters in both groups (as formalized, e.g., by Glas & Verhelst 1995; Eggen & Verhelst 2006) in order to define a common scale. The reason why we need anchoring is that it is necessary to separate DIF from true differences in the mean abilities between the groups (often termed impact) by means of somehow conditioning on an estimate of the ability (Lord 1980; Van der Flier, Mellenbergh, Adèr & Wijn 1984; DeMars 2010). From a practical point of view, we cannot know in advance which items are the ones that have DIF and which are the ones that do not. Ideally, those items that end up being selected into the anchor should be DIF free, because otherwise the false alarm rate of the DIF tests increases, as shown, e.g., by Wang et al. (2012), but in practice there is no way to check in an empirical setting whether the anchor selection worked properly.

In the DIF literature, we can find several assumptions and notions about DIF and DIF detection that are not always made very explicit. For example, anchoring methods may only work properly if DIF is balanced (as discussed for the "all other" and "equal mean" strategies above), or assume implicitly or explicitly that the majority of items is DIF-free (as discussed explicitly by Kopf

et al. (2015a), but implicitly underlying several other anchor methods as well). Note that in a real data analysis (as opposed to a simulation study, where the DIF structure is known) neither assumption can be approved or empirically tested a priori, so that users should critically assess whether these assumptions are plausible in their case and how grave the consequences of a deviation from the assumption would be (such as a severely inflated false alarm rate for the equal mean and constant "all other" methods in case the DIF is not balanced). We will further discuss below that the assumption that the majority of items is DIF-free may seem particularly plausible for many tests, because we know how much time and effort the content experts have spent on putting the items together. However, from a methodological point of view it may restrict our theoretical thinking about the general concepts of anchoring and DIF, and is also critically discussed by Bechger & Maris (2015) and Pohl, Stets & Carstensen (2017).

When DIF is considered from the point of view of multidimensionality (e.g., Ackerman 1992; Roussos & Stout 1996), it becomes clear that the assumption that the majority of items is DIF free corresponds to the assumption that the majority of items measure the primary dimension of interest, and nothing else. In this framework, it has been shown that DIF can result from secondary dimensions, for which the distributions of two groups of test takers differ (for details see Ackerman 1992; Roussos & Stout 1996), and which some of the items measure in addition to the primary dimension. If only few individual items measure secondary dimensions, this is perfectly in line with the assumption that the majority of items measures the primary dimension and should be considered DIF free. If we think of scenarios, however, where clusters of items measure the same secondary dimensions (including scenarios where the primary dimension no longer provides the majority of items, as will be illustrated below), it would be helpful to be able to detect this kind of pattern.

### 1.3. Outlook on the contents of this manuscript

In this manuscript we will follow an approach that is different from the "traditional" anchor item selection methods described above. Rather than assessing only certain combinations of anchor items, the idea of this approach is to align the two scales by optimizing an objective function, that captures the discrepancy between the scales along a continuum of potential anchor points. Specifically, we propose to maximize the inequality of item-wise absolute distances – captured by the so-called Gini Index, an inequality criterion from economics – in order to find an anchor point where very few items (if any) exhibit DIF, while most other items do not. This approach has turned out to be closely related to – but was developed independently of – the alignment method by Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014), who employ the

so-called component loss function as the criterion for selecting an optimal anchor point.

Although the motivation of both approaches was to select anchor points without using anchor items, it turns out somewhat surprisingly that optimal anchor points may in fact correspond to single anchor items. This is shown mathematically for both criteria, the Gini Index and the component loss function, for the case of a Rasch model based on conditional maximum likelihood (CML) estimation in two groups. Note that neither Asparouhov and Muthén's work, nor the previous version of our own manuscript (Strobl, Kopf, Hartmann & Zeileis 2018) had pointed out this property, which greatly facilitates searching for the optimal anchor point solution.

Despite this simple optimal solution, in the following we will first explain the general idea of shifting the scales along a continuum of potential anchor points, which establishes more broadly the idea of an optimal point where the item parameters best interlock. Moreover, in addition to the globally optimal solution, the pattern of potential local optima can be particularly informative, as will be shown in several illustrations.

Now we will first introduce a little notation and review the fundamentals of anchoring. Then the new approach for finding anchor points will be introduced. Its usefullness will be illustrated by means of illustrative toy examples, an extensive simulation study as well as two application examples, where we will investigate DIF between female and male test takers. Due to space constraints many illustrations and results have been moved to online appendices, to which we wil refer the reader in due course. In particular, Appendix F provides the mathematical derivation of the possible locations of optima for both criteria. In the discussion, we will also point out the possibility to extend our approach to settings with more parameters and multiple groups.

## 2. Anchoring revisited

Due to its scale indeterminacy, i.e., the fact that the latent scale has no natural origin, a restriction is necessary for estimating the item parameters in the Rasch model. Commonly used restrictions are setting (arbitrarily) the first item parameter or the sum of all item parameters to zero (Glas & Verhelst 1995; Eggen & Verhelst 2006). When the aim is to compare the item parameters between two groups, the item parameters are first estimated separately. In the following, these initial item parameter estimates will be termed $\tilde{\beta}_j^{(g)}$ for group $g$ and item $j$. Since any linear restriction can easily be obtained from any other, it does not matter which particular restriction is applied in this first step. However, the restriction used for DIF detection is a critical choice, as illustrated in Appendix A.

## 2.1. The choice of the restriction

Considering the choice of a suitable restriction for comparing the item parameters of two groups, a variety of strategies has been suggested to choose a set of suitable anchor items. The sum of the item parameters of this set of anchor items is usually set to zero in both groups as the new restriction. In the following, we will introduce and explain some notation for describing the process of anchoring mathematically.

We start off with the initial item parameter estimates for each group, $\tilde{\beta}_j^{(g)}$. In the following, we will employ the CML approach for estimating the item parameters, but the general principle outlined here applies to any kind of item parameter estimates[2].

In our notation for describing the process of anchoring, let $\mathcal{A}$ denote the set of anchor items and $|\mathcal{A}|$ its cardinality, i.e., the number of anchor items in this set. The restriction that the sum of the anchor item parameters should be zero in group $g$ can then be expressed as $\sum_{j\in\mathcal{A}} \hat{\beta}_j^{(g)} \stackrel{!}{=} 0$. The final item parameter estimates $\hat{\beta}_j^{(g)}$ can be derived from the initial estimates $\tilde{\beta}_j^{(g)}$ by means of shifting all item parameters by

$$\hat{\beta}_j^{(g)} = \tilde{\beta}_j^{(g)} - \frac{\sum_{j\in\mathcal{A}} \tilde{\beta}_j^{(g)}}{|\mathcal{A}|}.$$

This shift ensures that the sum of the anchor item parameters is zero in each group. Of course all other item parameters are also shifted by the same amount, so that the overall pattern of the item parameters in each group is not altered, but moved as a whole to a position where it can best be compared to the pattern of item parameters in the other group.

More abstractly speaking, the process of anchoring corresponds to shifting all item parameters by a constant $c^{(g)}$

$$\hat{\beta}_j^{(g)} = \tilde{\beta}_j^{(g)} - c^{(g)},$$

where in all traditional anchoring approaches $c^{(g)} = c^{(g)}(\mathcal{A}) = \frac{\sum_{j\in\mathcal{A}} \hat{\beta}_j^{(g)}}{|\mathcal{A}|}$ depends on the choice of the anchor set $\mathcal{A}$ and can take all values that result from the different combinations of anchor items that are being in- or excluded in $\mathcal{A}$.

Conceptually, it is possible to uncouple the shift of the item parameters from a certain choice of anchor items. This can be accomplished by means of searching over an interval $[c_{\min}, c_{\max}]$ of values for $c^{(g)}$, including values that do not result from any specific combination of anchor items.

[2]Note, however, that other estimation approaches, such as marginal maximum likelihood estimation, may make it necessary to account for possible impact through the specification of the person parameter distribution, which may also affect their sensitivity for detecting certain DIF patterns (cf., e.g., Debelak & Strobl 2019).

Without loss of generality, rather than shifting the item parameters of both groups, we leave the item parameters of the first group at their initial estimates

$$\hat{\beta}_j^{(g_1)} = \tilde{\beta}_j^{(g_1)} \text{ with } c^{(g_1)} = 0,$$

where any arbitrary restriction can be used for the initial estimates $\tilde{\beta}_j^{(g_1)}$. The item parameters of the second group are then "moved past" the item parameters of the first group by means of shifting them by a constant $c$:

$$\hat{\beta}_j^{(g_2)} = \tilde{\beta}_j^{(g_2)} - c^{(g_2)} \text{ with } c^{(g_2)} = c.$$

For the boundaries of the interval $[c_{\min}, c_{\max}]$ we can then use values such that the item parameter ranges of both groups are safely overlapping:

$$[c_{\min}, c_{\max}] = \left[\min(\tilde{\beta}^{(g_1)}) - \max(\tilde{\beta}^{(g_2)}), \max(\tilde{\beta}^{(g_1)}) - \min(\tilde{\beta}^{(g_2)})\right].$$

This means that the item parameters of the second group are moved fully past the item parameters of the first group, starting where the lowest item of the first group interlocks with the highest item of the second and moving on until the highest item of the first group interlocks with the lowest item of the second group. Below we will see in detail how the shift constant $c$ can be selected based on the data by searching over this interval (or over a more sparse grid, as derived in Appendix F).

For the final DIF test, we will then look at a test statistic based on the difference between the final item parameter estimates of the two groups on the shifted scale: $\hat{\beta}_j^{(g_1)} - \hat{\beta}_j^{(g_2)} = \tilde{\beta}_j^{(g_1)} - \tilde{\beta}_j^{(g_2)} - c$. Note that this comparison depends on the choice of $c$, which we will select in a suitable way, but not on the choice of the initial restrictions, because the selection of $c$ will make up for any shift in the $\tilde{\beta}^g$.

A common choice of such a test statistic for the final DIF test is that of the item-wise Wald test

$$t_j = \frac{\hat{\beta}_j^{(g_1)} - \hat{\beta}_j^{(g_2)}}{\hat{\text{se}}_j} = \frac{\tilde{\beta}_j^{(g_1)} - \tilde{\beta}_j^{(g_2)} - c}{\hat{\text{se}}_j},$$

with $\hat{\text{se}}_j = \sqrt{\widetilde{\text{Var}}(\tilde{\beta}^{(g_1)})_{j,j} + \widetilde{\text{Var}}(\tilde{\beta}^{(g_2)})_{j,j}}$. Note that we apply the item-wise Wald test to the conditional maximum likelihood estimates in the following (like in Glas & Verhelst 1995; Kopf et al. 2015a,b).

When we reconsider the idea of moving the item parameters of the second group past those of the first group, for us as human beings it is straightforward that some positions are smarter than others, but the crucial question is: Can we find an objective criterion to make this decision for

us automatically – both to avoid subjectiveness in our decision and to make it computationally feasible?

At first sight it may seem like $c$ could be optimized directly with respect to a test statistic like that of the Wald test displayed above, or with respect to some kind of norm $||d(c)||$ of the vector $d(c) = (d_1(c), \ldots, d_m(c))^\top$ of the item-wise absolute distances on the shifted scale

$$d_j(c) = |\hat{\beta}_j^{(g_1)} - \hat{\beta}_j^{(g_2)}| = |\tilde{\beta}_j^{(g_1)} - \tilde{\beta}_j^{(g_2)} - c|.$$

Measures based on these distances could capture what could be called the *overall amount* of DIF, for example by using the sum of squared (Euclidean) or absolute (Cityblock) distances (corresponding to the L2 or L1 norm) as the criterion. However, a norm-based criterion could become large both if there are many small differences or a few large differences in the vector $d(c)$. For DIF detection and interpretation, however, these would have very different meanings.

We will show in the next section that DIF detection can better be achieved by applying a measure of *inequality* instead of a measure of the *overall amount* of DIF to $d(c)$ by using, for example, the popular Gini Index as our criterion.

## 2.2. The Gini Index

As an objective criterion for automatically selecting anchor points, we suggest to use the Gini Index (Gini 1955). The Gini Index is a popular inequality measure, that is usually employed for assessing the distribution of wealth or income between the members of a society. It takes high values if, for example, a small minority of persons has a lot of wealth while the vast majority has very little. It is therefore used to compare different countries with respect to their distribution of wealth or income (e.g., Central Intelligence Agency 2017).

We will now show how the Gini Index can also be used as a means for selecting anchor points. This is most easily imagined when the majority of items displays no DIF. Then at the optimal anchor point, where the scales for the two groups are aligned as well as possible, most items will interlock (i.e., they will lie on top of or very close to each other for the two groups), while a minority of items will differ for the two groups and show DIF. So while initially the Gini Index was used to indicate whether a minority of *persons* has a lot of *wealth* and the majority has very little, we will use it here to find solutions where a minority of *items* has a lot of *DIF* (i.e., large absolute differences in their item parameter estimates between the groups) while the majority has very little or no DIF (i.e., small or no absolute differences in their item parameter estimates between the groups).

The Gini Index can be computed as

$$\mathrm{GI}(c) = \frac{2 \cdot \sum_{j=1}^{m} r_j(c) \cdot d_j(c)}{m \cdot \sum_{j=1}^{m} d_j(c)} - \frac{m+1}{m},$$

where $r_j(c)$ is the rank of the absolute item-wise distance $d_j(c)$ for item $j$, with $j = 1, \ldots, m$.

The optimal anchor point based on the Gini Index then corresponds to

$$c_{\mathrm{maxGI}} = \underset{c \in [c_{\min}, c_{\max}]}{\arg\max}\ \mathrm{GI}(c).$$

The Gini Index can take values between 0 and close to 1. The value zero corresponds to perfect equality, i.e., all items having the same absolute item-wise distances, in which case they can be shifted such that the two groups are perfectly aligned and no item displays DIF.[3] Values close to one, on the other hand, correspond to perfect inequality, where one item has all the DIF (i.e., a high absolute item-wise distance) while all other items have no DIF at all. In this case, the Gini Index reaches its maximum possible value of $1 - \frac{1}{m}$. For example, if one out of ten items had DIF and the remaining nine items would have no DIF at all, its maximum would be $1 - \frac{1}{m} = 1 - \frac{1}{10} = 0.9$.

Note that the value of the Gini Index in this example depends only on the number of items, not on the absolute amount of DIF. This property of the Gini Index, that it is independent of the absolute amount of wealth (i.e., it does not measure the absolute effect size of DIF, but the strength of the inequality of the distribution of DIF among the items) is further illustrated below.

We will show how selecting $c$ according to the Gini Index leads to shifts between the two groups that makes their item parameters well comparable. This approach can serve as the basis for any kind of graphical display as well as for formal DIF tests. We will also see that the Gini Index is able to detect multiple clusters of items in the case of unaccounted DIF-inducing multidimensionality.

### 2.3. The component loss function criterion used by Asparouhov and Muthén

Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014), coming from a factor analysis background, describe that their alignment method was first motivated by the task to estimate group-specific factor means and variances for many groups at a time, which the authors explain is not feasible by means of modification indices (Asparouhov & Muthén 2014). As a by-product, the result can also be used for measurement invariance analysis, i.e., for detecting DIF.

[3]Note that by definition, the Gini Index would be undefined in a case where all distances are exactly zero, because mathematically this would lead to a division by zero. In our implementation, we have redefined its value to zero in this case, because it also represents perfect equality.

Asparouhov & Muthén (2014) introduce their approach in a factor analysis notation and framework, but in Muthén & Asparouhov (2014) show how it translates to the case of a 2-parameter logistic IRT model. Here we will refer to the so-called simplicity function and component loss function (CLF) used by Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014), that we will explain in detail below. We will adopt the CLF as an alternative criterion for selecting optimal anchor points in our framework based on CML estimation for the Rasch model. We will illustrate below that, compared to the Gini Index, it has similar properties in some but distinct properties in other DIF settings. Moreover, we show mathematically in Appendix F that both the Gini and the CLF Criterion can only find optima in single items in this particular framework, which makes the selection computationally much more feasible.

Note that the application of this criterion in our framework, based on CML estimation for the Rasch model, means that certain properties of Asparouhov and Muthén's approach, that was originally described for a 2-parameter model and for optimizing means and variances, may not carry over (in particular any effects of DIF affecting group variances). Yet, concentrating on this simple case allows us to concentrate on some fundamental properties of the criteria and compare the results to the extensive existing literature on DIF testing in the Rasch model.

We will now translate the simplicity function and CLF used by Asparouhov and Muthén into our notation. At the core of both our and Asparouhov and Muthén's reasoning is the idea to find a criterion that can be optimized such that "there are a few large noninvariant measurement parameters and many aproximately invariant measuremente parameters rather than many medium-sized noninvariant measurement parameters" (Asparouhov & Muthén 2014, p. 497). This aim corresponds exactly to our initial idea when using the Gini Index: to find solutions where a minority of items has a lot of DIF while the majority has very little or no DIF. Asparouhov & Muthén (2014) motivate their approach by earlier suggestions for criteria for finding simple structure solutions in factor rotation. We will show in the following that inequality criteria behave very similarly and argue that it may be fruitful to further explore the mathematical and philosophical similarities and specifics of the criteria used here, as well as potential further criteria from both research areas.

Because Asparouhov & Muthén (2014) consider a 2-parameter model, their simplicity function $F$ (Equation 9, Asparouhov & Muthén 2014, p. 497) consists of a sum over both types of parameters (as well as over multiple groups). In the simpler case of the Rasch model with only one parameter (and the case of two groups), in our notation with $d_j(c)$ again representing the absolute distances in the difficulty parameters of item $j$ between the two groups at point $c$ in

the search grid, the simplicity function becomes

$$F(c) = \sum_{j=1}^{m} f_\epsilon \left(d_j(c)\right),$$

with $f_\epsilon \left(d_j(c)\right)$ denoting the CLF.

Asparouhov and Muthén use the particular form

$$f_\epsilon \left(d_j(c)\right) = \sqrt{\sqrt{d_j(c)^2 + \epsilon}}$$

for the CLF, where the small positive constant $\epsilon$ is only added to ensure continuous differentiability for making the optimization easier. Since we use a grid-based rather than a gradient-based search for the optimal value of $c$, it is not necessary to add $\epsilon$. We will therefore use the simplified CLF

$$f \left(d_j(c)\right) = \sqrt{\sqrt{d_j(c)^2}} = \sqrt{d_j(c)}$$

throughout this manuscript for mathematical coherence and simplicity. We have checked that this makes no notable difference for any of the empirical results. The use of this particular form of the CLF is motivated by Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014) through its being a "good choice" among component loss functions, that are being used in exploratory factor analysis to find rotations to simple structure solutions. The optimal anchor point based on the CLF then corresponds to

$$c_{\text{maxCLF}} = \underset{c \in [c_{\min}, c_{\max}]}{\arg\max} - \sum_{j=1}^{m} f \left(d_j(c)\right).$$

Note that throughout the main part of this manuscript we maximize and display $-\sum_{j=1}^{m} f \left(d_j(c)\right)$ (rather than minimizing $\sum_{j=1}^{m} f \left(d_j(c)\right)$ like Asparouhov and Muthén) and refer to this as the CLF Criterion in the following. We do this so that, both for the Gini Index and the CLF Criterion, larger values correspond to more inequal distributions of DIF, and maxima can be interpreted as optimal solutions. The shape of both criteria is illustrated in Appendix B.

Another difference to the original approach by Asparouhov and Muthén is that they concentrate on the use of the CLF in an automated process for the detection of a single global optimum, viewing multiple local optima as more of a nuisance, while we argue that those situations where distinct local optima occurr in addition to the global optimum are worth exploration by content experts. This is why, in addition to aggregated results, we will later display plots of the criterion values over the entire search interval for both the Gini Index and the CLF Criterion.

We would also like to point out that related notions of additional restrictions to be placed on the item parameter estimates to enhance comparability (e.g. Glas & Verhelst 1995; von Davier & von Davier 2007) and more or less algorithmic approaches for deciding which item parameters

should be allowed to differ between groups (e.g. Yamamoto, Khorramdel & von Davier 2013; Glas & Jehangir 2014; Oliveri & von Davier 2014) have been suggested and used for a long time by other authors. Pokropek, Lüdtke & Robitzsch (2020) also point out the connection to the literature on linking and equating. However, the approach of Asparouhov and Muthén is most closely related to the one presented here. Therefore in the following we will concentrate on investigating the properties of the Gini Index and the CLF Criterion for selecting anchor points in greater detail.

## 3. Illustration of the properties of Gini Index and CLF Criterion

The remainder of this paper, together with the Appendices C through E, provides several illustrations of the properties of the new anchor point selection approach based on the Gini Index under a variety of settings. First, we will show by means of a few toy examples how both the Gini Index and the CLF Criterion detect the optimal shift value for aligning the item parameters in both groups. Second, we will present results from an extensive simulation study, where these methods are compared to each other as well as to existing anchoring approaches from the literature. By means of additional illustrations we will show that the Gini Index shows a behavior that reflects our earlier considerations and that its criterion plot reflects additional information about the underlying pattern of the items particularly well. In Appendix E we will illustrate the practical usage of the approach by means of two empirical examples.

We will first consider a very simple DIF pattern in order to highlight the properties of the two criteria we can use for anchor point selection. In this first example we have simulated ten items, of which one item (item 4) has been simulated with DIF. We will first look at an illustration for the true values of the item parameters, i.e., without sampling variability, to highlight the general properties of the criteria.

Figure 1 (top left) shows the criterion plot of the Gini Index and the CLF Criterion over a grid of values for possible shifts $c$.[4] We see that both the Gini Index and the CLF Criterion have their global optimum at the same shift value of 0. The item parameter locations that correspond to this global optimum are displayed in Figure 1 (right column, top for CLF Criterion, bottom for Gini Index). We find that both criteria agree on a solution where all items but the fourth

[4]The values on the y-axis have been normalized for both criteria in order to be able to compare their shape in one plot. Note also that in this and the following illustrations of the criterion plots, we use an extensive grid over the search interval for better visibility. In the simulation studies, on the other hand, we use the sparse grid described in Appendix F for computational efficiency, as it has been shown mathematically to contain all possible locations for optima.

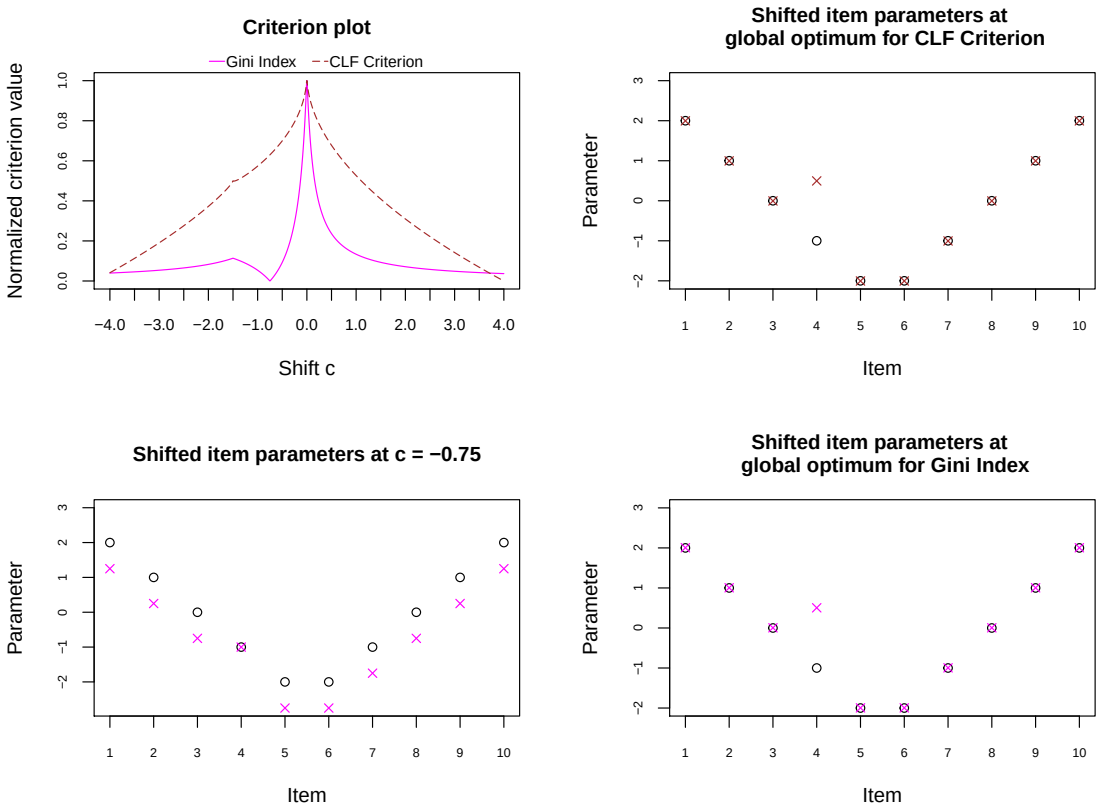


Figure 1: Criterion plot (top left), shifted item parameters according to global optima (right column) and shifted item parameters according to local maximum (bottom left) for toy example with one item displaying DIF of size 0.75, based on true item parameters.

item interlock, i.e., only item 4 shows DIF.

In the criterion plot in Figure 1 (top left), both criteria also show a smaller, local peak at the shift value $-0.75$, that is more notable for the Gini Index. The location of this second peak corresponds to a solution where the fourth item would interlock and all other items would show DIF, as illustrated in Figure 1 (bottom left). In this easy setting, both criteria – and we assume most readers – agree that the first solution, where item 4 is labeled to have DIF while all other items have no DIF, is preferable. However, we will later see scenarios where the decision is not so clear cut.

Additional illustrations are provided in Appendix C.

## 4. Simulation studies

We now present the results of two extensive simulation studies, where we compare the performance of the Gini Index and CLF Criterion to each other and to that of the three anchoring

methods from the literature that have been described above. For space constraints, here in the main text we present only a brief summary of the simulation setup and the key results. All further details are provided in Appendix D.

## 4.1. Simulation study I

*Simulation design*

The simulation design for this first study was chosen to be very similar to that of Kopf et al. (2015a) to ensure comparability with this extensive comparison study. We simulated data sets for two groups of subjects, the reference and the focal group, under the Rasch model. In most of the scenarios, a certain percentage of the items was simulated to show DIF between the groups. The direction of DIF was either balanced or unbalanced. There are also scenarios that were simulated completely under the null hypothesis with no DIF in any item. In each setting, 10000 replications were simulated.

*Results*

In the following we will report the false alarm rate, that is computed as the percentage of items that were simulated as DIF free, but erroneously show a significant test result, and the hit rate, that is computed as the percentage of items that were in fact simulated to have DIF and correctly show a significant test result.

First we checked the false alarm rates in the null case scenario where no DIF items were generated. The false alarm rates should correspond to the nominal type I error rate of 5%. Our results (omitted to save space) show that all methods roughly hold or fall below this nominal type I error rate in the null case scenario.

Now we will look at the results for scenarios with unbalanced DIF favoring one group. Figure 10 (first row) shows the false alarm rates, zoomed false alarm rates and hit rates for all methods in a scenario with a testlength of 40 items and 20% of these items being simulated with DIF. The results show that for this scenario all methods except for the "all other" method hold the nominal type I error rate. For the "all other" method, the false alarm rate notably increases with the sample size. This effect has already been discussed as a known problem of the "all other" method in unbalanced DIF settings in the introduction. All methods show hit rates that increase with the sample size as expected. The "iterative forward" method shows the highest hitrate, followed by the "constant four MPT" method, the Gini Index, the "all other" method and the CLF Criterion.

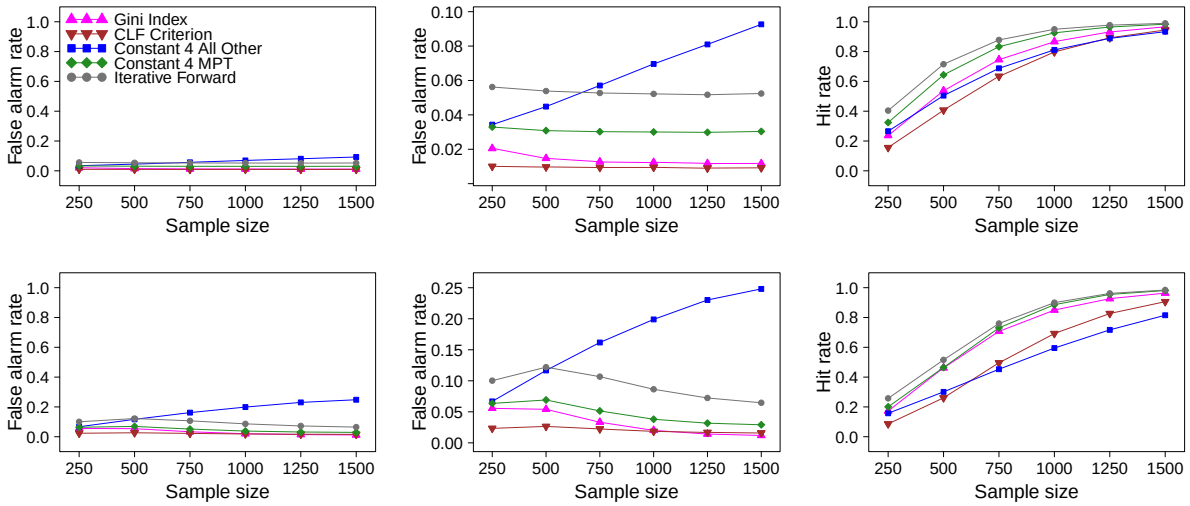


Figure 2: False alarm rates (y-axis from 0 to 1; left column), zoomed false alarm rates (y-axis from 0 to highest value; middle column) and hit rates (y-axis from 0 to 1; right column) for scenario with 20 percent (first row) and 40 percent (second row) DIF items favoring the focal group.

When the percentage of DIF items rises to 40% in the next scenario displayed in Figure 10 (second row), we note that the false alarm rates of some methods increase. Most notably, for the "all other" method, the false alarm rate increases even more strongly with an increasing sample size and goes up as high as 25%. For the other anchoring methods, as well as to a lesser degree for the Gini Index, we see a pattern where the false alarm rates show a slight inversely u-shaped pattern, that was similarly observed and explained by Kopf et al. (2015b). For the "iterative forward" method this also leads to a false alarm rate notably above the nominal 5% level for small and medium sample sizes, so that we should also interpret the hit rate of this method with caution. The hit rates of all methods increase with increasing sample size as expected. Again we find that the "iterative forward" method shows the highest hitrate (but also an increased false alarm rate), followed by the "constant four MPT" method and the Gini Index, and with some distance by the CLF Criterion and the "all other" method.

In addition to the first two DIF scenarios, where a minority of 20% or 40% of the items were simulated with DIF, we now consider a scenario where the majority of items, 60%, are simulated with DIF in favor of the focal group. Since these items are simulated with DIF of the same amount, they work together as a cluster that is in itself invariant.

When we would stick to the definition of the simulation design for this setting, the false alarm rates for most methods would strongly increase, while the hit rates would decrease similarly dramatically, because the methods would consider the majority cluster as the DIF free one.

However, as discussed above and further illustrated in Appendix D, from a philosophical point of view both solutions – considering the smaller or the larger item cluster as DIF free – are equally valid. We show in Appendix D that the Gini Index is particularly suited for identifying both solutions. Here we see a strong parallel to the works of Bechger & Maris (2015) and Pohl et al. (2017), who also critically discuss the general assumption that the majority of items is DIF free and instead aim at the detection of invariant item clusters.

In the simulation study, where we need to decide on a scoring rule to be able to compute the aggregated false alarm rates and hit rates, this reasoning cannot be entirely transported, but we can try to mimic it by using a scoring rule that counts either solution as correct. We refer to this scoring rule as "label-switching", because it resembles the fact that in cluster analysis we want to judge whether observations correctly end up in the same cluster in two runs, but the labeling of the clusters is arbitrary. When this label-switching scoring rule is used for computing the false alarm rates and hit rates for all methods (Figure 13), we see that the results return to what we saw for lower percentages of DIF items, namely that the methods show slightly increased (for the "iterative forward" method) or acceptable false alarm rates and increasing hit rates (except for the "all other" method, that has trouble with the unbalanced setting in general). The Gini Index now shows the highest hit rates, in particular notably higher than the CLF Criterion, as is further explained in Appendix D.

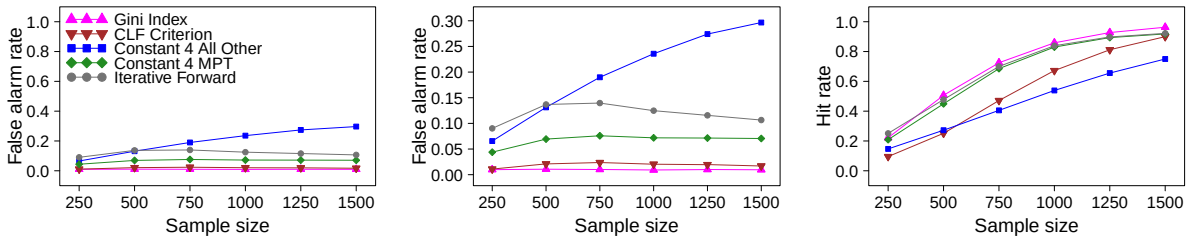


Figure 3: False alarm rates (y-axis from 0 to 1; left), zoomed false alarm rates (y-axis from 0 to highest value; middle) and hit rates (y-axis from 0 to 1; right) for scenario with 60 percent DIF items favoring the focal group and label switching allowed.

Appendix D also shows additional interesting results for the case of balanced DIF. In this setting both Gini Index and CLF Criterion were outperformed by the traditional anchor selection methods, some of which are particularly well suited for balanced DIF. However, our findings for this setting also futher support our notion that the globally optimal solution does not tell the whole story, and that solutions corresponding to local optima in the criterion plot should also be explored to better understand the DIF structure in the data.

### 4.2. Simulation study II

*Simulation design*

In order to further illustrate the connection between DIF and multidimensionality, we have conducted a second simulation study, that employs a multidimensional IRT model for data generation. The design of this study resembles the design for unbalanced DIF in Simulation study I as presented above. While there unidirectional DIF was generated by adding a fixed amount of DIF to certain item parameters, now the DIF is induced by letting certain items measure a secondary dimension in addition to the primary dimension (like described, e.g., in Roussos & Stout 1996). For details see again Appendix D. In each setting, again 10000 replications were simulated.

*Results*

As expected, we find the results (displayed in Appendix D) to be very similar to those in Figures 10 and 13 for Simulation study I, with only slightly higher false alarm and hit rates in some places. In particular, we see that the methods are again able to identify the pattern in the items even when the majority of items measures the secondary dimension when the label-switching scoring rule is applied. The Gini Index, together with the "iterative forward" and "constant four MPT" methods, again performs particularly well in this setting.

## 5. Empirical application examples

Appendix E provides two empirical application examples: One with a clear global optimum and one with an additional local optimum indicative of a DIF-inducing secondary dimension.

## 6. Summary and discussion

In this paper we have suggested a new approach for placing the item parameter estimates of a Rasch model for two groups of test takers on the same scale. We have suggested to use the Gini Index, an inequality criterion from economics, as the optimization criterion, and have compared its properties to those of the CLF Criterion, that is used in the alignment method by Asparouhov and Muthén.

We have shown by means of extensive simulations, illustrative toy examples and two application examples that the anchor point selection approach is able to identify locations on the item parameter continuum where the item parameter estimates for the two groups best overlap, and

that there can be more than one sensible solution. Therefore we recommend that, rather than reporting only the globally optimal solution, the entire criterion plot should be reported and inspected, because it provides valuable additional information about the item structure. This information should be taken into account for the decision how to proceed with specific test items.

Asparouhov & Muthén (2014) have stated that "[...] [I]f data are generated where a minority of the factor indicators have invariant measurement parameters and the majority of the indicators have the same amount of noninvariance, the alignment method will choose the noninvariant indicators as the invariant ones, singling out the other indicators as noninvariant." This corresponds exactly to the label-switching situations we have discussed above. However, we believe that, rather than considering this as a weakness of either aproach, we should consider it as an advantage and utilize the information on multiple solutions contained in the criterion plots.

The recent paper of Pokropek et al. (2020) investigates the effect of using different powers in the CLF – where Asparouhov and Muthén use a power of $\frac{1}{2}$ for the square root – and observe that for powers smaller than one, that show superior results in their study, local optima can occur. In our illustrations and simulation studies we have found local optima for both the Gini Index and the CLF Criterion, where those for the Gini Index were more distinct. While in many cases both criteria agreed on the global optimum, we also observed situations where this was not the case.

Depending on the simulation setting, the anchor point selection based on the Gini Index, that was suggested in this manuscript, performed equally well or even slightly better than existing anchor selection methods. However, both the Gini Index and the CLF Criterion were outperformed in the balanced DIF setting, for which some of the competitor methods are particularly well suited. Here it might also come into play that our mathematical results show that both the Gini Index and the CLF Criterion select single item anchors in the framework considered here. Single item anchors are more heavily affected by sampling error and the literature on anchor length implies that too short anchors diminish the power of the resulting DIF tests.

On the other hand, the fact that mathematically the set of possbible solutions is limited to single-item solutions makes it computationally easily feasible to extend this approach, e.g., to pairwise comparisons of multiple groups of test takers (such as several different language groups). In future research, we will also explore extensions to more general IRT models with different types of item parameters. Both extensions are possible for the Gini Index in the same way that is employed for the CLF criterion in Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014).

It should also be noted that we have seen in the illustrations of our results that, despite mathematically corresponding to single item anchors, all solutions represented as global or local optima in the criterion plot are well interpretable graphically and can help identify clusters of items representing, e.g., DIF-inducing secondary dimensions.

An interesting line of future research would be to compare the results of the anchor point selection approach to approaches that explicitly aim at identifying item clusters, such as Bartolucci (2007), Pohl et al. (2017), Pohl & Schulze (2020) and Schulze & Pohl (2020). As already mentioned above, the approach of Pohl and colleagues, that is based on the work of Bechger & Maris (2015) and the notion of differences in relative item difficulties, is closely related in philosophy to the approach presented here. We would expect that their item clusters should largely agree with our solutions corresponding to global or local optima, and believe that the Gini Index as an intuitive criterion, together with the possibility to graphically display the criterion plot, will be particularly helpful for test developers in understanding the patterns in their data and guiding their decisionmaking.

## Computational details

Our results were obtained using the R system for statistical computing (R Development Core Team 2019), version 3.6.2. Anchor point selection will be made available in the R package `psychotools`. For the Gini Index we used the implementation from the R package `ineq` (Zeileis 2014), for model fitting and DIF tests we employed existing functionality from the R package `psychotools` (Zeileis, Strobl, Wickelmaier, Komboz & Kopf 2020). Simulation Study II used the R package `mirt` (Chalmers 2012) for data generation.

## Acknowledgements

This research was supported in part by the Swiss National Science Foundation (00019_152548).

The authors would like to thank Raphael Hartmann and Tasnim Hamza for their work on previous versions of the R code for Simulation study I, Rudolf Debelak for a jump start on using `mirt` for Simulation study II, Matthias von Davier for pointing us to important references, the editors and anonymous reviewers for their very constructive feedback, as well as Thomas Augustin for his encouragement when this idea first came up a very long time ago.

# Online Appendices

## A. Illustration of the effect of restrictions

The initial item parameter estimates are obtained using an arbitrary restriction, typically setting the first item parameter or the sum of all item parameters to zero. However, as is illustrated in Figure 4, if these initial item parameter estimates were naively used for a comparison between the two groups, the choice of the restriction would indeed affect our conclusion. This example was set up such that the first three item parameters are the same for both groups while the fourth item parameter differs between the groups. This is obvious in the first column a) of Figure 4, where in the top row a direct comparison of the item parameters and in the bottom row the setup of a graphical test (Rasch 1960; Wright & Stone 1999) is displayed. In this first column a), the first item parameter is arbitrarily set to 0 in both groups. In the top row of plots, an item displays DIF if the item parameters of the two groups (symbolized by circles and crosses) do not interlock. In the graphical test in the bottom row of plots, an item displays DIF if it is not located on the diagonal. (To account for estimation error, significance tests and confidence ellipses have been suggested for the graphical test, but here for simplicity we only focus on the location of the item parameters and act as if their true values were known.) Considering the selection of anchor items, we see that item 1 was a good choice here because it shows no DIF itself and can be safely used to compare the other items.

In the second column b) of Figure 4, however, a different restriction was used: Here the sum of all item parameters was set to zero in both groups. In anchor terms, this would mean that all items were included in the anchor. Due to the DIF in item 4, this anchor is contaminated. When the scales are shifted according to this anchor, the between-group distance in item 4 decreases, but at the cost of all other items' distances increasing artificially.

Even more extremely, when the parameter for item 4 is set to 0 in both groups in the third column c) of Figure 4, it looks like item 4 had no DIF, but all other items now exhibit the amount of DIF originally inherent in item 4. Most readers would agree that this is not a good choice and all traditional anchor selection approaches would try to avoid this scenario. However, had this been our initial arbitrary restriction for estimating the item parameters, and had we not investigated its effect, we could have come to a very different conclusion than before.

At this point it is important to note that our interpretation of which conclusion is right or wrong strongly depends on the abovementioned assumption that it is a minority of items that exhibit DIF, not the majority. Without any additional assumption, given the scale indeterminacy, it

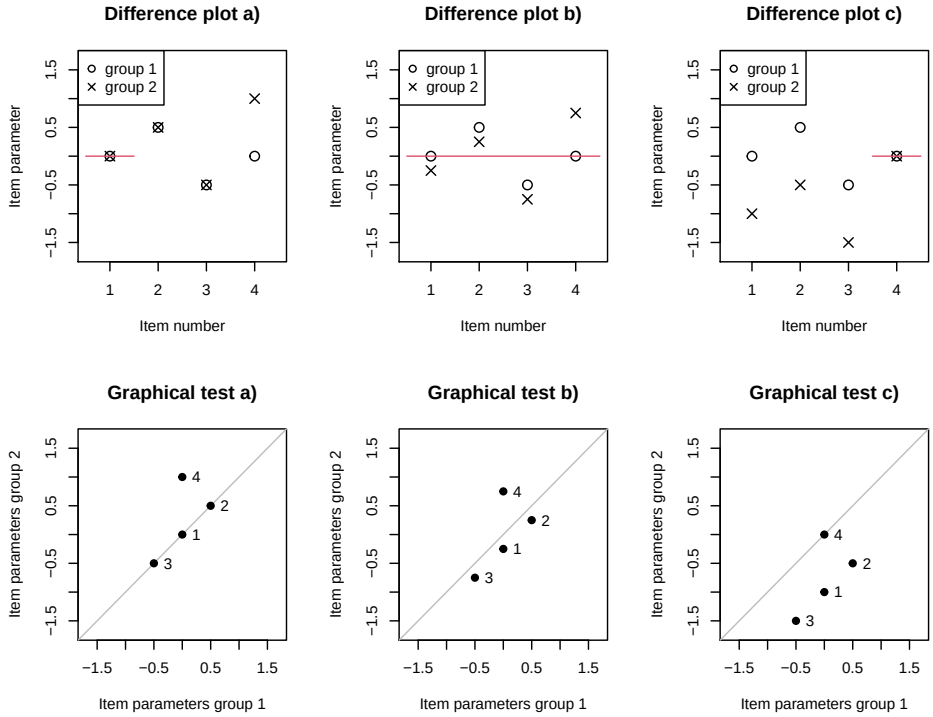


Figure 4: Illustration of comparisons of item parameters (top row) and graphical tests (bottom row) for different restrictions: a) $\beta_1^{(g)} = 0$, b) $\sum_j \beta_j^{(g)} = 0$, c) $\beta_4^{(g)} = 0$. Item numbers are displayed on the x-axis in the top row and next to the plotting symbols in the bottom row.

would not be possible to decide which scenario is the correct one.

As a side note, the didactically very well written textbook of Wright & Stone (1999) also implicitly follows this assumption. On p. 62 it shows an example of a graphical test where some items exhibit DIF. There, the authors move the original identity line, that seems to have been based on the arbitrary restriction used for the item parameter estimation, towards the location of the majority of items. The "second identity line" of Wright & Stone (1999) is exactly what a sensible anchoring approach would produce in this situation (even if the authors do not yet use this terminology and it sounds like the line was manually placed through the "major item stream").

## B. Comparison of the shape of the criteria

Before the properties of both criteria for DIF detection are illustrated and compared to those of traditional anchoring approaches, let us display the shape of both criteria as a function of the between-group absolute distances $d_j(c)$ for two items in Figure 5. We can see that the two criteria have similar but not identical shapes: The Gini Index reaches its maximum (visible as "rooftops" in Figure 5, left) in all points where one item parameter has an absolute between-

group distance of zero while the other item parameter has a value different from zero, which corresponds to the most inequal situation of one item being DIF free and the other item having "all the DIF" in a setup with only two items. On the diagonals, where the absolute between-group distances of the two items are equal, the Gini Index has the lowest values. Here the DIF would be equally distributed between all items, which would not be a reasonable choice for anchoring. This illustrates that the Gini Index behaves in a way that corresponds well to our intuition of DIF.

The CLF also shows relatively high values (visible as ridges of the "tent" in Figure 5, right) in points where one item parameter has an absolute between-group distance of zero while the other item parameter has a value different from zero, which corresponds to the most inequal situation of one item being DIF free and the other item having "all the DIF". Note, however, that the CLF Criterion produces higher values for solutions with lower absolute distances in the other item and does not decrease as strongly as the Gini Index on the diagonals where the absolute between-group distances of the two items are equal.

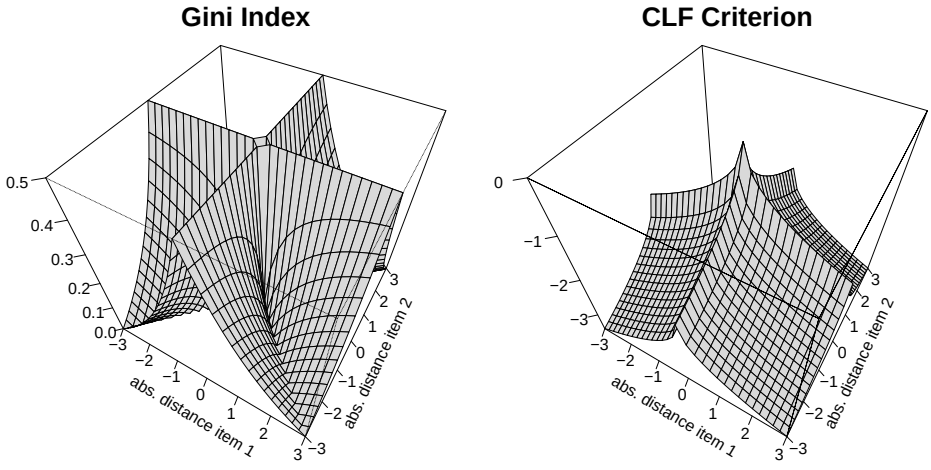


Figure 5: Shape of Gini Index and CLF Criterion for two items.

## C. Additional illustrations of the properties of the criteria

In addition to the illustration with the true item parameters in the main text, we now simulate item response data from these item parameters and display the results again for the estimated parameters. This gives a more realistic impression of the effect of sampling variability. Figure 6 (top left) again shows the criterion plot of the Gini Index (pink) and the CLF Criterion (brown) over a grid of values for the shift $c$. Note that the plots are more shaky now in the region around the maxmim due to sampling variability, but otherwise show the same pattern as before. Also

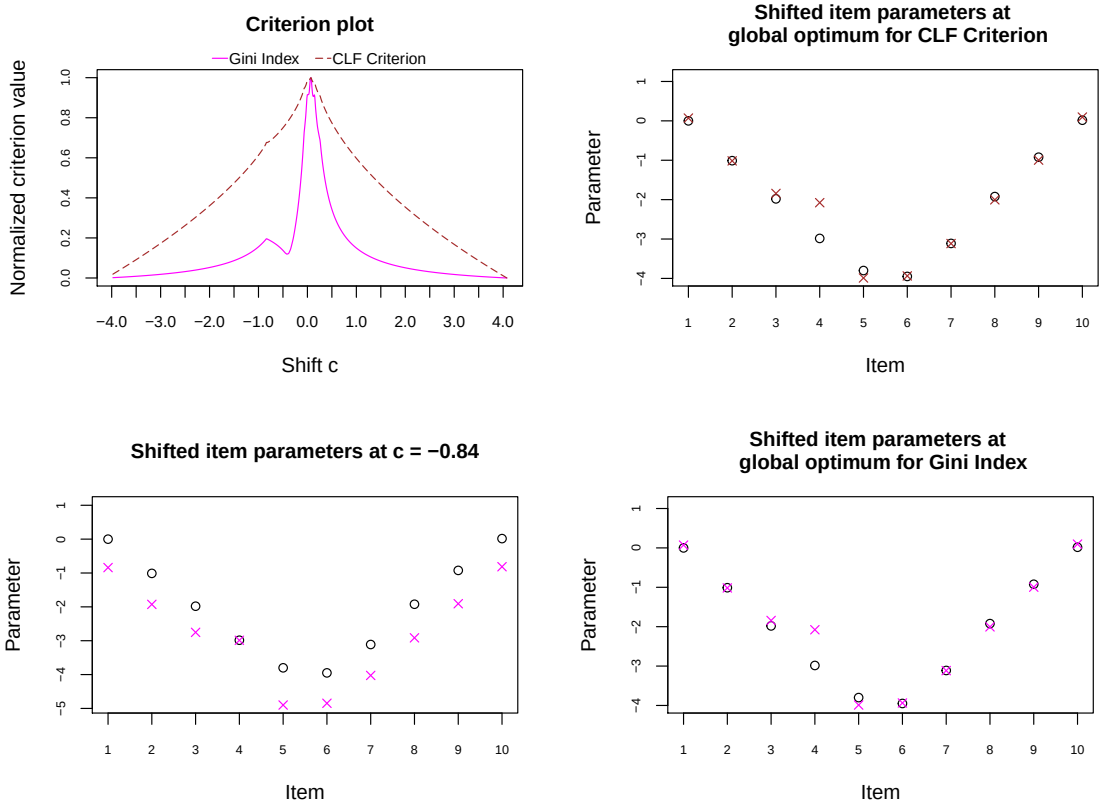


Figure 6: Criterion plot (top left), shifted item parameters according to global optima (right column) and shifted item parameters according to local maximum (bottom left) for toy example with one item displaying DIF of size 0.75, based on estimated item parameters.

the locations of the item parameter estimates corresponding to the global optimum in Figure 6 (right column) are affected by sampling variability now and cannot all be perfectly aligned, but again they show a solution where all items but the fourth item roughly interlock. In addition, there is again a second peak, also slightly shifted in position due to sampling variability of the item parameter estimates, where the fourth item would interlock (Figure 6 bottom left).

With a similar toy example we would also like to illustrate the abovementioned property, that in principle the Gini Index is independend of the absolute amount of wealth – or in our case DIF: When we increase the amount of DIF in item 4 (1.5 vs. 0.75), we can see in Figure 7 that the first peak stays in the same position, but the second peak shifts to the position where the fourth item would interlock, $c = -1.5$, which is logical. What is not visible in this figure, however, is the raw values of the criteria at their optima, because they are normalized for comparability. We therefore display the raw values of the criteria for different DIF patterns in Table 1 and compare the first two rows of Table 1, where still only item 4 has DIF, and the amount of this DIF is doubled in the second row. We see that for the Gini Index it is only relevant that one out of ten items contains the entire amount of DIF. Its value does not vary with the amount of

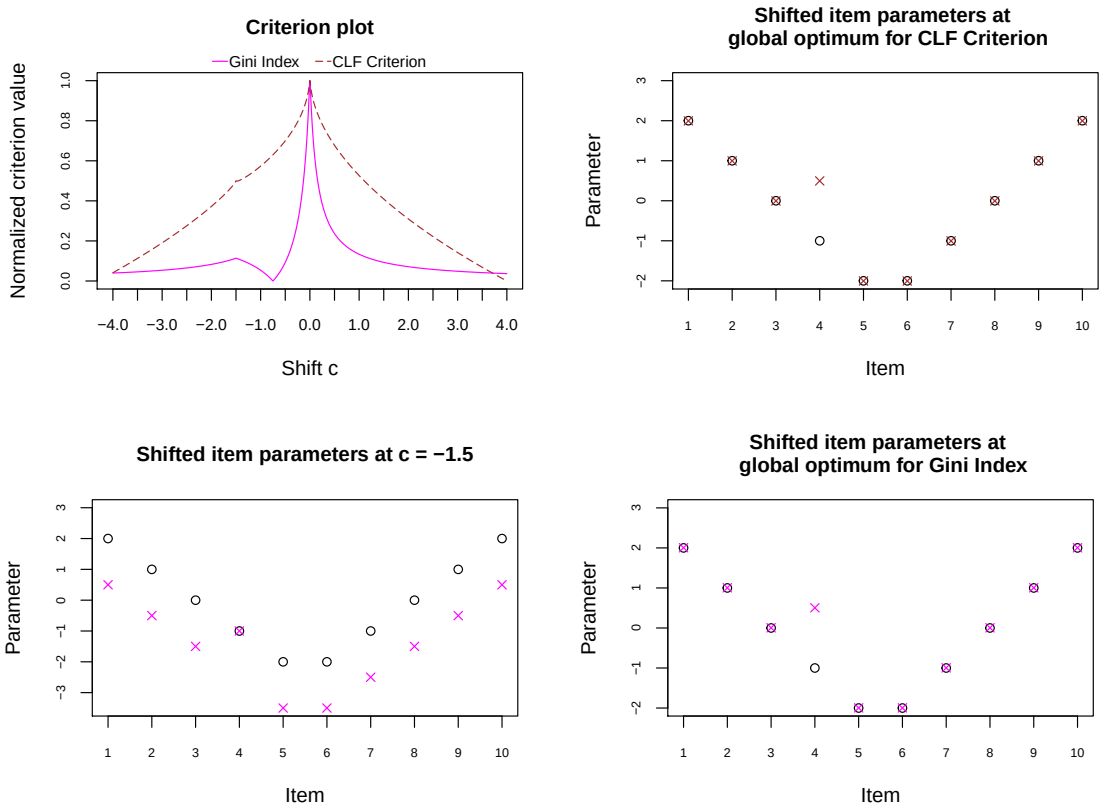


Figure 7: Criterion plot (top left), shifted item parameters according to global optima (right column) and shifted item parameters according to local maximum (bottom left) for toy example with one item displaying DIF of size 1.5, based on true item parameters.

DIF that this item contains. This is different for the CLF Criterion, that does vary based on the amount of DIF. If, however, several items share the same overall amount of DIF, like in the third and fourth row of Table 1 (where the DIF effects add up to 1.5, but are distributed over two or three items respectively), both criteria show lower values when the same overall amount of DIF is distributed between more items.

| DIF pattern | Gini Index | CLF Criterion |
|---|---|---|
| (0, 0, 0, 0.75, 0, 0, 0, 0, 0, 0) | 0.90 | −0.87 |
| (0, 0, 0, 1.5, 0, 0, 0, 0, 0, 0) | 0.90 | −1.22 |
| (0, 0, 0, 0.75, 0.75, 0, 0, 0, 0, 0) | 0.80 | −1.73 |
| (0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0, 0) | 0.70 | −2.12 |

Table 1: Gini Index and CLF Criterion for different DIF patterns.

For completeness we also show what the criterion plots (Figure 8, left) and item parameter locations (Figure 8, right) look like when no DIF is present. In this setting, all items have a

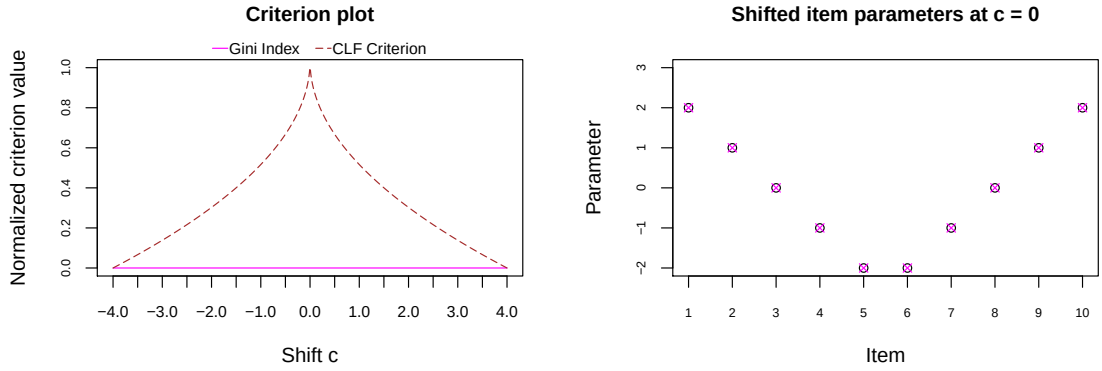


Figure 8: Criterion plot (left) and shifted item parameters according to global optimum (right) for toy example with no item displaying DIF, based on true item parameters.

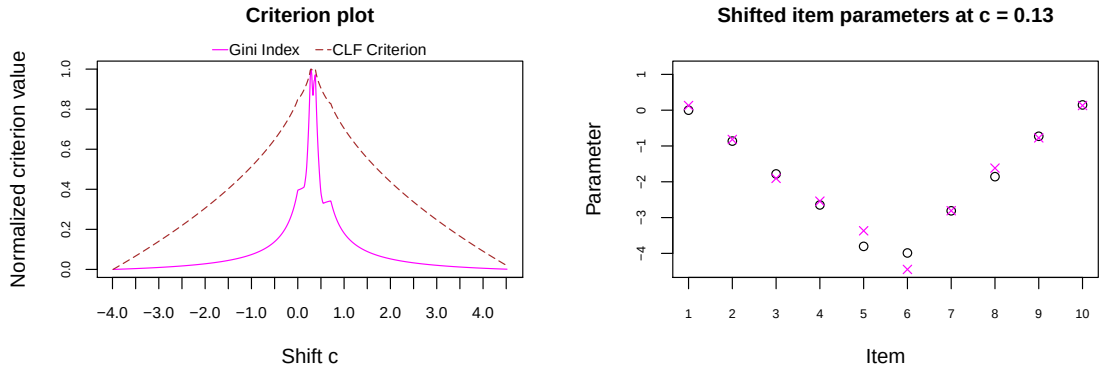


Figure 9: Criterion plot (left) and shifted item parameters according to global optima (right) for toy example with no item displaying DIF, based on estimated item parameters.

difference of zero between the groups. In this case, the Gini Index gives a flat criterion plot[5] and the CLF Criterion gives a single peak at the position where all differences between the item parameters are zero. Note, however, that this exact result is only possible when the true item parameters are used.

When we consider a more realistic situation and simulate item responses from the item parameters with no DIF, the criterion plot for the Gini Index looks quite different at first glance, as displayed in Figure 9 (left). Now we find that the criterion plot for the Gini Index also shows a clear peak because – due to the sampling variability – the item parameter estimates are no longer exactly identical between the two groups at any position. These empricial differences

[5]Remember that originally the Gini Index would be undefined at the position in the center of the criterion plot, where all differences are exactly zero. We have re-defined it to be zero at this position, because it also represents a perfectly equal distribution. The plot for the item parameter locations (Figure 8, right) corresponds to this position in the center of the criterion plot. Since the resulting item parameter locations are exactly the same for Gini Index and CLF Criterion, they are only displayed once here to save space.

are treated by both criteria like items that show a small amount of DIF, so that again the optimal solution for both criteria is one where most items interlock as well as possible between the groups, but some items show small differences in their item parameter estimates – in this case due to the sampling variablity.

# D. Simulation studies

These simulation studies compares the performance of the Gini Index and CLF Criterion to each other and to that of three anchoring methods from the literature.

## D.1. Simulation study I

The simulation design for this study was chosen to be very similar to that of Kopf et al. (2015a) to ensure comparability with this extensive comparison study.

*Simulation design*

We simulated data sets for two groups of subjects, the reference and the focal group, under the Rasch model. In most of the scenarios, a certain percentage of the items was simulated to show DIF between the groups, but there are also scenarios that were simulated completely under the null hypothesis with no DIF in any item. In each setting, 10000 replications were simulated.

**Person and item parameters** The person parameters were generated from a normal distribution with variance 1 and a mean of 0 for the reference group and of -1 for the focal group.

A set of 40 item parameters, that had been previously used by Wang et al. (2012) and Kopf et al. (2015a), were the basis for our study design: $\beta$ =(−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). These item parameters were used for all settings with a test length of 40 items.

In order to be able to manipulate the test length, for settings with test lengths of 20 or 60 items, the respective number of parameters was randomly drawn from the original set of 40 values (in the case of 20 in random order without replacement, in the case of 40 in random order, and in the case of 60 in random order with replacement).

**DIF-items** Depending on the percentage of DIF items, the first test length × DIF percentage of the items were simulated to display DIF by means of setting the difference in the item

parameters between reference and focal group to $+0.6$ or $-0.6$ depending on the intended direction of DIF.[6]

**IRT model** The item responses in each group were generated by means of the Rasch model.

**Manipulated variables** Similar to previous simulation studies such as Woods (2009); Wang et al. (2012) and Kopf et al. (2015a), the manipulated variables were the sample size, the test length, the direction of DIF, the percentage of DIF items and the anchoring methods.

**Sample size** The sample size for the simulated data sets was varied between 250 and 1500 in steps of 250. This overall sample size was divided equally between the two groups. (We also investigated settings with unequal samples sizes. For unequal group sizes the power was slightly diminished for all methods, but the comparisons between the methods were not affected by this factor. Therefore, in the interest of saving space, we present only results for equal group sizes, which also makes the following plots easier to read.)

**Direction of DIF and percentage of DIF items** The direction of DIF is either balanced (where each DIF-item favors either the reference or the focal group, but no systematic advantage for one group remains because the effects cancel out), or unbalanced with an advantage for the focal group. (We have also investigated settings with an advantage for the reference group, but the results are virtually the same and thus not displayed in the interest of saving space.)

The percentage of DIF items relative to the overall test length was set to either 0%, 20%, 40% or 60%.

**Anchoring methods** The following methods were compared in this study:

- the "constant four all other" method suggested by Woods (2009),
- the "constant four MPT" method suggested by Kopf et al. (2015a),
- the "iterative forward" method suggested by Kopf et al. (2015b),
- anchor point selection based on the CLF Criterion employed by Asparouhov & Muthén (2014) and Muthén & Asparouhov (2014), and
- anchor point selection based on the Gini Index suggested in this paper.

**Outcome variables** Like in Kopf et al. (2015b) and Kopf et al. (2015a) the item-wise Wald test based on the conditional maximum likelihood item parameter estimates (cp. Glas & Verhelst 1995; Kopf et al. 2015b) was used for the final DIF tests in our simulation studies. Note, however,

[6] This is equivalent to a random assignment, because the item parameters are drawn in a random order, but simplifies the interpretation of the following figures.

as will become clear below, that our results are of a general nature that straightforwardly generalizes to other DIF tests.

Due to the fact that one restriction is necessary for the item parameter estimation, for a test length of $m$ items only $m-1$ parameters can be estimated and tested freely. Therefore, one item cannot be formally tested for DIF in the final test. For the traditional anchor item selection methods, the item that was first selected into the anchor (and is thus considered the least likely to have DIF by the respective method) is not tested for DIF. For the anchor point selection approach, the item that shows the smallest item parameter difference between the groups in the global optimum of each criterion (again the one considered the least likely to have DIF by each criterion) is not tested for DIF.[7]

In the following plots we will report the average false alarm rate (that corresponds to the type I error) and the average hit rate (that corresponds to the power of the DIF tests) for the final DIF tests. The false alarm rate is computed as the percentage of items that were simulated as DIF free, but erroneously show a significant test result. The hit rate is computed as the percentage of items that were in fact simulated to have DIF and correctly show a significant test result.

*Results*

In the following, we will present the results for a test length of 40 items in all detail. The results for test lengths of 20 and 60 items showed very similar patterns and are thus omitted to save space. (For those results where the test length did have a small but notable effect, namely for the "constant four MPT" method and the "all other", this is mentioned below.)

**Null case: No DIF** First we have checked the false alarm rates in the null case scenario where no DIF items were generated. The false alarm rates should correspond to the nominal type I error rate of 5%. Our results (not displayed to save space) show that all methods hold this nominal type I error rate in the null case scenario, only the "iterative forward" methods shows a false alarm rate slightly above 5% in the scenario with a shorter test length of 20 items. The remaining methods show conservative false alarm rates (around 3%), with the CLF Criterion displaying the lowest (below 1%) and the Gini Index displaying the second lowest (around 2%) false alarm rate.

[7] Note again that we show mathematically in Appendix F that in all local and global optima the item with the smallest item parameter difference actually perfectly interlocks and thus constitutes an anchor item. If DIF tests were to be carried out for other shift positions, e.g., for illustration purposes, our implementation would also treat the item with the least item parameter difference as the one that should not be tested.

**Unbalanced DIF** Now we will look at the results for scenarios with unbalanced DIF favoring one group. Figure 10 (top row) shows the false alarm rates, zoomed false alarm rates and hit rates for all methods in a scenario with a testlength of 40 items and 20% of these items being simulated with DIF. The results show that for this scenario all methods except for the "all other" method hold the nominal type I error rate. For the "all other" method, the false alarm rate notably increases with the sample size. This effect is slightly more pronounced for shorter test lengths (results not shown for brevity) and has already been discussed as a known problem of the "all other" method in unbalanced DIF settings in the introduction. All methods show hit rates that increase with the sample size as expected. The "iterative forward" method shows the highest hitrate, followed by the "constant four MPT" method, the Gini Index, the "all other" method and the CLF Criterion.

When the percentage of DIF items rises to 40% in the next scenario displayed in Figure 10 (middle row), we note that the false alarm rates of some methods increase. Most notably, for the "all other" method, the false alarm rate increases even more strongly with an increasing sample size and goes up as high as 25%. For the other anchoring methods, as well as to a lesser degree for the Gini Index, we see a pattern where the false alarm rates show a slight inversely u-shaped pattern, that is more pronounced for longer test lengths (results not shown for brevity) and was similarly observed and explained by Kopf et al. (2015b). For the "iterative forward" method this also leads to a false alarm rate notably above the nominal 5% level for small and medium sample sizes, so that we should also interpret the hit rate of this method with caution. The hit rates of all methods increase with increasing sample size as expected. Again we find that the "iterative forward" method shows the highest hitrate (but also an increased false alarm rate), followed by the "constant four MPT" method and the Gini Index, and with some distance by the CLF Criterion and the "all other" method (that shows a lower increase in the hit rate for larger sample sizes, where it also showed a higher increase in the false alarm rate).

In addition to the first two DIF scenarios, where a minority of 20% or 40% of the items were simulated with DIF, we now consider a scenario where the majority of items, 60%, are simulated with DIF in favor of the focal group. The results in Figure 10 (bottom row) show that for this scenario, the false alarm rates strongly increase, while the hit rates decrease similarly dramatically. (The only method that is less affected by this is the "all other" method, but we have seen above that this method also has its problems, they only work as an alleged advantage in this particular setting.)

To understand what is going on in this setting, Figure 11 (top left) illustrates the criterion plot for the item parameter values used in the simulations with 60% DIF. We find that the criterion

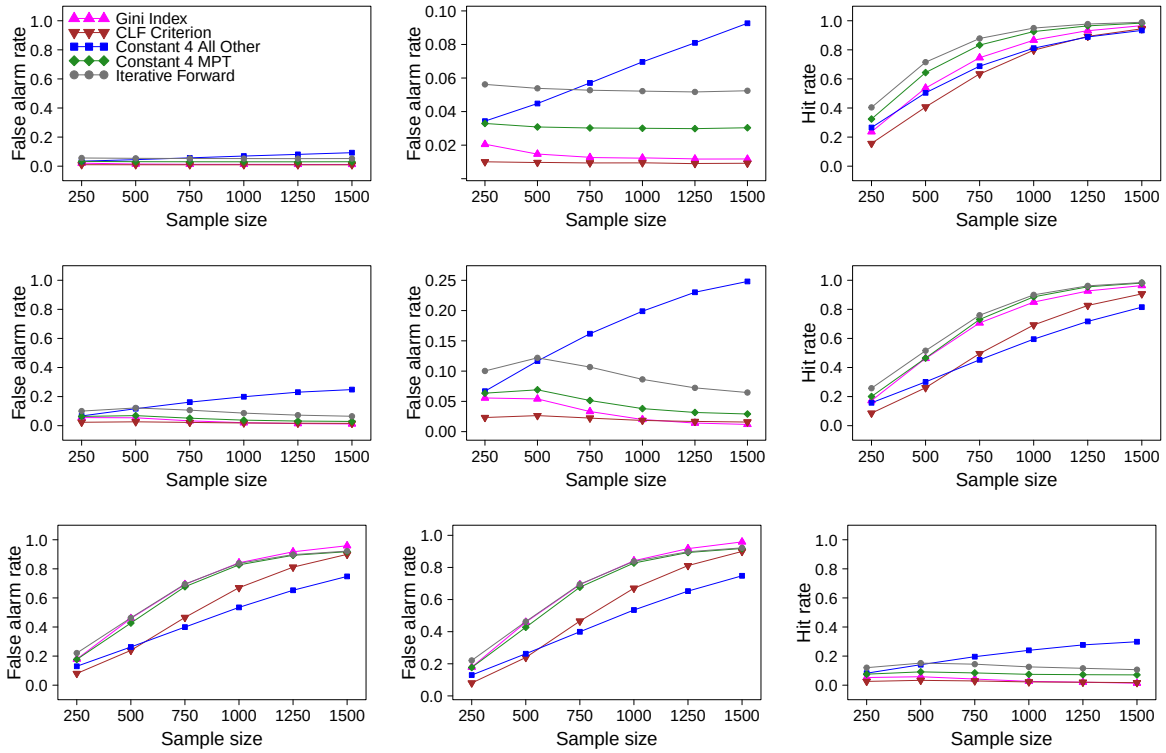


Figure 10: False alarm rates (y-axis from 0 to 1; left column), zoomed false alarm rates (y-axis from 0 to highest value; middle column) and hit rates (y-axis from 0 to 1; right column) for scenario with 20 percent (top row) 40 percent (middle row) and 60 percent (bottom row) DIF items favoring the focal group.

plot has two peaks, corresponding to the displays of the shifted item parameters in Figure 11 (right column and bottom left).[8] At the global optimum (right column), the majority of items interlock. However, this majority of items has – by the definition of the simulation desing – been simulated with DIF of the same amount. This makes the items work together as the bigger cluster and results in the high false alarm rates in Figure 10 (bottom row). The second peak, on the other hand, where the smaller cluster of items interlocks, would be the correct one according to the simulation design. However, as we have discussed above, from a philosophical point of view both solutions can be equally valid. As we have discussed in the introduction, this is a situation where, since mathematically both solutions are equivalent, we have to decide based on considerations about the content of the items, and what the scale is supposed to measure, which solution is to be favored.

Here we also see a strong parallel to the works of Bechger & Maris (2015) and Pohl et al. (2017), who also critically discuss the general assumption that the majority of items is DIF free and

[8]Please note that in the simulations presented here, the item clusters always consist of neighboring items, because this makes them easier to detect visually. As will become clear from the empirical application examples, however, the methods work equally well when clusters are formed by non-neighboring items.

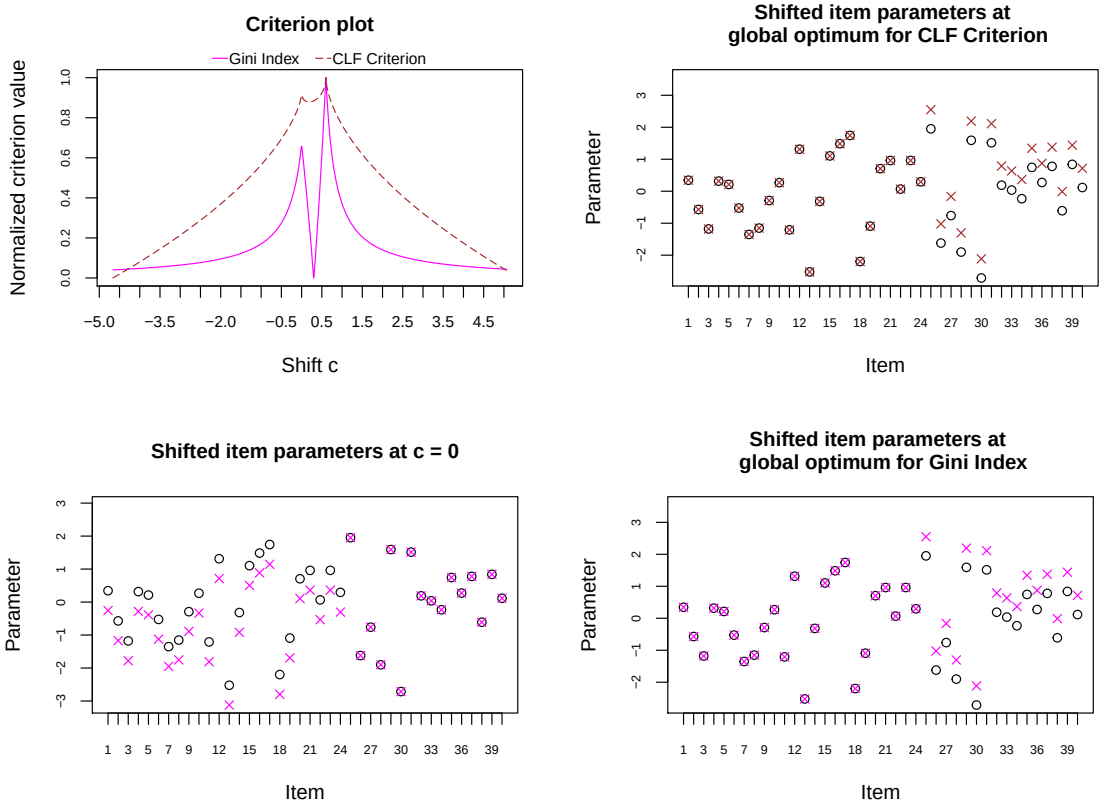


Figure 11: Criterion plot (top left), shifted item parameters according to global optima (right column) and shifted item parameters according to local maximum (bottom left) for simulation setting with 60 percent DIF items favoring the focal group, based on true item parameters.

aim at the detection of invariant item clusters. In our approach, multiple invariant item clusters correspond to multiple local optima in the criterion plot.

Let us first look again at the results where we stick to the definition of the simulation design and label the minority of items as DIF free and the majority of items as having DIF. The results for this view are displayed in Figure 10 (bottom row). We see that all anchoring methods show an increased false alarm rate now, because they tend to label the majority of items to be DIF free, which is counted as wrong under this view. Notably, the "all other" method, that showed an increased false alarm rates in the earlier settings, now has the lowest false alarm rate for larger sample sizes. The reason for this is that this method assumes that DIF was balanced, so that it selects a solution closer to the simulation design. Similarly, the CLF Criterion shows a lower false alarm rate than the other methods for reasons we will shortly illustrate.

Considering the hit rates (Figure 10, bottom row, right panel) we see that now the methods do not improve with increasing sample size (except for the "all other" method, that improves as an artifact of its erroneous assumption of balanced DIF). From the pattern we notice that the hit rates in this setting (with 60% DIF items) follow the same pattern as the false alarm rates in

the earlier setting (with 40% DIF items), because the methods consider the majority of items as DIF free and accordingly mislabel the other items.

Let us further explore why the false alarm rates of the CLF Criterion were apparently also less affected by this extreme scenario: When we look at the criterion plots for the Gini Index and the CLF Criterion (based on the simulated items parameters from Figure 11, top), we find that for the CLF Criterion the two peaks corresponding to the two solutions have more similar criterion values than for the Gini Index. Therefore, in the simulations from this setting with random variation, the CLF Criterion may be more likely to pick the left peak in some simulation runs (an exemplary illustration of one simulation run, where for the CLF Criterion the empirical global optimum corresponds to the left peak, is provided in Figure 12). Since picking the left peak corresponds to labeling the minority of the items as DIF free – which is against our intuition but in line with the scoring rule currently considered – the CLF Criterion is less affected in its false alarm rate and hit rate in Figure 10.

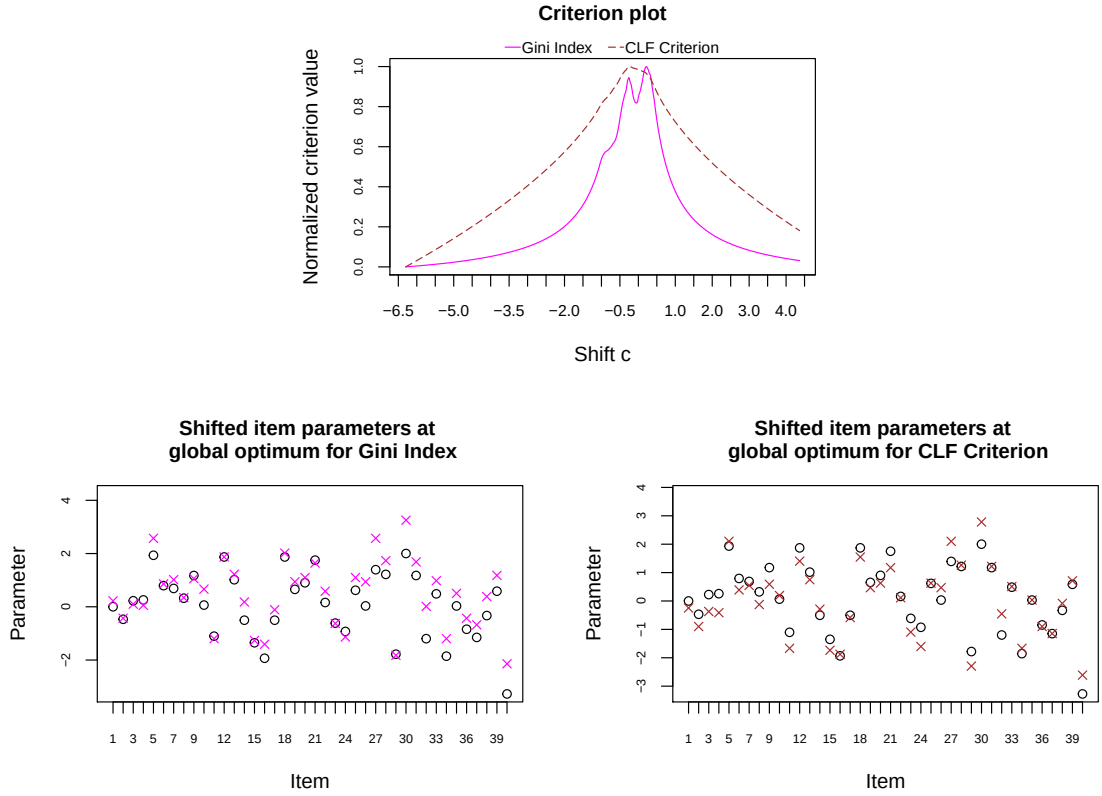


Figure 12: Criterion plot (top) and shifted item parameters according to global optima (bottom) for simulation setting with 60 percent DIF items favoring the focal group, based on estimated item parameters.

We argue throughout this manuscript that this additional information provided by the criterion

plot is extremely valuable and should be taken into account – together with the item content and the question what the scale is supposed to measure – when deciding which items should be labeled as DIF items and how to proceed with this information. In the simulation study, where we need to decide on a scoring rule to be able to compute the aggregated false alarm rates and hit rates, this reasoning cannot be entirely transported, but we can try to mimic it by using a scoring rule that counts either solution as correct. This scoring rule counts both the solution where the cluster of items labeled as DIF items in the simulation design is identified as DIF items, and the solution where exactly the other cluster of items is identified as DIF items as correct, but still counts any item assignment not corresponding to either of the two solutions as wrong. We refer to this scoring rule as "label-switching", because it resembles the fact that in cluster analysis we want to judge whether observations correctly end up in the same cluster in two runs, but the labeling of the clusters is arbitrary.

When this label-switching scoring rule is used for computing the false alarm rates and hit rates for all methods (Figure 13), we see that the results return to what we saw for lower percentages of DIF items, namely that the methods show slightly increased (for the "iterative forward" method) or acceptable false alarm rates and increasing hit rates (except for the "all other" method, that has trouble with the unbalanced setting in general). The Gini Index now shows the highest hit rates, in particular notalby higher than the CLF Criterion. We assume that this is due to the fact that the criterion plot of the Gini Index more clearly distinguishes between the two peaks, which under the label-switching scoring rule both result in low false alarm rates and high hit rates, whereas the CLF Criterion shows similarly high criterion values for the entire peak area in Figure 12 (top), so that it may be more likely in a simulation with random variability to select a shift value in between, that corresponds to neither one of the two correct solutions.

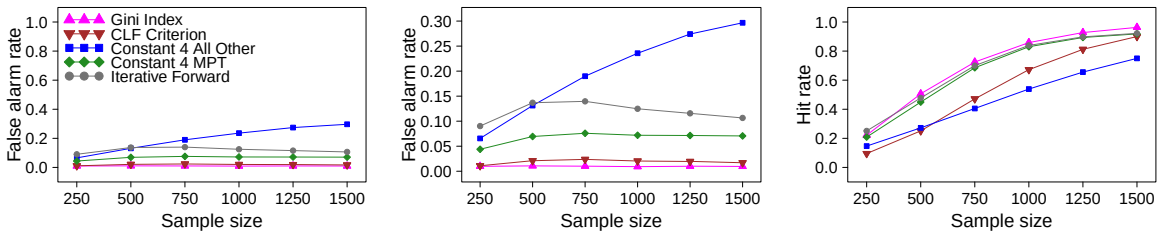


Figure 13: False alarm rates (y-axis from 0 to 1; left), zoomed false alarm rates (y-axis from 0 to highest value; middle) and hit rates (y-axis from 0 to 1; right) for scenario with 60 percent DIF items favoring the focal group and label switching allowed.

**Balanced DIF** Next we will look at the results for scenarios with balanced DIF, where half of the DIF items favor the reference and half of the DIF items favor the focal group.

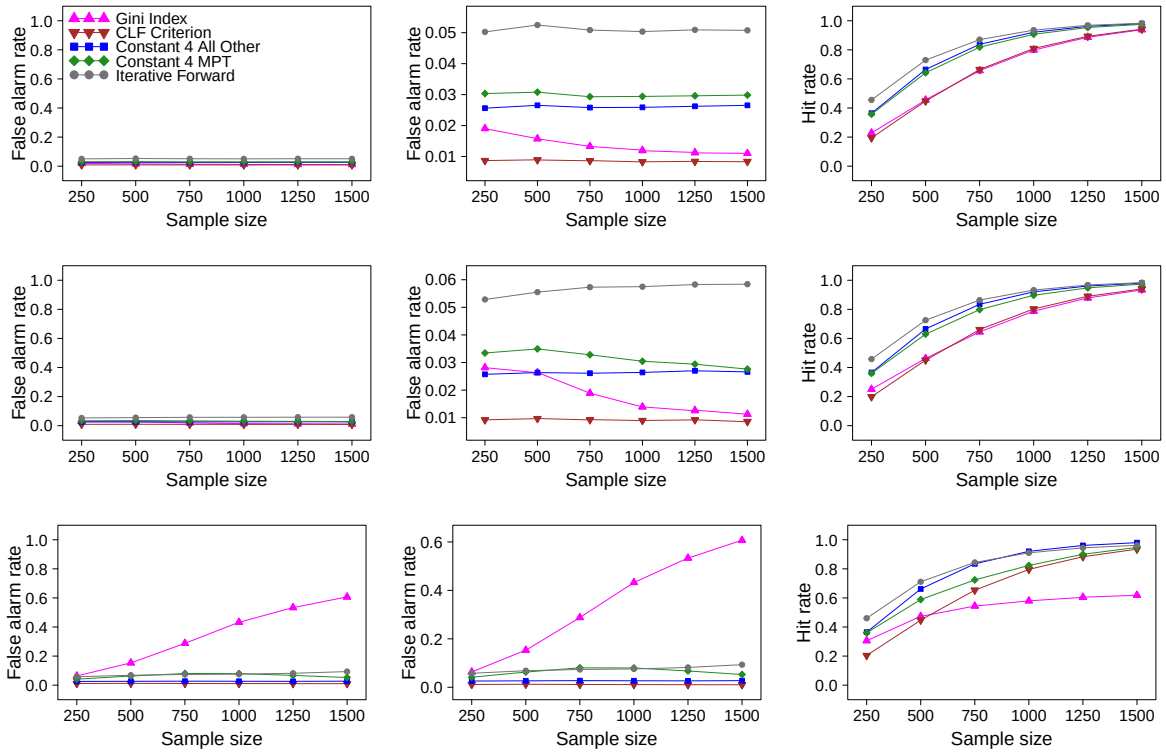


Figure 14: False alarm rates (y-axis from 0 to 1; left column), zoomed false alarm rates (y-axis from 0 to highest value; middle column) and hit rates (y-axis from 0 to 1; right column) for scenario with 20 percent (top row) 40 percent (middle row) and 60 percent (bottom row) DIF items in balanced setting.

Figure 14 (top row) shows the false alarm rates, zoomed false alarm rates and hit rates for all methods in a scenario with a testlength of 40 items and 20% of these items being simulated with balanced DIF. The results show that for this scenario all methods roughly hold the nominal type I error rate and show hit rates that increase with the sample size as expected. The "iterative forward" method again shows the highest hitrate (but note that it also shows a slightly increased false alarm rate), followed by the the "all other" method, the "constant four MPT" method, and, with some distance, the Gini Index and the CLF Criterion. A very similar picture can be found in Figure 14 (middle row) for 40% DIF items.

When the DIF percentage is increased to 60% in Figure 14 (bottom row), however, we notice that the Gini Index shows a strongly inflated false alarm rate and a diminished hit rate, while the other methods, including the CLF Criterion, are not as much affected. These results may at first look surprising, but can also be well explained when we again look at the simulated item parameters and the criterion plot in Figure 15.

What is important to understand this simulation scenario is that in the case of 60% DIF items in the balanced setting, the items actually form three clusters: 30% of the items are simulated with

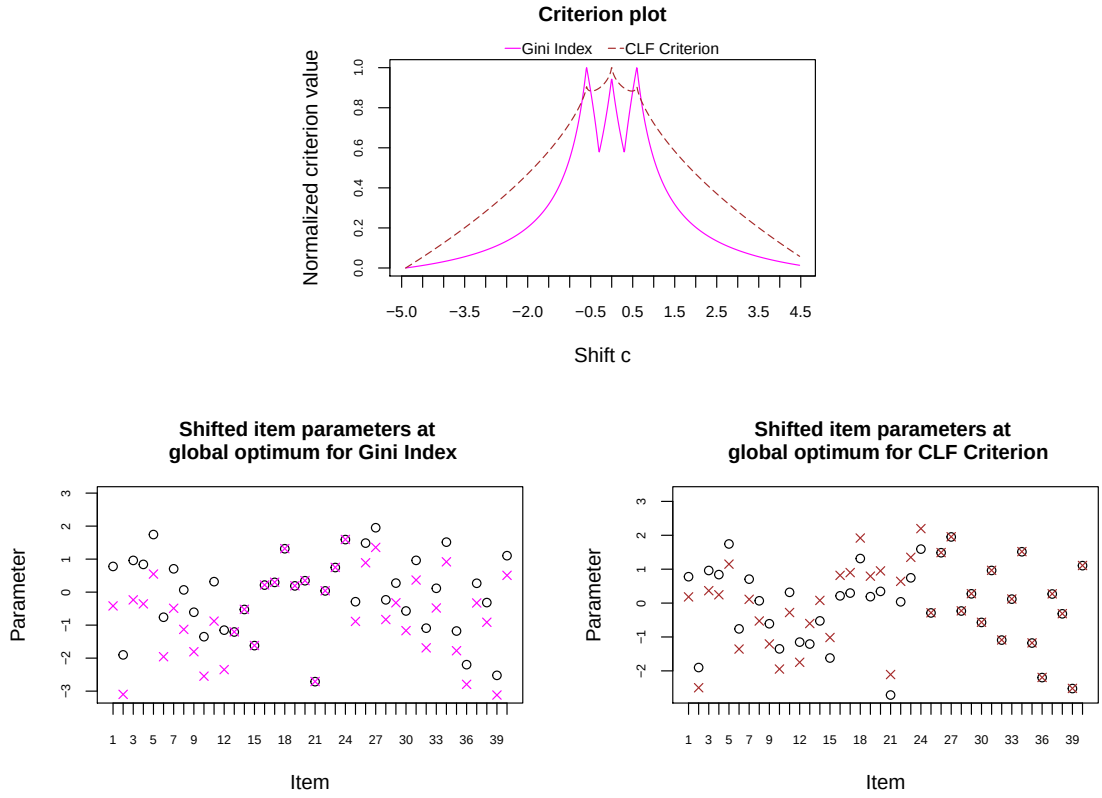


Figure 15: Criterion plot (top) and shifted item parameters according to global optima (bottom) for simulation setting with 60 percent DIF items in a balanced setting, based on true item parameters.

DIF favoring the reference group, 30% with DIF of the same size favoring the focal group, and 40% of the items are simulated without DIF. One solution for shifting the item parameters would recover this pattern, with the largest cluster of 40% of the items interlocking. This solution corresponds to the central peak in the criterion plot in Figure 15 (top). We see that the CLF Criterion has the highest peak at this solution. The item parameter locations corresponding to this solution are displayed in Figure 15 (bottom right), where we see that the the items in the largest cluster (on the right hand side of the plot) interlock. Due to the high central peak of the CLF criterion plot, this solution will be found in most simulation runs. Therefore, this solution leads to non-increased false alarm rates under the original scoring rule for the CLF Criterion.

The Gini criterion, on the other hand, also shows two clearly distinguishable peaks for the other two solutions, where either one of the clusters containing 30% of the items interlock, and a slightly lower central peak. Over the simulation runs with random variability, it will jump back and forth between all three solutions, in many cases selecting not the central peak but one of the other two solutions, one of which is displayed in Figure 15 (bottom left). However, in the original scoring rule underlying Figure 14 (bottom row), either of these solutions is counted as

wrong. This causes the strongly increased false alarm rate of the Gini Index in this setting. It even increases with sample size because on average the peaks become more distinguishable as sample size increases and item parameter estimates are less variable.

If, however, we again adapt our view and count either of the three possible solutions as correct by means of using the label-switching scoring rule for all methods, we see in Figure 16 that the false alarm rate of the Gini Index returns to a very low value (while the "iterative forward" and "constant four MPT" methods show slightly increased false alarm rates), and the hit rate of the Gini Index is now close or equal to that of the CLF Criterion.

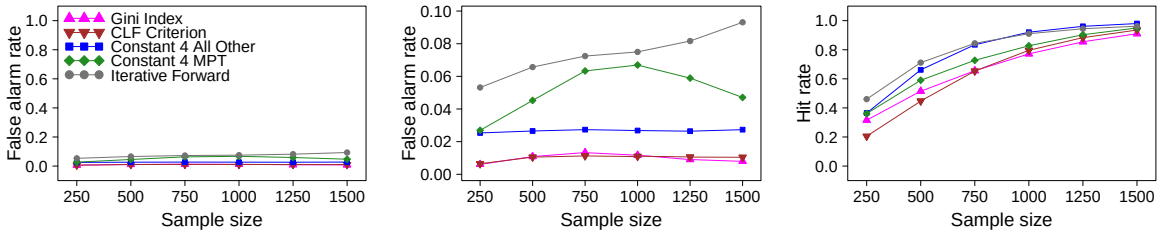


Figure 16: False alarm rates (y-axis from 0 to 1; left), zoomed false alarm rates (y-axis from 0 to highest value; middle) and hit rates (y-axis from 0 to 1; right) for scenario with 20 percent DIF items in balanced setting with label switching allowed.

To shed further light on the different behaviors of the CLF Criterion and Gini Index for this setting, we will increase the DIF percentage even further. This will highlight how both crireria treat different combinations of item cluster sizes and item parameter differences.

When we produce an even more extreme setting with 70% of the items being simulated with balanced DIF, this results again in three item clusters, now with 35% of the items favoring the reference group, 30% neutral items, and 35% of the items favoring the focal group. In this scenario (Figure 17, left; only the criterion plot is presented to save space) the item cluster corresponding to the central solution is the smallest cluster, which makes the other two solutions even more plausible alternative solutions. The Gini Index, that already showed higher peaks for these two solution before, now shows this pattern even more pronounced. Interestingly, however, the CLF Criterion still produces the highest peak for the central solution, which labels the smallest item cluster as DIF free at the cost of labelling the two other item clusters as having DIF. This was an advantage in the simulation study with the original scoring rule, but does not well reflect our intuition.

Even if the three groups were of exactly equal sizes (results not shown for brevity) we would find a similar pattern for the criterion plot as in Figure 17 (left), with the CLF Criterion preferring the central peak and the Gini Index preferring the other two solutions.

Figure 17 (right) further illustrates this finding: On the x-axis it displays the percentage of items in the two non-central clusters taken together. On the y-axis it displays the relative height difference between the central peak and the non-central peaks for each criterion (as a percentage of the height difference for the minimum possible amount of balanced DIF for each criterion, to allow comparability). Values above zero (i.e., above the black horizontal line) indicate that the criterion prefers the central solution for a given percentage of items, while values below zero (i.e., below the black horizontal line) indicate that the criterion prefers the non-central solutions. While the Gini Index prefers the non-central solutions already at a percentage close to 60% (corresponding to cluster sizes of 30%, 40% and 30%), the CLF criterion still prefers the central solution when the percentage is close to 80% (corresponding to cluster sizes of 40%, 20% and 40%).

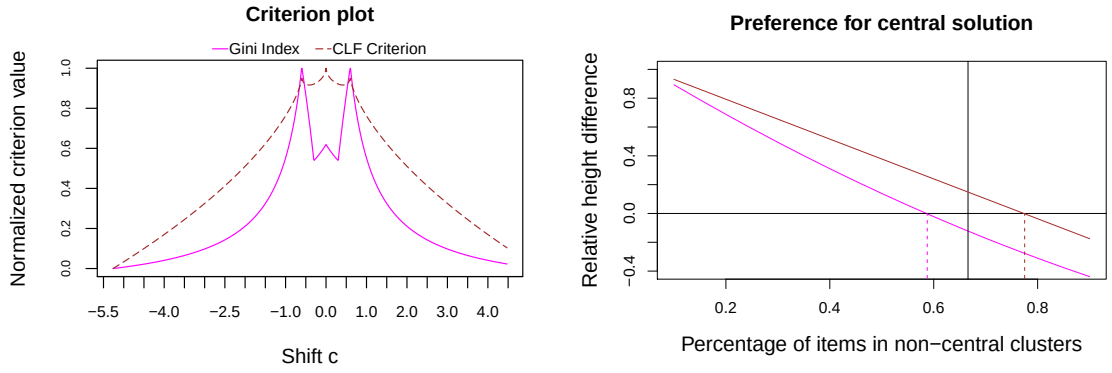


Figure 17: Criterion plot for simulation setting with 70 percent DIF items in a balanced setting, based on true item parameters (left) and illustration of preference for central solution as a function of the item cluster sizes (right).

Intuitively, one could argue that a neutral criterion without any additional assumptions should treat all three solutions as equal when the cluster sizes and DIF amounts are the same. This rationale is represented in Figure 17 (right) by the vertical black line at 66% (corresponding to equal cluster sizes of 33% each). This illustration shows that neither of the two criteria investigated here corresponds exactly to this rationale, but that the behavior of the Gini Index is somewhat closer to it. Over a certain range of percentages, the Gini Index aready prefers the non-central solutions, where the DIF of the non-interlocking items goes in the same direction, while the CLF Criterion keeps preferring the central solution, where the DIF in the non-interlocking items cancels out. With respect to the optimal solutions it generates, this behavior of the CLF Crierion resembles the assumption of balanced DIF in traditional anchoring methods. For practical decisions derived from either crierion, these findings again support

our notion that the globally optimal solution does not tell the whole story, and that solutions corresponding to local optima in the criterion plot should also be explored.

## D.2. Simulation study II

In order to further illustrate the connection between DIF and multidimensionality, we have conducted a second simulation study, that employs a multidimensional IRT model for data generation.

*Simulation design*

The design of this study resembles the first part of Simulation Study I for unbalanced DIF. While there unidirectional DIF was generated by adding a fixed amount of DIF to certain item parameters, now the DIF is induced by letting certain items measure a secondary dimension in addition to the primary dimension (like described, e.g., in Roussos & Stout 1996).

**Person and item parameters** The person parameters were generated from a bivariate normal distribution with means of 0.5 and 0.5 in the reference group and - 0.5 and -0.5 in the focal group, a variance of 1 for each dimension in each group and a covariance of 0.5 between the two dimensions in each group.

A set of 40 intercept parameters was randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 0.5. All items measured the first dimension with a fixed discrimination of 1.

**DIF-items** Depending on the percentage of DIF items, the first test length $\times$ DIF percentage of the items also measured the second dimension with a fixed discrimination of 1. Given that the person parameter distributions (in this simple case only the means) on the secondary dimension differed between the groups, this induces (uniform) DIF in these items (for further details see Roussos & Stout 1996; Ackerman 1992).

**IRT model** The item responses in each group were generated by means of a compensatory dichotomous multidimensional IRT model with the abovementioned specifications (for details on the generating model see Reckase 2009; Chalmers 2012), which corresponds to the multidimensional counterpart of a Rasch model.

**Manipulated variables** Similar to Simulation Study I and previous studies, the manipulated variables were the sample size and the percentage of DIF items, with the same levels as in Simulation Study I. In each setting, again 10000 replications were simulated.

*Results*

As expected, we find the results in Figures 18 and 19 to be very similar to those in Figures 10 and 13 for Simulation Study I, with slightly higher false alarm and hit rates in some places but the same general results, that have already been described in detail above.

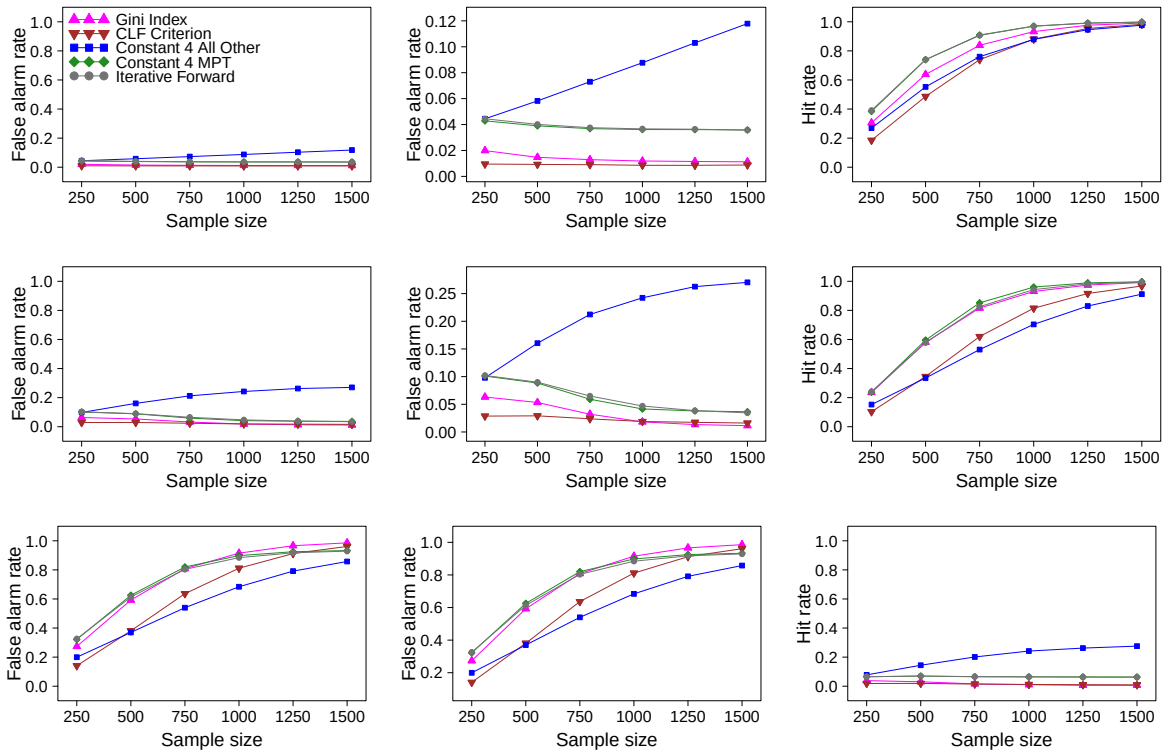


Figure 18: False alarm rates (y-axis from 0 to 1; left column), zoomed false alarm rates (y-axis from 0 to highest value; middle column) and hit rates (y-axis from 0 to 1; right column) for scenario with 20 percent (top row) 40 percent (middle row) and 60 percent (bottom row) DIF items induced by a secondary dimension.

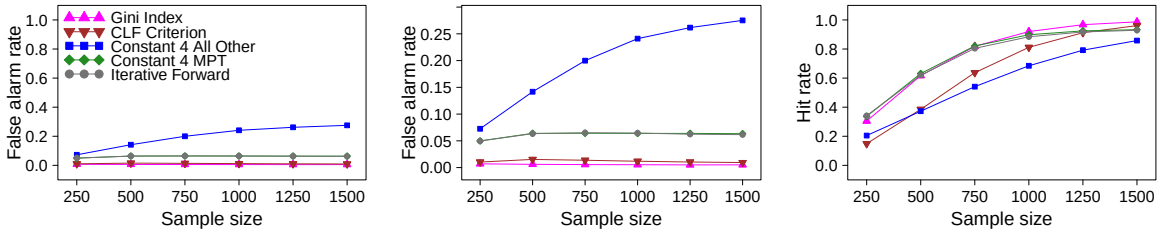


Figure 19: False alarm rates (y-axis from 0 to 1; left), zoomed false alarm rates (y-axis from 0 to highest value; middle) and hit rates (y-axis from 0 to 1; right) for scenario with 60 percent DIF induced by a secondary dimension and label switching allowed.

In particular, we find that when 60% of the items measure the secondary dimension, the original scoring rule again results in very low hit rates (Figure 18, bottom). With the label-switching

scoring rule (Figure 19), however, we can see that the methods are again able to identify the pattern in the items with increasing sample size, and that the Gini Index, together with the "iterative forward" and "constant four MPT" methods, again performs particularly well in this setting.

Plotting a criterion plot (not shown to save space) for this setting would also show two peaks here, similar to the one displayed in Figure 12, because again by design the items form two clusters.

## E. Empirical application examples

For further illustration the anchor point selection method will be applied to two empirical data sets from an online quiz for testing one's general knowledge and from a personality item pool.

### E.1. Application example I: General knowledge quiz

An online quiz for testing one's general knowledge was conducted by the German weekly news magazine DER SPIEGEL in 2009. Overall, about 700,000 respondents participated in this general knowledge quiz and also answered a set of sociodemographic questions. The quiz consisted of a total of 45 items from five different domains: politics, history, economy, culture, and natural sciences. For each domain, four different sets of nine items were available, that were randomly assigned to the participants. A thorough discussion and analysis of the original data set is provided in Trepte & Verbeet (2010).

Here we consider an exemplary sample of university students enrolled in the federal state of Bavaria, who had been assigned questionnaire number 20. This sample contains 1075 cases (417 male and 658 female) and is freely available in the `psychotree` R package (Zeileis, Strobl, Wickelmaier, Komboz & Kopf 2018), where also the wording of all 45 items contained in this quiz is documented.

The result of the anchor point selection for the general knowledge quiz data is depicted in Figure 20. From the criterion plot (Figure 20, left), we can see that there is a clear global maximum for both the Gini Index and the CLF Criterion, and only a slight second bump.[9] When we display the shifted item parameters at the global optimum (Figure 20, right), we see that some item parameters interlock visibly between the groups, while others show smaller or larger distances. To judge which of these distances correspond to significant DIF, the symbols

[9]The slight second bump is caused by a small group of items that in the globally optimal solution show a lower difficulty for female participants to a similar degree, and thus also form a small cluster.

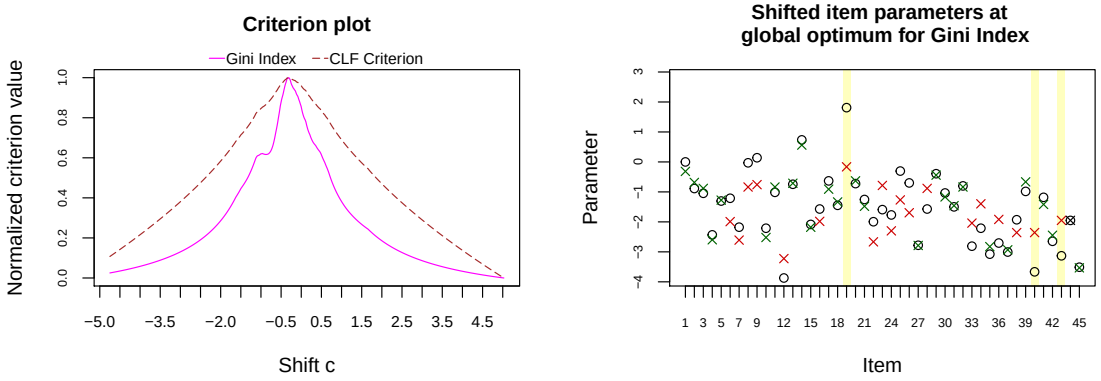


Figure 20: Criterion plot (left) and shifted item parameters according to global optimum (right), with items displaying statistically significant DIF indicated in red and items not displaying statistically significant DIF indicated in green, based on estimated item parameters for the general knowledge quiz data. The item parameters symbolized by circles belong to the female group, those symbolized by crosses to the male group. Items indicated by yellow highlighting are referred to in the text.

for the item parameters in the male group are colored according to the results of the Wald test, with items marked in red displaying significant DIF and items marked in green not displaying significant DIF.

We did not account for multiple testing here, so that the results should not be overinterpreted. However, when we look at those items that exhibit the largest amount of DIF (indicated in Figure 20, right, by yellow highlighting) we find that item 19 ("Who is this? - Picture of Dieter Zetsche, CEO of Mercedes-Benz.") shows a higher difficulty for female participants, whereas items 40 ("What is also termed Trisomy 21? - Down syndrome.") and 43 ("Which kind of bird is this? - Blackbird.") show a lower difficulty for female participants. It is plausible that these items exhibit DIF with respect to the variable gender, for example because they are of differently high interest for male and female participants.

## E.2. Application example II: International personality item pool

In this second example we would like to highlight again the notion of DIF induced by an unaccounted secondary dimension, on which the groups differ. For this example, we use data from 2800 subjects (1881 female and 919 male), taken as a subsample from the International Personality Item Pool (ipip.ori.org) data. This data set is freely available in the `psych` R package (Revelle 2018), where also the wording of all items is documented.

For didactic reasons we include only the first three factors, Agreeableness, Conscientiousness

and Extraversion. Each factor is measured by five items, some of which are inversely phrased.[10] The item responses were originally encoded using a six point scale ranging from *very inaccurate* to *very accurate*. For this exemplary analysis with the binary Rasch model, the responses have been recoded to a binary format, with the three lower categories being recoded as 0 and the three upper categories being recoded as 1.

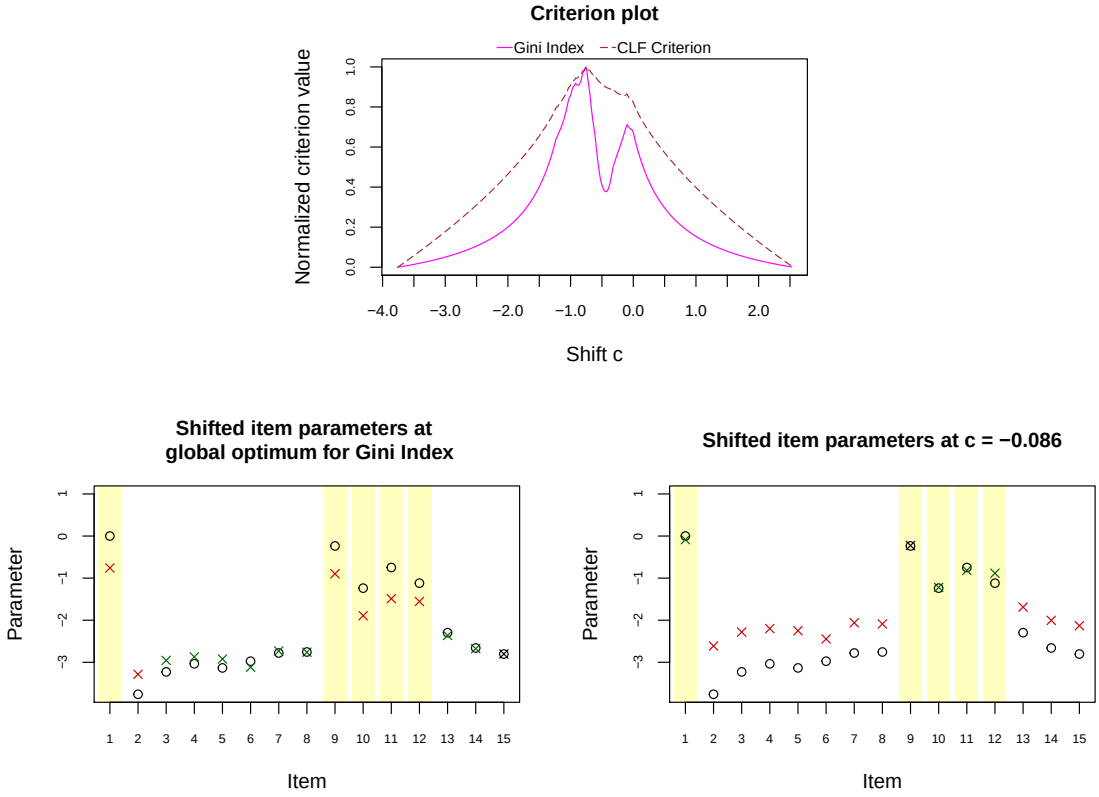


Figure 21: Criterion plot (top), shifted item parameters according to global (bottom left) and local optimum (bottom right), with items displaying statistically significant DIF indicated in red and items not displaying statistically significant DIF indicated in green, based on estimated item parameters for the first three factors of the international personality item pool. The item parameters symbolized by circles belong to the female group, those symbolized by crosses to the male group. Items indicated by yellow highlighting are referred to in the text.

From the criterion plot in Figure 21 (top), we can see that there is again one dominant global maximum for both the Gini Index and the CLF Criterion, but – in contrast to the first application example – now we also see a clearly pronounced second peak to the right of the global optimum, again more notably in the criterion plot for the Gini Index. The shifted item pa-

[10]For the remaining factor Neuroticism there are no inversely phrased items. For Openness the two inversely phrased items do not function in the same way that we will see for the first three factors, possibly because their content does not induce gender specific social desirability.

rameters at both these locations are shown in Figure 21 (bottom, left for global, right for local optimum). As in the previous section, items marked in green do not show significant DIF, while items marked in red do show significant DIF according to the Wald test.

Given that the first 15 self report items are supposed to cover three different factors, that we inappropriately placed on one joint scale, we might have expected to encounter three clusters of items based on these dimensions. That could have been the case if the two groups, males and females, would show systematically different distributions on these dimensions (cf. Roussos & Stout 1996).

What we do find, however, is a less obvious pattern: Those items that show notable DIF in the same direction (harder to agree to for female respondents) in the globally optimal solution (Figure 21, bottom left) and interlock or are very close in the locally optimal solution (Figure 21, bottom right), are the inversely phrased items: items 1 ("A1: Am indifferent to the feelings of others"), 9 ("C4: Do things in a half-way manner."), 10 ("C5: Waste my time."), 11 ("E1: Don't talk a lot.") and 12 ("E2: Find it difficult to approach others."). These items are indicated in Figure 21 (bottom) by yellow highlighting.[11]

In this example, the inversely phrased items form a cluster corresponding to the second peak in the criterion plot, that is clearly distinguishable for the Gini Index. An interpretation of this cluster from the multidimensionality perspective of DIF is that these inversely phrased items form a method factor (c., e.g., Weijters, Baumgartner & Schillewaert 2013) on which female and male respondents differ, possibly due to gender specific social desirability.

## F. Mathematical derivation of possible locations of optima

Let $a_j = \tilde{\beta}_j^{(g_1)} - \tilde{\beta}_j^{(g_2)}$ be the distance between the initial item parameter estimates in the two groups for item $j$, with $j = 1, \ldots, m$. The item-wise absolute distance on the shifted scale $d_j(c)$ used in the main text then corresponds to $d_j(c) = |a_j + c|$. For brevity, we will denote $d_j(c)$ as $d_j$.

Without loss of generality, we assume that the $a_j$ are sorted in non-decreasing order and $a_m = 0$. We will in the following use $s_j$ for the sign of $a_j + c$ and $r_j$ for the rank of $d_j$ in non-decreasing order. Note that $s_j$ is the first derivative of $d_j$ wrt. $c$.

[11] Note that item 2 ("A2: Inquire about others' well-being.") also shows notable DIF in the globally optimal solution (Figure 21, bottom left), but this items is not part of the same cluster as the inversely phrased items because for item 2 the DIF goes in the other direction (easier to agree to for female respondents).

## F.1. Locations of optima for the Gini Index

In our simplified notation, the Gini Index can be written as as

$$\mathrm{GI}(c) = \frac{2 \cdot \sum_{j=1}^{m} r_j \cdot d_j}{m \cdot \sum_{j=1}^{m} d_j} - \frac{m+1}{m}.$$

This formula can be separated in a trivial part and an interesting part by writing $\mathrm{GI}(c) = \frac{2}{m} f(c) - \frac{m+1}{m}$ with $f(c)$ defined as

$$f(c) = \frac{\sum_{j=1}^{m} r_j d_j}{\sum_{j=1}^{m} d_j}$$

The numerator of the first derivative of $f$ wrt. $c$ is given by

$$num(f'(c)) = \left(\sum_{j=1}^{m} r_j s_j\right)\left(\sum_{k=1}^{m} d_k\right) - \left(\sum_{j=1}^{m} s_j\right)\left(\sum_{k=1}^{m} r_k d_k\right) \tag{1}$$

Resorting the summands by $s_j d_k$, this can be written as

$$num(f'(c)) = \sum_{j,k=1}^{m} (r_j - r_k) s_j d_k = \sum_{j,k=1}^{m} (r_j - r_k) s_j s_k (a_k + c).$$

Observe that this value is non-continuous at every position where either any sign $s_j$ switches, or where the ranks of two neighboring values change, i.e., if $d_j = d_k$. Note that since all values start negative, the second condition can be written as $a_j + c = -c - a_k$ (regardless of the order of $j$ and $k$). So all positions $c$ where $f'$ is not continuous are all points

$$c = -a_j \tag{2}$$

or

$$c = -\frac{a_j + a_k}{2}$$

for any $a_j$ and $a_k$; these are $\frac{m(m+1)}{2}$ positions.

Between any two such positions, $s_j$ and $r_j$ are constant, $f$ is linear in the denominator, and the denominator is always positive.

An important and non-trivial observation is that the numerator is constant with respect to $c$, since

$$\begin{aligned} num(f'(c)) &= \sum_{j,k=1}^{m} (r_j - r_k) s_j s_k (a_k + c) \\ &= \left(\sum_{j,k=1}^{m} a_k (r_j - r_k) s_j s_k\right) + c \left(\sum_{j,k=1}^{m} (r_j - r_k) s_j s_k\right) \\ &= \left(\sum_{j,k=1}^{m} a_k (r_j - r_k) s_j s_k\right) \end{aligned}$$

So all extrema of $f$ are at the points where $f'$ is not continuous. Note that the ranges $c < 0$ and $c > -a_1$ asymptotically approach zero and contain no extrema other than the minima at $\pm\infty$.

Further note that at any point where the rank switches, two neighboring ranks $r_k$ and $r_l$ are exchanged; assume $k < l$ wlog. At this position, $s_k$ is still negative while $s_l$ is positive. The second term of Equation 1 is continuous here as $d_k = d_l$, in the first term the sum of all $d$'s is positive. As the rank $r_l$ is increased by one while $s_l$ is positive, and vice versa for $k$, $f'(c)$ steps to a higher value, potentially indicating a minimum, but no maximum.

So, the Gini index can only take a maximum at the points where $c = -a_j$ (Equation 2) and thus $d_j$ is zero for any $j$. Since these are only $m$ positions, we can easily find the maximal Gini index by testing all $c = a_j$ values and choosing the largest among them, in $O(m)$ steps.

### F.2. Locations of optima for the CLF Criterion

In our simplified notation and for the two-group case with only one item parameter, i.e., for the Rasch model, the CLF Criterion[12] reduces to

$$\mathrm{CLF}(c) = \sum_{j=1}^{m} \sqrt{d_j}.$$

To compare with the Gini index, we consider the extrema of this index, too. The first and second derivatives are given by

$$\begin{aligned} \mathrm{CLF}'(c) &= \frac{1}{2} \sum_{j=1}^{m} s_j \frac{1}{\sqrt{d_j}} \\ \mathrm{CLF}''(c) &= -\frac{1}{4} \sum_{j=1}^{m} d_j^{-\frac{3}{2}} \end{aligned}$$

We observe that the derivatives are not defined at any point where $d_j = 0$. The second derivative is negative everywhere in between these points. So in these areas, we have no minima, and exactly one maximum. At the points where $d_j = 0$, the first derivative reaches a pole with a sign switch from negativ to positive, reflected by a local minimum (with a cusp) at the index itself. Consequently, all minima are at $c = a_j$, and the global minimum can be found most efficient by picking the smallest $\mathrm{CLF}(c)$ among those.

This means that for both criteria, we need to search only through the positions $a_1, \ldots, a_m$ to find all optima.

[12] For mathematical simplicity, we have not changed the sign of the CLF in this appendix. Therefore, the interesting optima for the CLF Criterion in the appendix are minima, while those for the Gini Index are maxima.

## References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

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

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, *72*(2), 141.

Bechger, T. M. & Maris, G. (2015). A statistical test for Differential Item Pair Functioning. *Psychometrika*, *80*(2), 317–340.

Central Intelligence Agency (2017). The world factbook: Distribution of family income – Gini Index.

Chalmers, R. P. (2012). `mirt`: A multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://CRAN.R-project.org/package=mirt.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the Likelihood Ratio test for detection of Differential Item Functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Debelak, R. & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement*, *79*(2), 385–398.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961–972.

Egberink, I. J. L., Meijer, R. R., & Tendeiro, J. N. (2015). Investigating measurement invariance in computer-based personality testing: The impact of using anchor items on effect size indices. *Educational and Psychological Measurement*, *75*(1), 126–145.

Eggen, T. & Verhelst, N. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, *71*(2), 303–322.

Fischer, G. & Molenaar, I. (Eds.). (1995). *Rasch Models: Foundations, Recent Developments and Applications*. New York: Springer-Verlag.

Gini, C. (1912, reprinted 1955). Variabilità e mutabilità. In E. Pizetti & T. Salvemini (Eds.), *Memorie Di Metodologica Statistica*. Rome: Libreria Eredi Virgilio Veschi.

Glas, C. & Jehangir, K. (2014). Chapter 5: Modeling country-specic Differential Item Functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment*. New York: Chapman and Hall/CRC.

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. New York: Springer-Verlag.

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103.

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

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.

Muthén, B. & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 978.

Oliveri, M. & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21.

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

Pohl, S., Stets, E., & Carstensen, C. (2017). Cluster-based anchor item identification and selection. Technical Report 68, Leibniz Institute for Educational Trajectories, National Educational Panel Study, Bamberg.

Pokropek, A., Lüdtke, O., & Robitzsch, A. (2020). An extension of the invariance alignment method for scale linking. *Psychological Test and Assessment Modeling*, *62*(2), 305–334.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, London: The University of Chicago Press.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. https://CRAN.R-project.org/package=psych.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Schulze, D. & Pohl, S. (2020). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling: A Multidisciplinary Journal*. (In press).

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*(3), 184–199.

Strobl, C., Kopf, J., Hartmann, R., & Zeileis, A. (2018). Anchor point selection: An approach for anchoring without anchor items. Working Paper 2018-03, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.

Teresi, J. A. & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement equivalence of the patient reported outcomes measurement information system (PROMIS) short forms". *Psychological Test and Assessment Modeling*, *58*(1), 37–78.

Trepte, S. & Verbeet, M. (Eds.). (2010). *Allgemeinbildung in Deutschland – Erkenntnisse Aus Dem SPIEGEL Studentenpisa-Test*. Wiesbaden: VS Verlag.

Van der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*(2), 131–145.

von Davier, M. & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology*, *3*(3), 115–124.

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*(4), 687–708.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*.

Westers, P. & Kelderman, H. (1992). Examining Differential Item Functioning due to item difficulty and alternative attractiveness. *Psychometrika*, *57*(1), 107–118.

Woods, C. M. (2009). Empirical selection of anchors for tests of Differential Item Functioning. *Applied Psychological Measurement*, *33*(1), 42–57.

Wright, B. D. & Stone, M. (1999). *Measurement Essentials*. Wilmington: Wide Range Inc.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Chapter 17: Scaling PIAAC cognitive data. In W. T. Irwin Kirsch (Ed.), *Technical Report of the Survey of Adult Skills (PIAAC)*. OECD.

Zeileis, A. (2014). *ineq: Measuring Inequality, Concentration, and Poverty*. https://CRAN.R-project.org/package=ineq.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2018). *psychotree: Recursive Partitioning Based on Psychometric Models*. https://CRAN.R-project.org/package=psychotree.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2020). *psychotools: Infrastructure for Psychometric Modeling*. https://CRAN.R-project.org/package=psychotools.

**Affiliation:**

Carolin Strobl

Department of Psychology

Universität Zürich

Binzmühlestrasse 14, Box 27

CH-8050 Zürich, Switzerland

E-mail: carolin.strobl@uzh.ch

**University of Innsbruck - Working Papers in Economics and Statistics**
**Recent Papers** can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2018-03 **Carolin Strobl, Julia Kopf, Raphael Hartmann, Achim Zeileis:** Anchor point selection: An approach for anchoring without anchor items

2018-02 **Michael Greinecker, Christopher Kah:** Pairwise stable matching in large economies

2018-01 **Max Breitenlechner, Johann Scharler:** How does monetary policy influence bank lending? Evidence from the market for banks' wholesale funding

2017-27 **Kenneth Harttgen, Stefan Lang, Johannes Seiler:** Selective mortality and undernutrition in low- and middle-income countries

2017-26 **Jun Honda, Roman Inderst:** Nonlinear incentives and advisor bias

2017-25 **Thorsten Simon, Peter Fabsic, Georg J. Mayr, Nikolaus Umlauf, Achim Zeileis:** Probabilistic forecasting of thunderstorms in the Eastern Alps

2017-24 **Florian Lindner:** Choking under pressure of top performers: Evidence from biathlon competitions

2017-23 **Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, Achim Zeileis:** Estimation methods for non-homogeneous regression models: Minimum continuous ranked probability score vs. maximum likelihood

2017-22 **Sebastian J. Dietz, Philipp Kneringer, Georg J. Mayr, Achim Zeileis:** Forecasting low-visibility procedure states with tree-based statistical methods

2017-21 **Philipp Kneringer, Sebastian J. Dietz, Georg J. Mayr, Achim Zeileis:** Probabilistic nowcasting of low-visibility procedure states at Vienna International Airport during cold season

2017-20 **Loukas Balafoutas, Brent J. Davis, Matthias Sutter:** How uncertainty and ambiguity in tournaments affect gender differences in competitive behavior

2017-19 **Martin Geiger, Richard Hule:** The role of correlation in two-asset games: Some experimental evidence

2017-18 **Rudolf Kerschbamer, Daniel Neururer, Alexander Gruber:** Do the altruists lie less?

2017-17 **Meike Köhler, Nikolaus Umlauf, Sonja Greven:** Nonlinear association structures in flexible Bayesian additive joint models

2017-16 **Rudolf Kerschbamer, Daniel Muller:** Social preferences and political attitudes: An online experiment on a large heterogeneous sample

2017-15 **Kenneth Harttgen, Stefan Lang, Judith Santer, Johannes Seiler:** Modeling under-5 mortality through multilevel structured additive regression with varying coefficients for Asia and Sub-Saharan Africa

2017-14 **Christoph Eder, Martin Halla:** Economic origins of cultural norms: The case of animal husbandry and bastardy

2017-13 **Thomas Kneib, Nikolaus Umlauf:** A primer on bayesian distributional regression

2017-12 **Susanne Berger, Nathaniel Graham, Achim Zeileis:** Various versatile variances: An object-oriented implementation of clustered covariances in R

2017-11 **Natalia Danzer, Martin Halla, Nicole Schneeweis, Martina Zweimüller:** Parental leave, (in)formal childcare and long-term child outcomes

2017-10 **Daniel Muller, Sander Renes:** Fairness views and political preferences - Evidence from a large online experiment

2017-09 **Andreas Exenberger:** The logic of inequality extraction: An application to Gini and top incomes data

2017-08 **Sibylle Puntscher, Duc Tran Huy, Janette Walde, Ulrike Tappeiner, Gottfried Tappeiner:** The acceptance of a protected area and the benefits of sustainable tourism: In search of the weak link in their relationship

2017-07 **Helena Fornwagner:** Incentives to lose revisited: The NHL and its tournament incentives

2017-06 **Loukas Balafoutas, Simon Czermak, Marc Eulerich, Helena Fornwagner:** Incentives for dishonesty: An experimental study with internal auditors

2017-05 **Nikolaus Umlauf, Nadja Klein, Achim Zeileis:** BAMLSS: Bayesian additive models for location, scale and shape (and beyond)

2017-04 **Martin Halla, Susanne Pech, Martina Zweimüller:** The effect of statutory sick-pay on workers' labor supply and subsequent health

2017-03 **Franz Buscha, Daniel Müller, Lionel Page:** Can a common currency foster a shared social identity across different nations? The case of the Euro.

2017-02 **Daniel Müller:** The anatomy of distributional preferences with group identity

2017-01 **Wolfgang Frimmel, Martin Halla, Jörg Paetzold:** The intergenerational causal effect of tax evasion: Evidence from the commuter tax allowance in Austria

University of Innsbruck

Working Papers in Economics and Statistics

2018-03

Carolin Strobl, Julia Kopf, Raphael Hartmann, Achim Zeileis
Anchor point selection: An approach for anchoring without anchor items

**Abstract**
For detecting differential item functioning (DIF) between two groups of test takers, their item parameters need to be aligned in some way. Typically this is done by means of choosing a small number of so called anchor items. Here we propose an alternative strategy: the selection of an anchor point along the item parameter continuum, where the two groups best overlap. We illustrate how the anchor point is selected by means of maximizing an inequality criterion. It performs equally well or better than established approaches when treated as an anchoring technique, but also provides additional information about the DIF structure through its search path. Another distinct property of this new method is that no individual items are flagged as anchors. This is a major difference to traditional anchoring approaches, where flagging items as anchors implies - but does not guarantee - that they are DIF free, and may lull the user into a false sense of security. Our method can be viewed as a generalization of the search space of traditional anchor selection techniques and can shed new light on the practical usage as well as on the theoretical discussion on anchoring and DIF in general.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)