Strobl, Carolin; Kopf, Julia; Hartmann, Raphael; Zeileis, Achim

Working Paper

Anchor point selection: An approach for anchoring without anchor items

Working Papers in Economics and Statistics, No. 2018-03

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

Suggested Citation: Strobl, Carolin; Kopf, Julia; Hartmann, Raphael; Zeileis, Achim (2018) : Anchor point selection: An approach for anchoring without anchor items, Working Papers in Economics and Statistics, No. 2018-03, University of Innsbruck, Research Platform Empirical and Experimental Economics (eeecon), Innsbruck

This Version is available at:
https://hdl.handle.net/10419/184981

universität innsbruck

Research Platform
Empirical and Experimental Economics

Anchor point selection: An approach for anchoring without anchor items

Carolin Strobl, Julia Kopf, Raphael Hartmann, 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: An approach for anchoring without anchor items

Carolin Strobl
Universiät Zürich

Julia Kopf
Ludwig-Maximilians-Universität München

Raphael Hartmann
Albert-Ludwigs-Universität Freiburg

Achim Zeileis
Universität Innsbruck

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.

Keywords: item response theory (IRT), Rasch model, differential item functioning (DIF), anchor items, invariant item clusters.

1. Introduction

One of the major advantages of item response theory (IRT) 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, and Wollack 1996; Magis and De Boeck 2011) as well as its acceptance from the side of the test takers and policy makers.

Once DIF items are identified, they need to be improved or excluded from the final test form (cf., e.g., Westers and Kelderman 1992). But, in order to identify them, first the item parameters of the groups need to be aligned in some way that allows to compare the individual item parameters between the groups on one common scale. 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 (see, e.g., Teresi and 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 the search strategy to select those items.

Studies on the anchor length show 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 and Hagquist 2012) in the other items and thus drive up the false alarm rate of the following DIF tests. Since the effect of the anchor length also depends on the number of actual DIF items, which in pratice is unknown, an anchor length of three to five, most often four, items has been suggested as a compromise (cf. Shih and Wang 2009; Wang, Shih, and Sun 2012; Egberink, Meijer, and Tendeiro 2015). Note, however, that this number is somewhat arbitrary because its performance in a selection of simulation settings may not generalize well to all practical problems.

The different strategies that have been proposed for selecting the anchor items have been classified by Kopf, Zeileis, and Strobl (2015b) into several groups, including approaches that use all other items as preliminary anchors to judge the quality of a candidate item for a final anchor of a certain length (employed, e.g., by Wang 2004; Woods 2009) as well as iterative approaches (employed, e.g., by Candell and Drasgow 1988; Kopf, Zeileis, and Strobl 2015a; Kopf *et al.* 2015b).

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 and 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 achieved by placing the same restriction on the item parameters in both groups (as formalized, e.g., by Glas and Verhelst 1995; Eggen and Verhelst 2006) in order to define a common scale.

In the DIF literature, we can find several additional assumptions, some of which are made very explicitly while others seem to be taken for granted. Moreover, some of these assumptions may be more realistic than others, as well as more necessary than others, where the latter may also depend on the chosen anchoring approach. These implicit or explicit assumptions include the notions

- that some items in a test may have DIF, but that there are also items in the test that do not have any DIF at all (which may or may not be realistic),
- that, more specifically, it is the minority of items that have DIF and the majority of items that do not (which we would hope holds in a real test developed with lots of effort by content experts – but may restrict our theoretical thinking about the general concepts of anchoring and DIF, and is also critically discussed by Bechger and Maris 2015; Pohl, Stets, and Carstensen 2017),
- that in practice we do not know which items are the ones that have DIF and which are the ones that do not (or, thinking more continuously rather than in categories: which have more and which have less DIF),
- that DIF needs to be separated from actual ability differences by means of somehow conditioning on an estimate of the ability (Lord 1980; Van der Flier, Mellenbergh, Adèr, and Wijn 1984; DeMars 2010),

- and that those items that end up being selected into the anchor should be DIF free (which is a desirable property, 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 – yet there is a high risk that researchers consider those items that have been selected into the anchor as "safe").

The aim of this paper is to raise the question whether it is even necessary – or desirable – to select a set of individual items to form the anchor, when all we need is some restriction to align the scales between the groups. Therefore, here we suggest an approach for automatically selecting a point on the item parameter continuum, rather than a set of items, for achieving this alignment. We will do this by means of maximizing a criterion from poverty research, the Gini index, that is usually used to judge the inequality of wealth in a society, but will be applied to the item parameter differences here.

In this paper, we focus on the case of items that exhibit DIF between two groups of test takers, which are often termed reference group and focal group, as well as on the Rasch model as the underlying IRT model. This will allow us to focus on the properties of our newly suggested method under circumstances that are easy to understand and can be related to the existing literature on DIF and anchoring, that also has a strong focus on the two group case and the Rasch model. However, the extension of this approach to more general cases is possible and subject of ongoing research.

In the following, we will first introduce a little notation and two ways of illustrating the two group anchoring problem in a very simplified setting, that will help us understand and discuss the fundamentals. Then the new approach for finding the anchor point will be introduced. Its usefullness will be illustrated by means of an extensive simulation study, a few additional simulated toy examples (that are displayed graphically in the main body of the paper and by means of movie-like slide shows in online supplements), as well as an application example from a general knowledge quiz, where we will investigate DIF between female and male test takers.

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 and Verhelst 1995; Eggen and Verhelst 2006). When the aim is to compare the item parameters in two (or more) groups, the item parameters are first estimated separately. Since any linear restriction can easily be obtained from any other, it does not matter which particular restriction is applied in this first step. In the following, the initial item parameter estimates are termed $\tilde{\beta}_j^g$ for group g and item j.

However, as is illustrated in Figure 1, if these initial item parameter estimates were naively used for a comparison between the two groups, the choice of the restriction would 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 1, 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

and Stone 1999) are displayed. In this first column a), the first item parameter is arbitrarily set to 0 in both groups. In the direct comparison in the top row of plots, an item displays DIF if the item parameters are not aligned. 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 1, however, a different restriction was used: Here the sum of all items 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 item 4 is set to 0 in both groups in the third column c) of Figure 1, 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 this assumption, given the scale indeterminancy, it would not be possible to decide which scenario is the right one (see Section 3.2 for further discussion).

As a side note, the didactically very well written textbook of Wright and 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 already 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 and 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").

2.1. The choice of the restriction

Considering the choice of a suitable restriction for comparing the item parameters of two groups more technically, the minimum requirement is that the same restriction be used in both groups (Glas and Verhelst 1995). Since we have seen that the initial arbitrary restriction may be contaminated by DIF, however, a variety of strategies has been suggested to choose a set of suitable anchor items, whose sum is usually set to zero as the new restriction.

Let $\mathcal{A}$ be this set of anchor items and $|\mathcal{A}|$ its cardinality. The restriction can then be expressed as $\sum_{j\in\mathcal{A}} \hat{\beta}_j^g \stackrel{!}{=} 0$ and the final item parameter estimates $\hat{\beta}_j^g$ can then be derived from $\tilde{\beta}_j^g$ by means of shifting all item parameters by

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

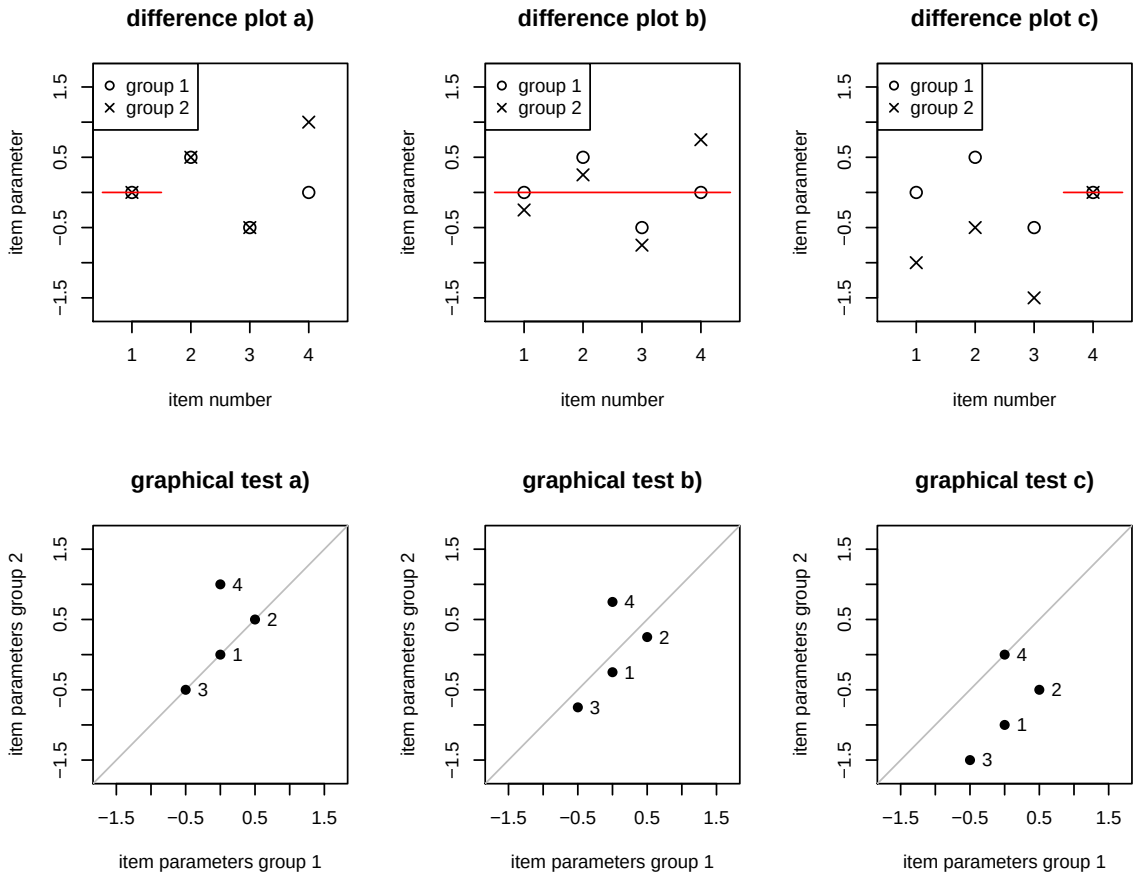

Figure 1: Illustration of comparisons of item parameters (top row) and graphical tests (bottom row) for different restrictions: a) $\beta_1^g = 0$, b) $\sum_j \beta_j^g = 0$, c)$\beta_4^g = 0$. Item numbers are displayed on the x-axis in the top row and next to the plotting symbols in the bottom row.

More abstractly speaking, this 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 $\mathcal{A}$ and can only take a limited number of different values, that result from the different combinations of anchor items that are being in- or excluded in $\mathcal{A}$.

Philosophically speaking, this approach to select items in or out of the anchor set strongly corresponds to the deterministic notion that an item either has DIF or not. This view may make sense for the true item parameters – and may also be an indicator of the political necessity to exclude all items with significant DIF and consider all remaining items as fair. However, even if the true item parameters were indeed equal between groups, their estimates would not be exactly equal due to random sampling fluctuation. In addition, one could consider it as more realistic that even the majority of "good" items may have some minor group difference even in their true parameters, while a minority of items does exhibit more severe DIF of a larger extent (here it would be useful to think of some reasonable measures

of effect size or practical impact on person parameter estimates or classification decisions to quantify this extent, like described, e.g., by Teresi and Jones 2016; Teresi *et al.* 2017).

Another adverse side effect of choosing designated anchor items is that – once they made it into the anchor – they are usually considered to be DIF free by definition (cf., e.g., Woods 2009), despite the fact that anchor selection strategies themselves are not flawless (see, e.g., the thorough discussion on anchor contamination in Kopf *et al.* 2015b).

Going back to our starting point, that we need some restriction or anchor only to fix the origin of the scale, the question is: Do we even need to select designated anchor items to choose this restriction, or could we just pick a single point in the latent continuum to fix our origin and avoid selecting a set of anchor items entirely?

We will show in the following that, yes, it is possible to uncouple the shift of the item parameters to a position that is ideal for DIF detection from a certain choice of anchor items. This can be accomplished by means of searching over a wider range of values for c^g, including values that do not result from any specific combination of anchor items. Therefore, in the following, c^g is no longer restricted to be equal to $\frac{\sum_{j \in \mathcal{A}} \hat{\beta}_j^g}{|\mathcal{A}|}$, but will be found by searching over an interval $[c_{\min}, c_{\max}]$. This interval will be chosen such that the item parameter ranges of both groups are safely overlapping, as described in more detail below.

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 (as illustrated in the movie-like slide shows in the online supplements) by means of shifting them by a constant

$$\hat{\beta}_j^{g_2} = \tilde{\beta}_j^{g_2} - c^{g_2} \text{ with } c^{g_2} = c,$$

which will be selected in a way that is suitable for DIF detection, as described below.

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 $\hat{\text{se}}_j = \sqrt{\widetilde{Var}(\tilde{\beta}^{g_1})_{j,j} + \widetilde{Var}(\tilde{\beta}^{g_2})_{j,j}}$. Note that we apply the item-wise Wald test to the conditional maximum likelihood estimates in the following (like in Glas and 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, as illustrated in Figure 1 as well as in the movie-like slide shows in the online supplements, for us as human beings it is straightforward to see that some positions are smarter than others and that the most suitable anchor point is the value for c where the item

parameter estimates for the majority of items will interlock. This is also the rationale behind the manually inserted "second identity line" through the "major item stream" of Wright and Stone (1999). 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 and – as argued above – the situation with few items having large differences is usually more intuitive.

To illustrate this, imagine again a scenario like in the first column a) of Figure 1, where one item shows a large distance and the rest of the items overlap, as opposed to a scenario like in the second column b) of Figure 1, where all items are shifted and show some distance between the groups. Both types of scenarios can produce the same sum of (squared/absolute) distances, but again it is the notion that only a minority of items has DIF makes most readers prefer the first scenario as the basis for DIF detection. Therefore, we would like the criterion for finding the most suitable value of c to respect this notion.

We will show in the next section that this can 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 the value of the anchor point we suggest to use the Gini index (Gini 1912, reprinted 1955). The Gini index is a popular inequality measure, that is usually employed for assessing the distribution of wealth 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, and lists are published how the ranking developes (cp., for example, Central Intelligence Agency 2017). Scandinavian countries, for example, tend to have a low Gini index indicating a rather equal income distribution, while the US, for example, has a medium Gini index and some countries that have been politically troubled in their recent history, such as Haiti, have a high Gini index.

We will now show how the Gini index can also be used as a means for selecting the anchor point. This is most easily imagined when the majority of items displays no DIF. Then at the 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* while the

majority has very little, we will use it here to see whether a minority of *items* has a lot of *DIF* (i.e., large absolute differences in their item parameters between the groups) while the majority has very little or no DIF.

The Gini index of the absolute item-wise distances $d_j(c)$ for the $j = 1, \ldots, m$ items can be computed via their order statistics $d_{(1)}(c), d_{(2)}(c), \ldots, d_{(m)}(c)$ as

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

The anchor point then corresponds to

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

We will show in the following illustrations how selecting c according to the Gini index leads to a shift between the two groups that makes their item parameters directly comparable. This approach can be used for any kind of graphical display as well as for DIF tests.

Currently we are using a grid search to find the maximum of the Gini index and illustrate the search path in the following sections. To make sure the grid search is feasible but includes all interesting positions, we use $[c_{\min}, c_{\max}] = \left[\min(\tilde{\beta}^{g_1}) - \max(\tilde{\beta}^{g_2}), \max(\tilde{\beta}^{g_2}) - \min(\tilde{\beta}^{g_1})\right]$, so 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). In between, the scales move past each other along a fine grid. Future research will also investigate more efficient optimization approaches.

In the illustrations below a grid with 1000 points is used and we will see that this is fine enough to include all locations that are relevant for the interpretation of our results. Conceptually, it is important to point out that the search also includes as special cases those locations that would result from selecting a particular set of anchor items (or, if the grid was really too coarse, points that are very close to them). In addition to these locations, however, it also includes plenty of other locations that are not defined through any particular set of anchor items. In this sense, the suggested strategy provides a generalization of the search space of traditional anchor selection techniques.

3. Illustrations

This section provides several illustrations of the properties of the new anchor point selection method under a variety of settings.

At first, we will show the results of an extensive simulation study, where the result of the anchor point selection is compared to two exemplary existing anchoring approaches from the literature. Here, we only use the final result, the anchor point itself, corresponding to the location of the global maximum of the Gini index, as the anchor and "throw away" the search path in order to be able to compare the performance of our new approach to that of traditional anchoring approaches.

In a second set of simulations, we will illustrate by means of a few additional toy examples how the search path, as well as the value of the Gini index at its maximum, provide additional information about the DIF pattern inherent to the data.

At the end of this section we illustrate the practical usage of the approach by means of an empirical example from a general knowledge quiz.

3.1. Simulation study I: Using the anchor point as a traditional anchor selection method in DIF testing

The anchor point selected by our new approach can be used for traditional anchoring in DIF tests just like any anchor selection strategy with designated anchor items. The following study compares its performance as a traditional anchor method to two other anchor methods that have been proposed in the literature, as well as to two baseline conditions.

From the variety of anchor methods available in the literature, we have included only the following two as examples in order to keep the results simple and be able to focus on the properties of the new method. The first exemplary comparison method is the anchor method suggested by Woods (2009), that is classified as the "constant four all other" method by the taxonomy of Kopf *et al.* (2015b), is based on the all other strategy: In the initial step, each item is tested for DIF using all remaining items as anchors. The four anchor 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 rather common approach, but has been shown to exhibit an increased false alarm rate in certain settings (Kopf *et al.* 2015b,a). It serves as a comparison method here to see whether the new approach is affected by similar problems. It should be pointed out, however, that Woods (2009) provides a very thorough discussion of anchor length choice, while here we consider only a fixed anchor length of four items because this is not the focus of our study.

The second exemplary comparison method is the more recent "constant four mean p-value threshold" 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 as single anchor (for a more detailed description see Kopf *et al.* 2015a). This method has been shown to be one out of two top performing methods in the extensive comparison study of Kopf *et al.* (2015a) and thus serves as a strong competitor here.

Besides the two competitor methods by Woods (2009) and Kopf *et al.* (2015a) and the anchor point selection method, we have included two baseline conditions. These conditions, termed "perfect ..." in the results section, serve only for illustration purposes and cannot be used for real data, because they are based on the information which items were simulated with and without DIF, which is of course not available in real data. Still these approaches will help us evaluate the performance of the actual anchoring methods, as is explained in more detail below.

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 and Verhelst 1995; Kopf *et al.* 2015b, and the original references therein) was used for the final DIF test, but as will become clear below, our results are of a general nature that straightforwardly generalizes to other DIF tests.

Simulation design

The simulation design was 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, but there are also scenarios that were simulated completely under the null hypothesis with no DIF in any item. Each data set represents one of 500 replications from one simulation setting.

- *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 without replacement, in the case of 60 with replacement).

- *DIF items*

  A randomly chosen set of items were simulated to display DIF by 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.

- *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 and the anchor methods.

- *Sample sizes*

  The sample size for the simulated data sets was varied between 500 and 3000 in steps of 500. 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, so that the comparisons between the methods are not affected by this factor. Therefore, and in the interest of saving space, here we present only results for equal group sizes because this makes the following plots easier to read.)

- *Directions and proportions of DIF*

  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), an advantage for the reference group or an advantage for the focal group.

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

- *Anchoring methods*

  The following methods were compared in this study:

  - the "constant four all other" method suggested by Woods (2009),
  - the "constant four mean p-value threshold" method suggested by Kopf *et al.* (2015a),
  - the "anchor point selection" method suggested in this paper,
  - the "perfect random selection", that randomly selects four anchor items from the set of items that (in the simulation) are known to be DIF-free,
  - the "perfect anchor point", that selects the anchor point from the grid that directly optimizes the false alarm rate and hit rate.

  Note that both "perfect ..." methods rely on information that is not available in practical settings and only serve as (unrealistic) baseline conditions here, as further explained below.

Outcome variables

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

Further specifications

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. To account for this in a way that is comparable between the anchoring methods, the item that was first selected into the "constant four mean p-value threshold" anchor by Kopf *et al.* (2015a), meaning that it showed the least indication of DIF, was left out of the final DIF test for all methods. For convenience the same anchor was also used as the initial restriction $\tilde{\beta}^{g_1}$ in the specification of the grid. Note again that this choice of the initial restriction has no effect on the result of the anchor point selection or the item-wise distances for the selected anchor point, but may lead to small differences in the standard errors $s\hat{e}_j$ of the Wald statistics.

Computational details

Our results were obtained using the R system for statistical computing (R Development Core Team 2017), version 3.3.0. The code for the anchor point selection was implemented by ourselves and will be made available in an R package. For the Gini index we used the implementation from the R package `ineq` (Zeileis 2014).

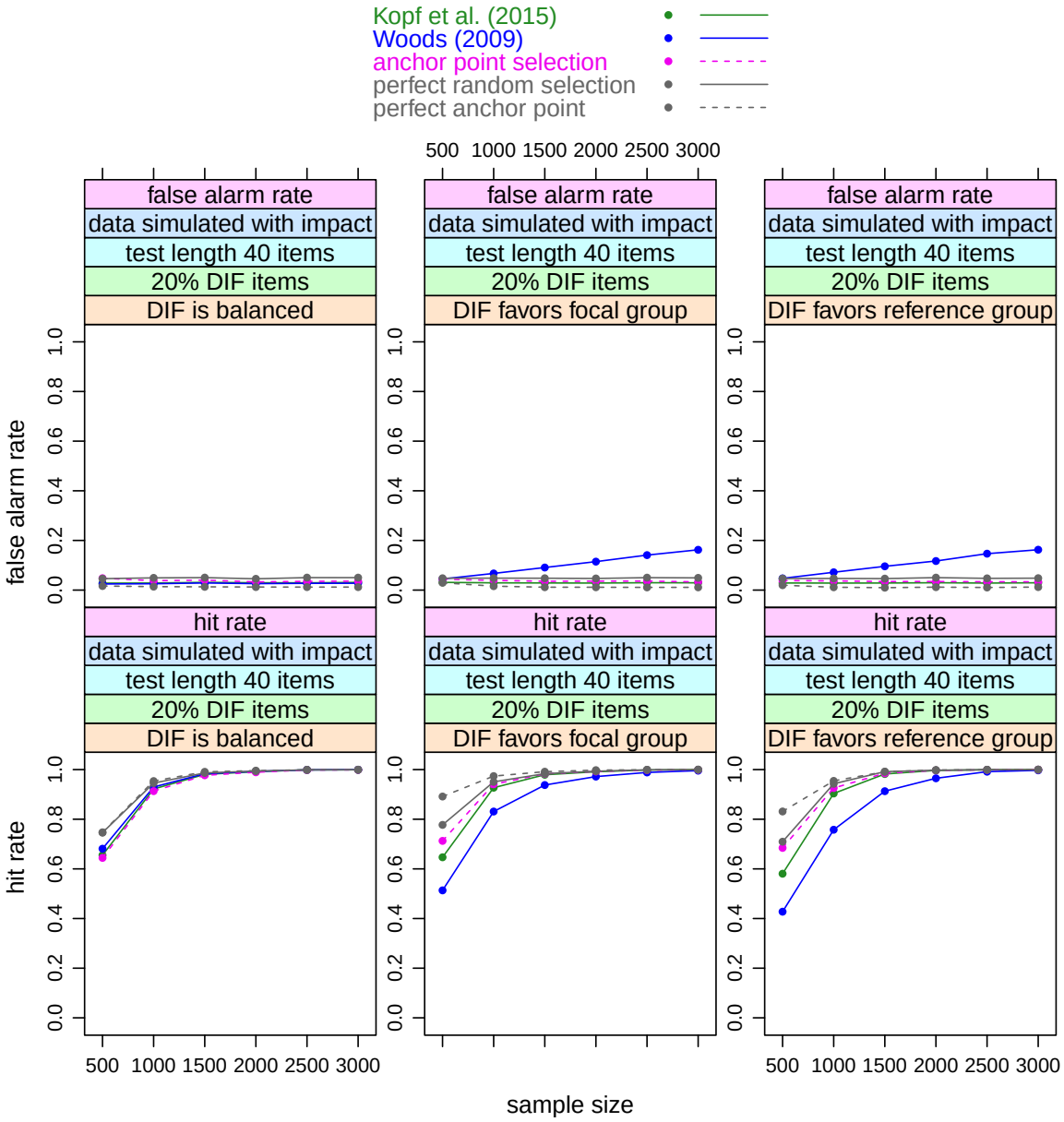

Figure 2: False alarm rate and hit rate for the simulation setting with a test length of 40 items and 20% DIF items.

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 are provided in an online supplement for comparison.

Figure 2 shows the results for a test length of 40 items and a DIF percentage of 20% (corresponding to 8 DIF items). The top row of Figure 2 shows the false alarm rate (percentage of items that were simulated as DIF free, but erroneously show a significant test result), the bottom row shows the hit rate (percentage of items that were in fact simulated to have DIF and correctly show a significant test result).

We see that in the balanced DIF setting all methods hold the nominal false alarm rate of 5%. In the unbalanced DIF conditions favoring the focal or the reference group, the method of Woods (2009) shows an increased false alarm rate, that also increases with the sample size, while the other methods again hold the nominal rate.

With respect to the hit rates, all methods perform very similarly in the balanced DIF setting, with the (unrealistic) "perfect ..." methods on top and the Woods (2009) method slightly outperforming the other two methods. In both unbalanced settings the (unrealistic) "perfect ..." methods again perform best, followed by the newly suggested "anchor point selection" method and then by the methods of Kopf *et al.* (2015a) and Woods (2009). (Note, however, that since the Woods (2009) method did not hold the nominal false alarm rate, its hit rate should be interpreted with caution.)

The fact that the (unrealistic) "perfect ..." methods perform best is not surprising, but an important proof of concept. The "perfect random selection" randomly samples four anchor items from the set of items that – in the simuation setting, unlike in reality – are known to be DIF-free. This condition serves as a baseline for evaluating the false alarm rate and hit rate of the underlying DIF test when we randomly select an appropriate set of four items as an anchor, by definition ruling out any anchor contamination.

The "perfect anchor point" selects the anchor point from the grid that directly optimizes the false alarm rate and hit rate, rather than our empirical Gini criterion. Again, this is only possible in the simulation setting, not in real data. However, this condition serves as a baseline for how well the anchor point selection could in principle perform. If this performance is better than that of the other anchoring approaches, we can be sure that our grid is fine enough to include the ideal solution and any inferior performance of the anchor point selection method is due to uncertainty arising from the empirical data.

In Figure 3 we see a similar picture for a test length of 40 items and a DIF percentage of 40% (corresponding to 16 DIF items). For the smallest sample sizes, however, we find that all methods have a slightly inflated false alarm rate, particularly in the unbalanced DIF settings, with the anchor point selection method and the Woods (2009) method showing the most notable inflation.

We find the same tendency also for the other test lengths combined with high percentages of DIF items and small samples sizes (in Figure 13 in the online supplement for a test length of 20 items and a DIF percentage of 40% and, less pronounced, in Figure 16 in the online supplement for a test length of 60 items and a DIF percentage of 40%). For a DIF percentage of 20% the inflation of the false alarm rate is negligible for either test length. The inflation in the settings with 40% DIF items is most pronounced for the "anchor point selection" method, but we see the same pattern to a reduced extent for the Kopf *et al.* (2015a) method and even for the "perfect anchor point".

This inflation effect of the anchor point method (and to a lesser extend the Kopf *et al.* (2015a) method) appears only in combinations with small sample sizes and high DIF percentages. Since the true DIF percentage is unknown in practice, one should generally keep in mind that for small sample sizes the final DIF tests derived from these methods may show more false positive results than the nominal level of the test would suggest, but that this problem disappears for larger sample sizes.

For the Woods (2009) method, on the other hand, the inflated false alarm rate present in all unbalanced settings does not go away but even increases for larger sample sizes. Since it can

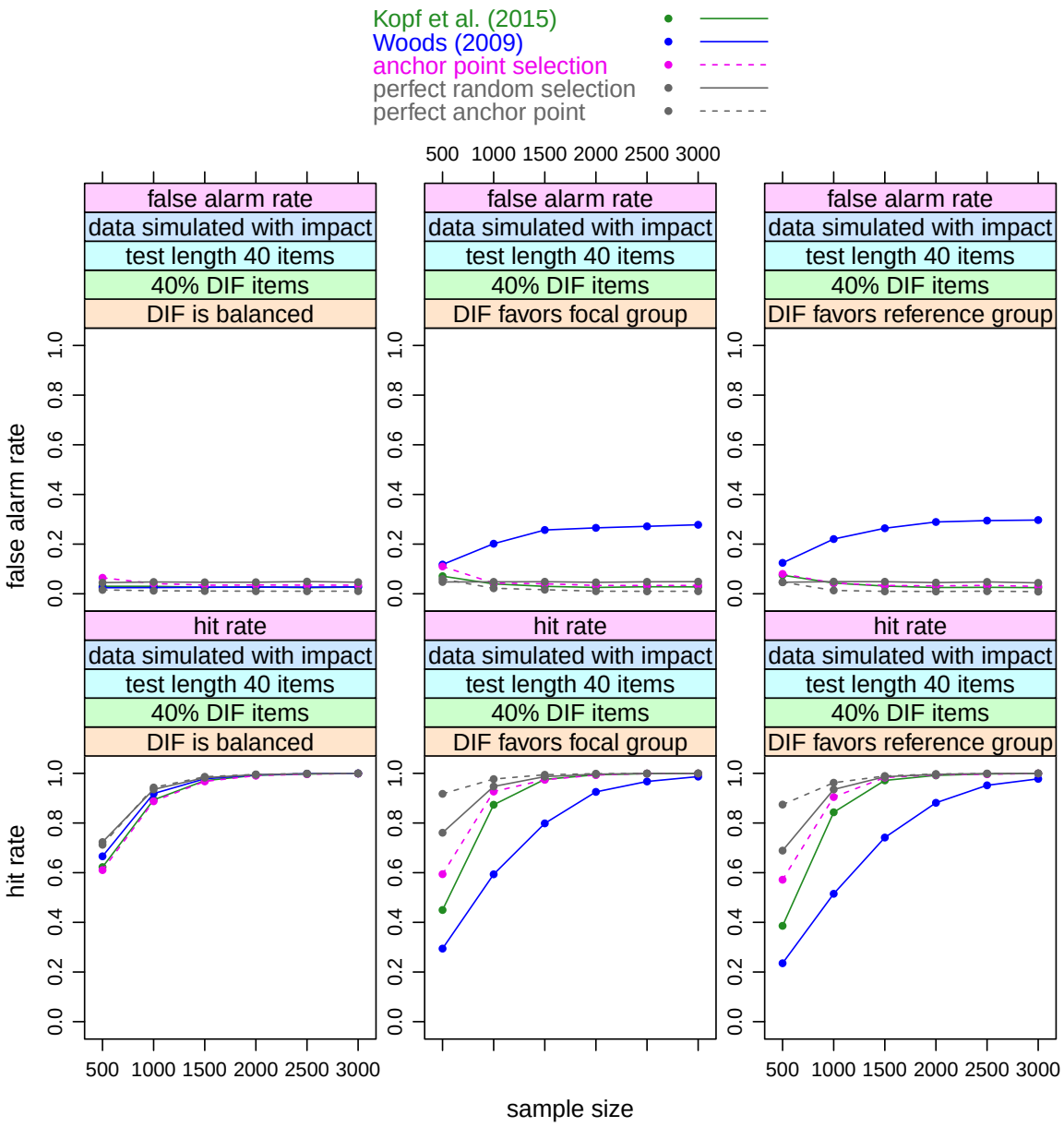

Figure 3: False alarm rate and hit rate for the simulation setting with a test length of 40 items and 40% DIF items.

go up to 0.4 in certain settings (cf. Figure 16 in the online supplement), and in reality we have no means of knowing whether this is the situation that underlies our data, we do not recommend this method, despite the fact that in some other scenarios this method shows the highest hit rate of the three empirically applicable methods.

With respect to the hit rates, Figure 3 shows again that for the balanced DIF setting all methods perform very similarly, with the (unrealistic) "perfect ..." methods on top and the Woods (2009) method slightly outperforming the other two empirically applicable methods.

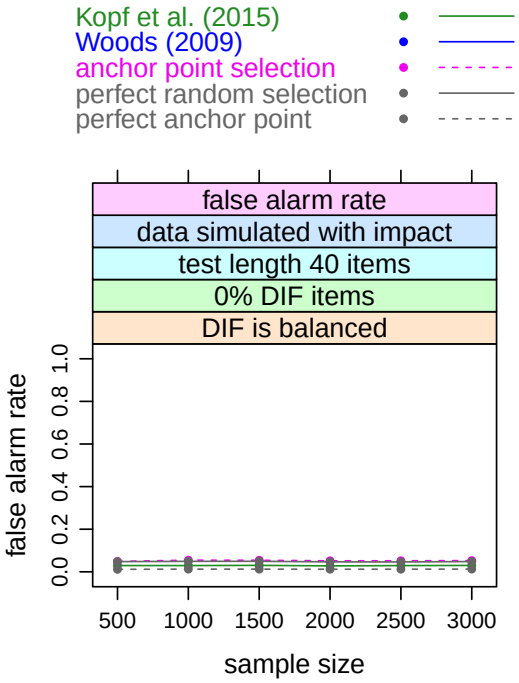

Figure 4: False alarm rate and hit rate for the simulation setting with a test length of 40 items and 0% DIF items.

In both unbalanced settings again the (unrealistic) "perfect ..." methods perform best, followed by the newly suggested "anchor point selection" method and then by the methods of Kopf *et al.* (2015a) and Woods (2009). Here it should again be pointed out that the hit rate for the anchor point method should not be overinterpreted in those scenarios with a high DIF percentage for the smallest sample sizes (first points in the respective plots), where its false alarm rate is inflated. However, the order of the hit rates remains the same for the larger sample sizes where its false alarm rate is no longer inflated.

For the other test lengths (presented in the online supplement) we find the same order of the methods with respect to their hit rates, with the "anchor point selection" on par or slightly inferior to the other two empirically applicable methods in the balanced setting and on par or slightly superior to the other two empirically applicable methods in the unbalanced settings, and sometimes even on par with the (unrealistic) "perfect random selection". Again, note that the hit rate for the anchor point method should not be overinterpreted in those scenarios with a high DIF percentage for the smallest sample sizes (first points in the respective plots), where its false alarm rate is inflated.

For large sample sizes (from about 2000 observations), the performances of the "perfect ..." methods, the "anchor point selection" and the method of Kopf *et al.* (2015a) converge, while the Woods (2009) method keeps showing an inflated false alarm rate and a lower hit rate.

For completeness, we also present a null hypothesis scenario with no DIF items for every test length. As Figure 4 shows, all methods perform very similarly and hold the nominal false alarm rate.

In summary, the results of this first simulation study show that the new anchor point selection method, when treated as an anchor selection approach, performs comparably well to a modern method like that of Kopf *et al.* (2015a) and more reliably than the traditional method of Woods (2009) in unbalanced settings. It does show an inflated false alarm rate in settings where a small sample size coincides with a high percentage of DIF items, particularly in combination with a short test length. However, unlike for the Woods (2009) method, this effect disappears when sample size increases. We conclude that the newly suggested anchor point selection method is well suited as an anchor selection approach in DIF testing as long as the sample sizes and test lengths are not too small.

3.2. Simulation study II: Getting additional information out of the anchor point selection search path

Now we want to show a few prototypical simulated settings that best illustrate the additional information that can be drawn from the search path of the anchor point selection method.

The data were again generated by means of a Rasch model and with two groups of 1000 simulated participants each.

For the item parameters in the reference group, we chose a very simple pattern: $\beta = (2, 1, 0, -1, -2, -2, -1, 0, 1, 2)$. This setting is not meant to be realistic, but only to be easily visually recognizable in the following plots.

The item parameters in the focal group deviate from the item parameters in the reference group by means of simple patterns: either only one item parameter deviates by a certain amount (for the first three illustrations) or there are groups of item parameters that deviate by the same amount (for the last three illustrations).

In Figure 5 we see the results for an idealized scenario with one DIF item and a DIF magnitude of 2. The top plot of Figure 5 displays the item locations for the anchor point selection, with items that show significant DIF in the final DIF test being colored in red and items that do not show significant DIF for this item location colored in green. The bottom plot of Figure 5 shows the search path of the anchor point selection, with its maximum highlighted by the red line. The magnitude of the Gini index at its maximum for this scenario is 0.743.

We find that in this scenario there is one clear maximum visible in the search path[1], that corresponds to the item parameter location where clearly one item shows DIF and all other items interlock.

When we compare this to the results in Figure 6 for an idealized scenario with again one DIF item but now a higher DIF magnitude of 4, we find that again there is one clear maximum visible in the search path, that corresponds to the item parameter location where clearly one item shows DIF and all other items interlock. However, now the magnitude of the Gini index at its maximum is higher at 0.807.

For comparison, Figure 7 shows the results for a scenario with no DIF at all. The maximum Gini value again corresponds to the item parameter location where the groups align (here with no item showing DIF and all items interlocking). Now also the magnitude of the Gini index at its maximum is lower at 0.430. This shows that the magnitude of the Gini index at its maximum is an indicator of the amount of DIF, as is further discussed below.

[1] With a few similarly high peaks right next to it, which are due to sampling fluctuation and the fine grained grid, and lead to very similar item locations as well as the same test decisions.

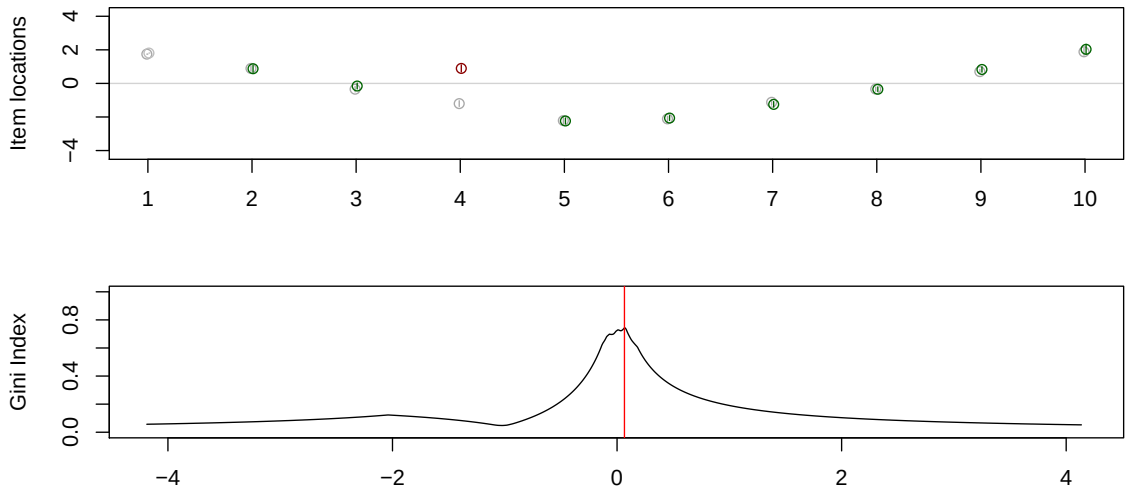

Figure 5: Top plot: Item locations for selected anchor point. Bottom plot: Search path of anchor point selection. Scenario with a single DIF item and DIF size 2. Global maximum of Gini index = 0.743.

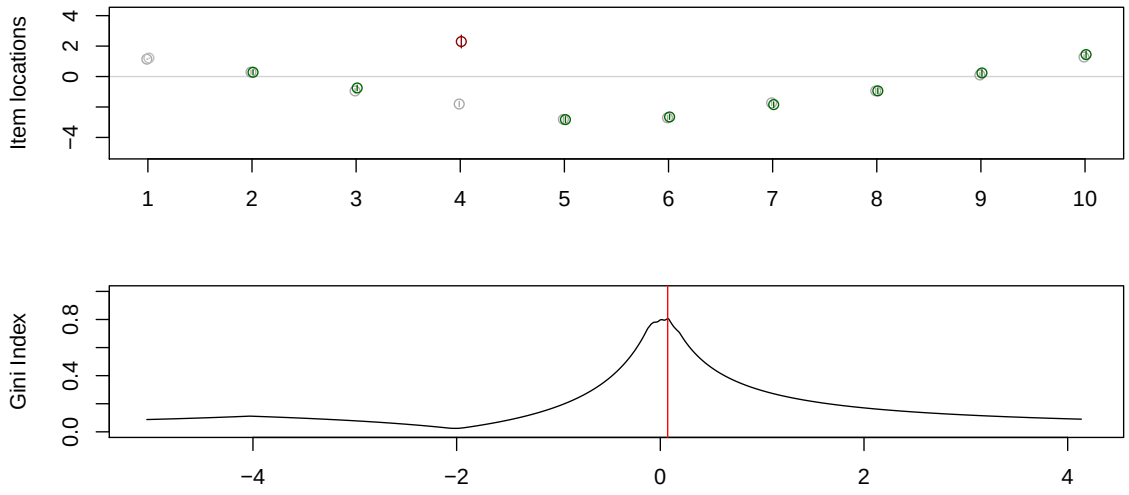

Figure 6: Top plot: Item locations for selected anchor point. Bottom plot: Search path of anchor point selection. Scenario with a single DIF item and DIF size 4. Global maximum of Gini index = 0.807.

In the following illustrations, we want to go further and point out another characteristic of the search path, that may be interesting at least from a theoretical point of view. This illustration has been stimulated by discussions with students, who, when introduced to the concept of DIF, often come to the point where they realize that if half the items of a test had DIF of the same size and direction, we could not decide by means of statistical reasoning which of the two possible alignments is "correct". In lack of a better term, we will call this a "tie" in the following. As discussed earlier, large parts of the DIF and anchoring literature

address this issue by making the implicit or explicit assumption that the majority of items function correctly and only a minority of items exhibits DIF. While this assumption seems to be plausible in real test construction settings, where a lot of expertise is available for constructing fair items, it is not testable empirically – and we might profit conceptually if we relax this assumption. We will get back to this topic shortly.

In the next scenario, we again use a DIF size of 2, but now not only one but three out of the ten items are affected by the DIF. Figure 8 shows the results for this scenario. We find that again there is one clear global maximum, B, visible in the search path, that corresponds to the item parameter location where clearly the first three items show DIF and all other items interlock. However, now there is also another smaller local maximum, A, which corresponds to the item parameter location where the first three items interlock and all other items show DIF.

The magnitude of the Gini index at its global maximum B in this scenario is 0.655. Compared to the earlier scenarios, this value is clearly higher than the result in the scenario with no DIF, but lower than in the scenario with a single item showing DIF of the same size 2. Here the property of the Gini index as an inequality measure comes into play, that assigns a higher inequality to the case where one item "owns all the DIF" (see again the comparison to one rich person owning all the wealth in Section 2.2) than to the case where a group of items "share the DIF". This shows that the Gini index as a measure of inequality captures not only the item location where the groups best align, but its magnitude also reflects the amount and distribution of DIF that is still present at this maximal position. In future research, this property could be further investigated and potentially be used as an additional information in DIF research.

In the next setting we go even further to illustrate the "tie" phenomenon. We simulated two equally sized groups of 5 items each, where the first five items were shifted by a DIF size of 2. Figure 9 shows the results for this scenario. We find that now there is not one clear maximum visible in the search path, but there are two maxima of virtually equal hight (differences are only due to sampling fluctuation). The first maximum, A, corresponds to the location where the first five items interlock and the rest show DIF, and the second maximum, B, to the opposite location.

Philosophically speaking, this "tie" scenario cannot be decided by statistical reasoning – but traditional anchoring approaches would still try to select one set of anchor items from either group of items, which here would be an arbitrary choice based only on small random differences in the item parameter estimates. The search path of the anchor point selection, on the other hand, shows visually that there are two solutions that are equally appropriate, and that some additional reasoning (such as the judgement of content experts) is necessary to resolve this tie. This is an important additional information and a very intuitive illustration of the properties of our search approach.

Here we also see a strong parallel to the works of Bechger and Maris (2015) and Pohl *et al.* (2017), who critically discuss the widely used assumption that the majority of items is DIF free and aim at the detection of invariant item clusters. In our approach, multiple invariant item clusters seem to correspond to multiple local maxima of the Gini index. Therefore, the information on the number and location of item clusters is already inherent in the search path. In this sense, the search path may also be helpful for guiding the decision whether for a particular test the final DIF tests should be based on the assumption that the majority

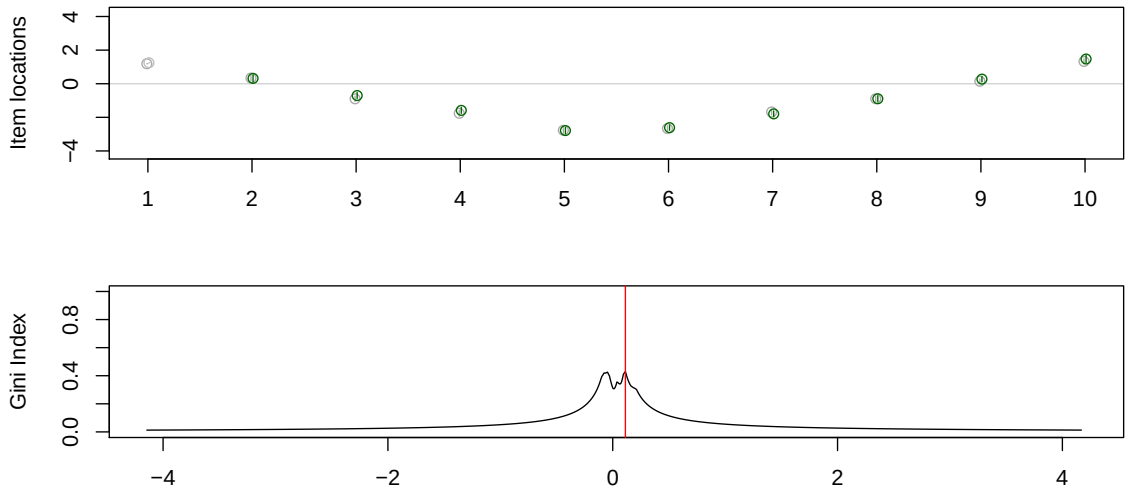

Figure 7: Top plot: Item locations for selected anchor point. Bottom plot: Search path of anchor point selection. Scenario with no DIF. Global maximum of Gini index = 0.4307.

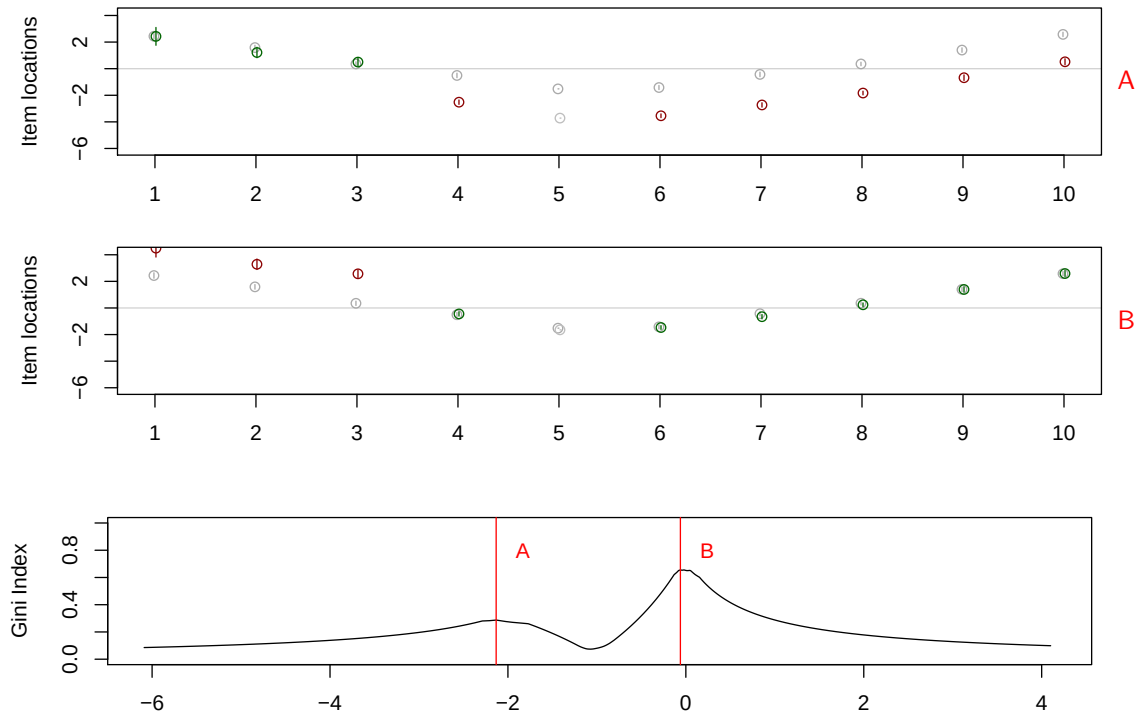

Figure 8: Top plot: Item locations for local maximum A. Middle plot: Item locations for global maximum B. Bottom plot: Search path of anchor point selection. Scenario with a smaller group of three DIF items and DIF size 2. Global maximum of Gini index = 0.655.

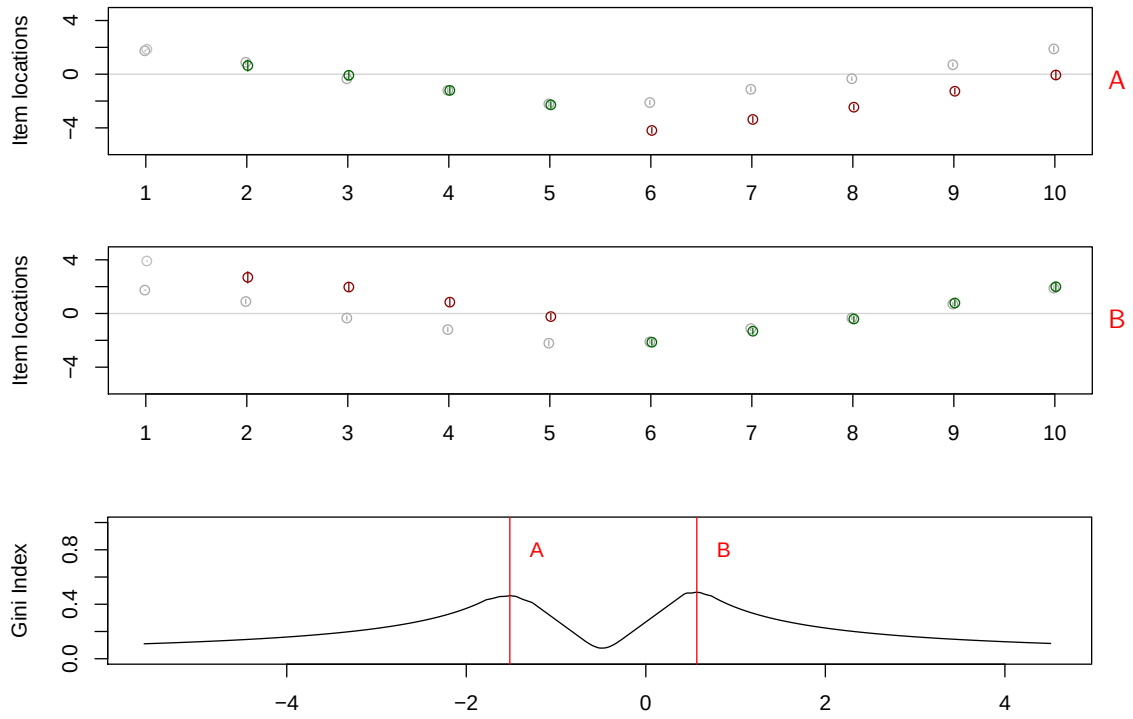

Figure 9: Top plot: Item locations for local maximum A. Middle plot: Item locations for local maximum B. Bottom plot: Search path of anchor point selection. Scenario with two equally sized groups and DIF size 2. Global maximum of Gini index = 0.489.

of items is DIF free (which we consider sensible in the case of a clear global maximum) or whether this assumption should be given up in the presence of multiple invariant item clusters (corresponding to mutliple local maxima, as further discussed below).

As a final illustration, another scenario with three groups of similar sizes was simulated, that straightforwardly leads to a search path with three local maxima[2], as illustrated in Figure 10, as we would expect.

The toy examples we have presented in this section may appear artificial, but they highlight an important theoretical property of our method: If a test does not only contain a few individual DIF items, but consists of groups of items that have DIF of the same type Bechger and Maris (corresponding to the item clusters of 2015, and to local maxima in our approach), the search path gives us a chance to detect this kind of pattern.

Such item clusters may be due to a second (or third) trait that these items measure in addition to the trait of interest. As has been pointed out by Ackerman (1992), if groups of test takes differ in their aptitude on these additional dimensions (such as test takers who are not native speakers of the test language having trouble with word problems in a math test), this can manifest itself as DIF. Another possible source of DIF are response styles (cf., e.g., Bolt and

[2]Please note that the last item in the third plot (C) is colored in red because it shows significant DIF. This is a false positive result based on sampling fluctuation, which is in accordance with the type one error rate of the Wald test. Still the item parameter locations displayed in the third plot (C) are correctly identified as a local maximum in accordance with the simulation settings.

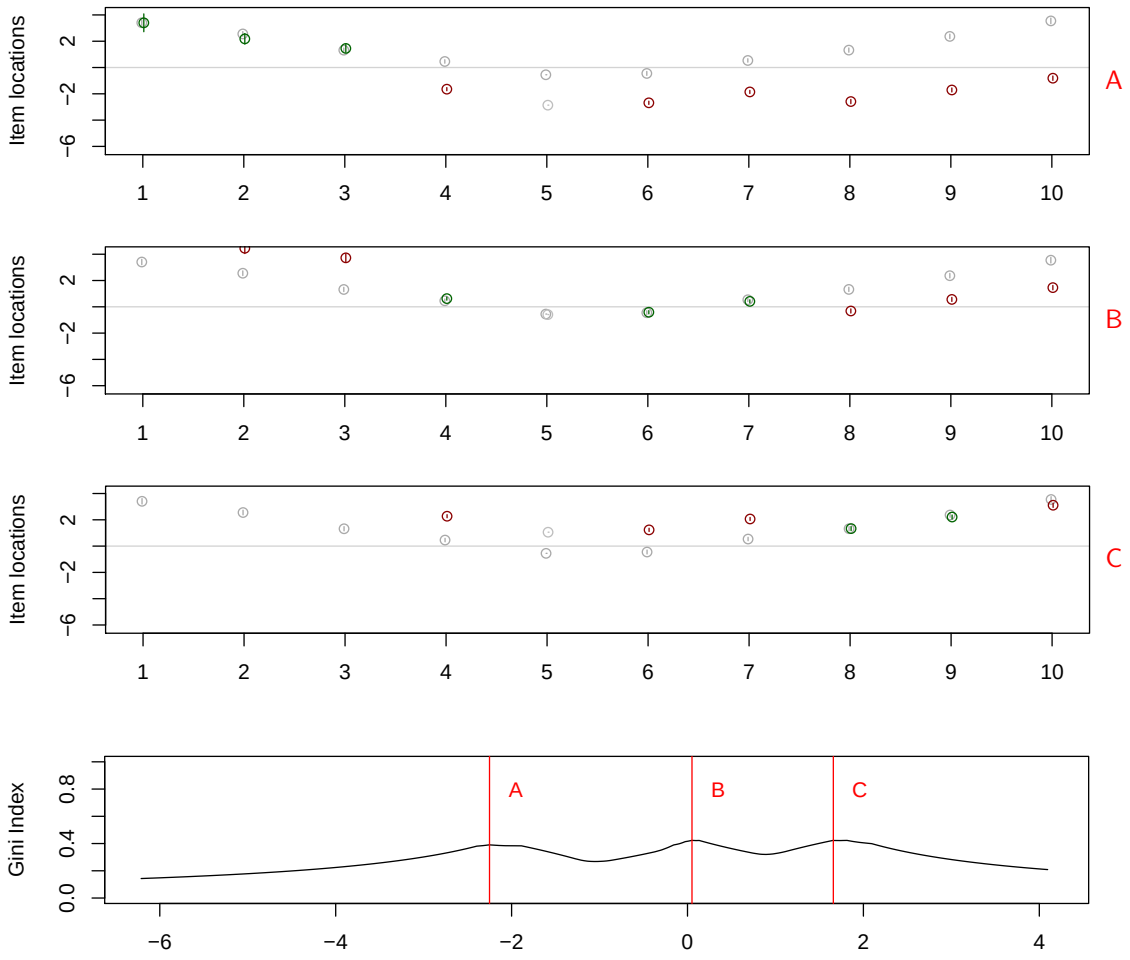

Figure 10: Top plot: Item locations for local maximum A. Second plot: Item locations for local maximum B. Third plot: Item locations for local maximum C. Bottom plot: Search path of anchor point selection. Scenario with three equally sized groups and DIF size 2. Global maximum of Gini index = 0.423.

Johnson 2009; Wetzel, Böhnke, Carstensen, Ziegler, and Ostendorf 2013).

What our toy examples have shown is that – while traditional anchoring approaches aim at the majority of items – our search path conserves the information that (and where) there are groups of items that go together and measure something different from the intended trait, which may be a valuable information for the decision how to proceed with the DIF items: kick them out, modify their mode of presentation, work towards a multidimensional test if the additional traits are not considered as nuisance, and/or further investigate and try to address the influence of response styles.

3.3. Application example: Students' PISA data

For further illustration we will now apply the anchor point selection method to empirical data from an online quiz for testing one's general knowledge, that was conducted by the German

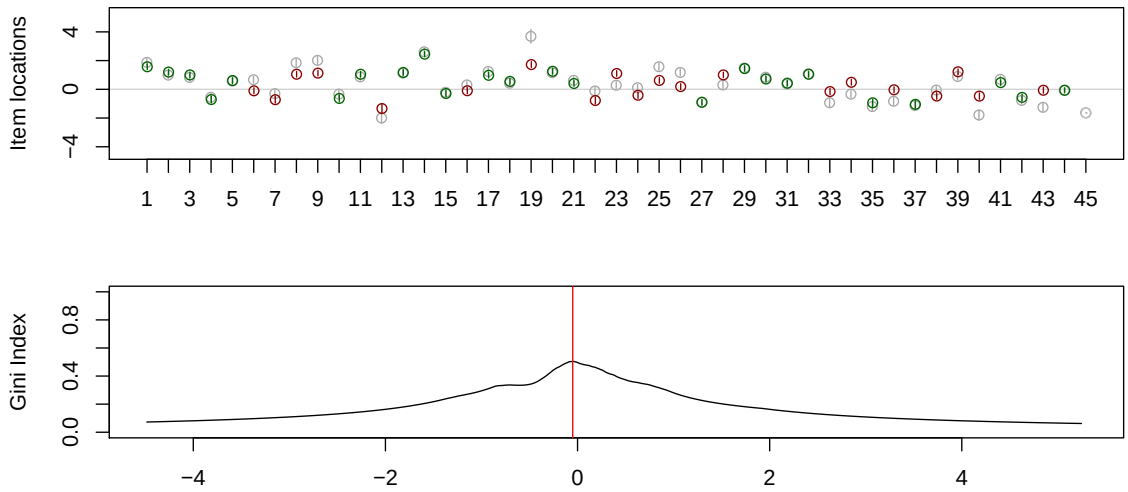

Figure 11: Top plot: Item locations for selected anchor point. Bottom plot: Search path of anchor point selection. Students' PISA data. Global maximum of Gini index = 0.504.

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 and 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, where also the wording of all 45 items contained in this questionnaire is documented.

The result of the anchor point selection for the Students' PISA data is depicted in Figure 11. From the path in the bottom plot in Figure 11, we can see that there is a clear, single maximum for the Gini index, so there are no larger groups of items that function the same way and may form an additional dimension.

The item locations at this maximum are shown in the top plot in Figure 11. As in the previous section, items colored in green do not show significant DIF, while items colored in red do show significant DIF according to the Wald test. The entire movie showing the anchor point selection for this example can again be found in an online supplement.

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 11 by the largest absolute distance between the item parameters in the male group, represented by the colored symbols, and the female group, represented by the grey symbols) 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.

This example also shows that the item location derived from the anchor point selection provides a helpful means of visually aligning the item parameters for interpretation.

4. Summary and discussion

In this paper we have suggested a new approach for anchoring the item parameters of two groups of test takers for DIF testing, that is not based on an anchor in its original sense. This new approach does not select a set of designated anchor items, but searches for the point on the item parameter continuum where the groups best overlap. The resulting item parameter locations at the maximum criterion value can be used as a traditional anchor in DIF testing, but the entire search path also provides additional information about the item structure.

We have shown by means of extensive simulation studies that the new anchor point selection method performs comparably well or more reliably than existing anchor selection approaches when its single best result is used as an anchor for DIF testing. It does show an inflated false alarm rate in settings where a small sample size coincides with a high percentage of DIF items and a short test length, but this effect disappears when sample size increases.

In addition to being usable as an anchor selection strategy, however, our method also provides a lot of extra information through its search path, as well as through the magnitude of the inequality criterion itself. While the latter should be investigated more thoroughly in future research, we have already shown here by means of a set of simulated toy examples that the method is able to identify clusters of items that function similarly and may represent additional trait dimensions.

We believe another advantage of the anchor point selection approach is that it does not flag individual items as anchors – a practice that all traditional anchoring approaches share, and that may lead to wrongly assuming those items were indeed DIF free.

In future research we will work on extending this approach to multiple groups of test takers (such as different language groups) and other IRT models as well as a more efficient optimization approach.

Acknowledgements

The first author would like to thank Thomas Augustin for his encouragement when she first had a very vague idea of this a very long time ago.

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

References

Ackerman TA (1992). "A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective." *Journal of Educational Measurement*, **29**(1), 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x.

Andrich D, Hagquist C (2012). "Real and Artificial Differential Item Functioning." *Journal of Educational and Behavioral Statistics*, **37**(3), 387–416. doi:10.3102/1076998611411913.

Bechger TM, Maris G (2015). "A Statistical Test for Differential Item Pair Functioning." *Psychometrika*, **80**(2), 317–340. doi:10.1007/s11336-014-9408-y.

Bolt DM, Johnson TR (2009). "Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style." *Applied Psychological Measurement*, **33**(5), 335–352. doi:10.1177/0146621608329891.

Candell GL, Drasgow F (1988). "An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory." *Applied Psychological Measurement*, **12**(3), 253–260. doi:10.1177/014662168801200304.

Central Intelligence Agency (2017). "The World Factbook: Distribution of Family Income – Gini Index." URL https://www.cia.gov/library/publications/the-world-factbook/fields/2172.html.

Cohen AS, Kim SH, Wollack JA (1996). "An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning." *Applied Psychological Measurement*, **20**(1), 15–26. doi:10.1177/014662169602000102.

DeMars CE (2010). "Type I Error Inflation for Detecting DIF in the Presence of Impact." *Educational and Psychological Measurement*, **70**(6), 961–972. doi:10.1177/0013164410366691.

Egberink IJL, Meijer RR, Tendeiro JN (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. doi:10.1177/0013164414520965.

Eggen T, Verhelst N (2006). "Loss of Information in Estimating Item Parameters in Incomplete Designs." *Psychometrika*, **71**(2), 303–322. doi:10.1007/s11336-004-1205-6.

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

Gini C (1912, reprinted 1955). "Variabilità e Mutabilità." In E Pizetti, T Salvemini (eds.), *Memorie Di Metodologica Statistica*. Libreria Eredi Virgilio Veschi, Rome.

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

Kopf J, Zeileis A, Strobl C (2015a). "Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches." *Educational and Psychological Measurement*, **75**(1), 22–56. doi:10.1177/0013164414529792.

Kopf J, Zeileis A, Strobl C (2015b). "A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection." *Applied Psychological Measurement*, **39**(2), 83–103. doi:10.1177/0146621614544195.

Lord F (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, New Jersey. doi:10.4324/9780203056615.

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. doi:10.1080/00273171.2011.606757.

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.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Shih CL, Wang WC (2009). "Differential Item Functioning Detection Using the Multiple Indicators, Multiple Causes Method with a Pure Short Anchor." *Applied Psychological Measurement*, **33**(3), 184–199. doi:10.1177/0146621608321758.

Teresi JA, Jones RN (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. doi:10.1037/e689602007-001.

Teresi JA, Ocepek-Welikson K, Toner JA, Kleinman M, Ramirez M, Eimicke JP, Gurland BJ, Siu A (2017). "Methodological Issues in Measuring Subjective Well-Being and Quality-of-Life: Applications to Assessment of Affect in Older, Chronically and Cognitively Impaired, Ethnically Diverse Groups Using the Feeling Tone Questionnaire." *Applied Research in Quality of Life*, **12**(2, SI), 251–288. doi:10.1007/s11482-017-9516-9.

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

Van der Flier H, Mellenbergh GJ, Adèr HJ, Wijn M (1984). "An Iterative Item Bias Detection Method." *Journal of Educational Measurement*, **21**(2), 131–145. doi:10.1111/j.1745-3984.1984.tb00225.x.

Wang WC (2004). "Effects of Anchor Item Methods on the Detection of Differential Item Functioning within the Family of Rasch Models." *Journal of Experimental Education*, **72**(3), 221–261. doi:10.3200/jexe.72.3.221-261.

Wang WC, Shih CL, Sun GW (2012). "The DIF-Free-Then-DIF Strategy for the Assessment of Differential Item Functioning." *Educational and Psychological Measurement*, **72**(4), 687–708. doi:10.1177/0013164411426157.

Westers P, Kelderman H (1992). "Examining Differential Item Functioning Due to Item Difficulty and Alternative Attractiveness." *Psychometrika*, **57**(1), 107–118. doi:10.1007/bf02294661.

Wetzel E, Böhnke JR, Carstensen CH, Ziegler M, Ostendorf F (2013). "Do Individual Response Styles Matter?" *Journal of Individual Differences*, **34**(2), 69–81.

Woods CM (2009). "Empirical Selection of Anchors for Tests of Differential Item Functioning." *Applied Psychological Measurement*, **33**(1), 42–57. doi:10.1177/0146621607314044.

Wright BD, Stone M (1999). *Measurement Essentials*. Wide Range Inc., Wilmington.

Zeileis A (2014). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-13, URL https://CRAN.R-project.org/package=ineq.

A. Additional results for simulation study I

A.1. Test length 20

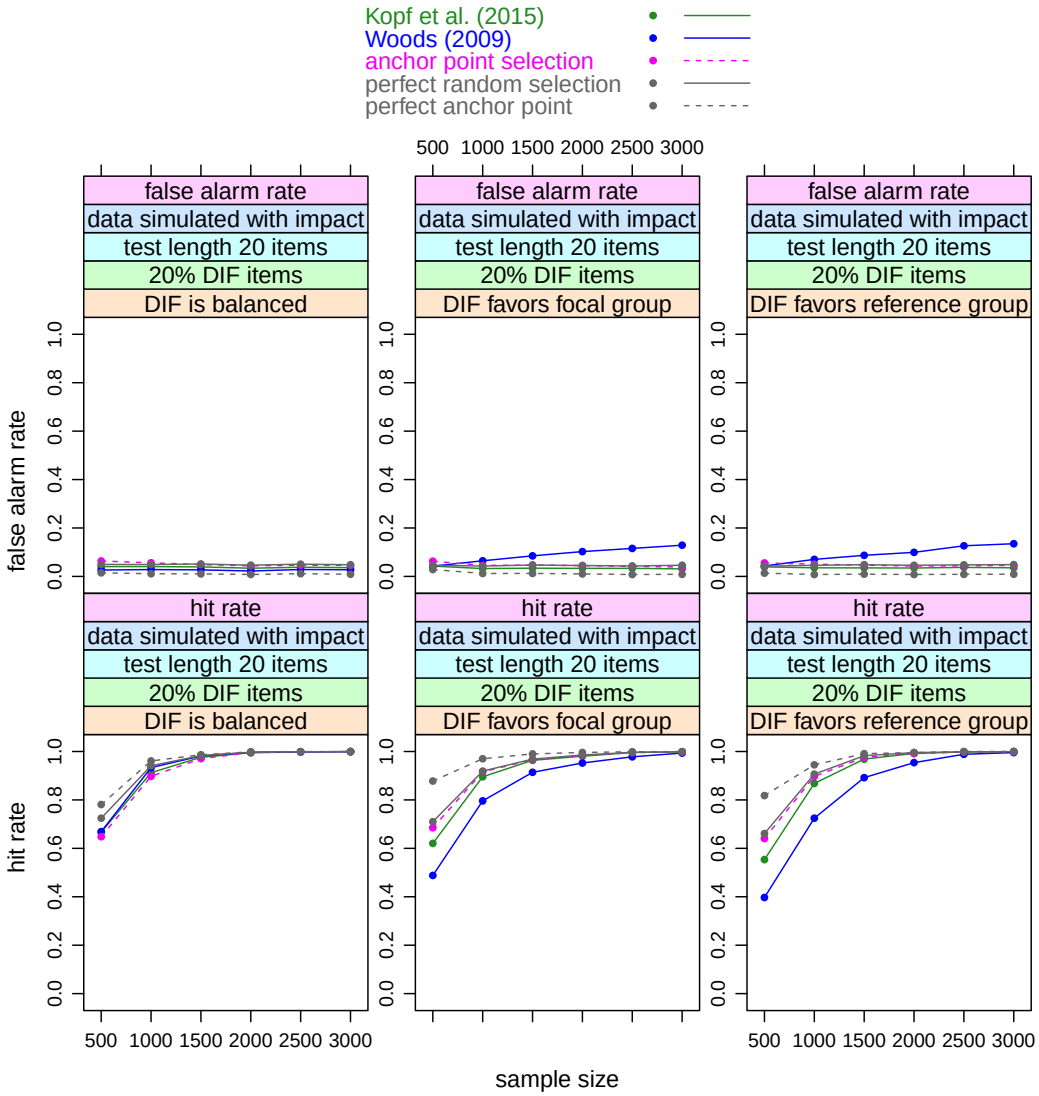

Figure 12: False alarm rate and hit rate for the simulation setting with a test length of 20 items and 20% DIF items.

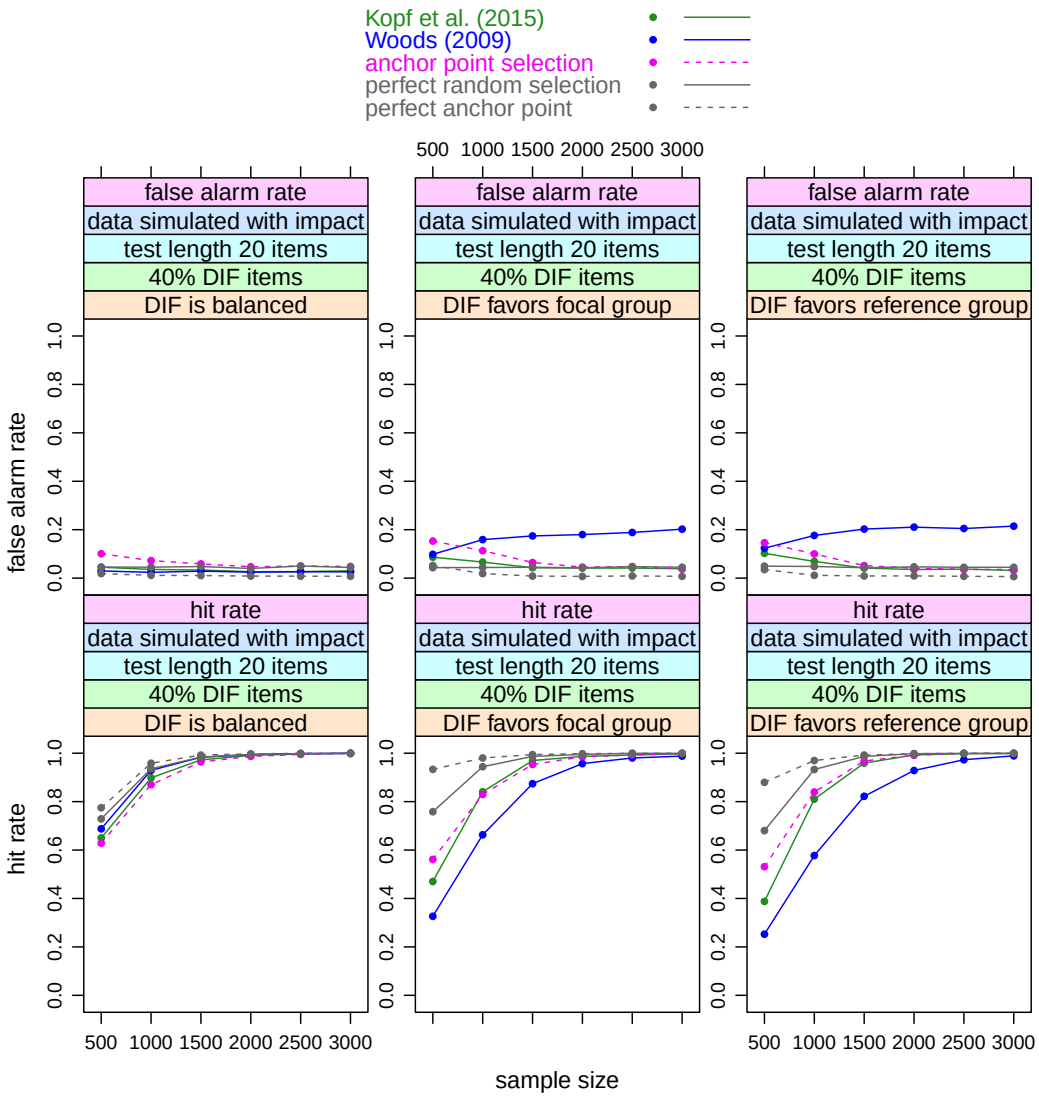

Figure 13: False alarm rate and hit rate for the simulation setting with a test length of 20 items and 40% DIF items.

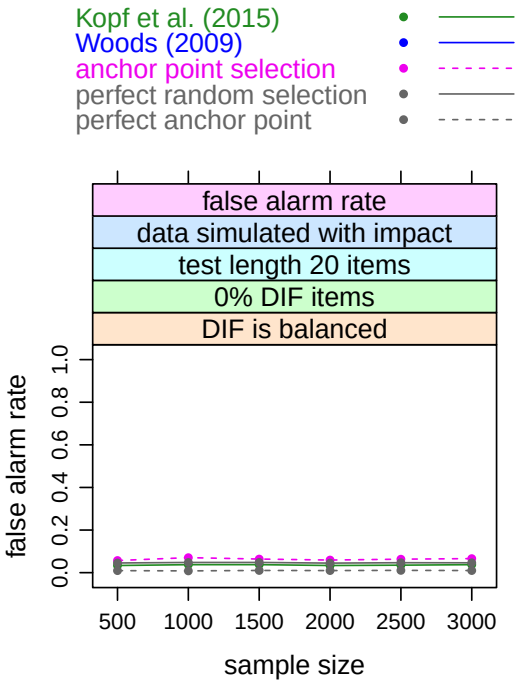

Figure 14: False alarm rate and hit rate for the simulation setting with a test length of 20 items and 0% DIF items.

A.2. Test length 60

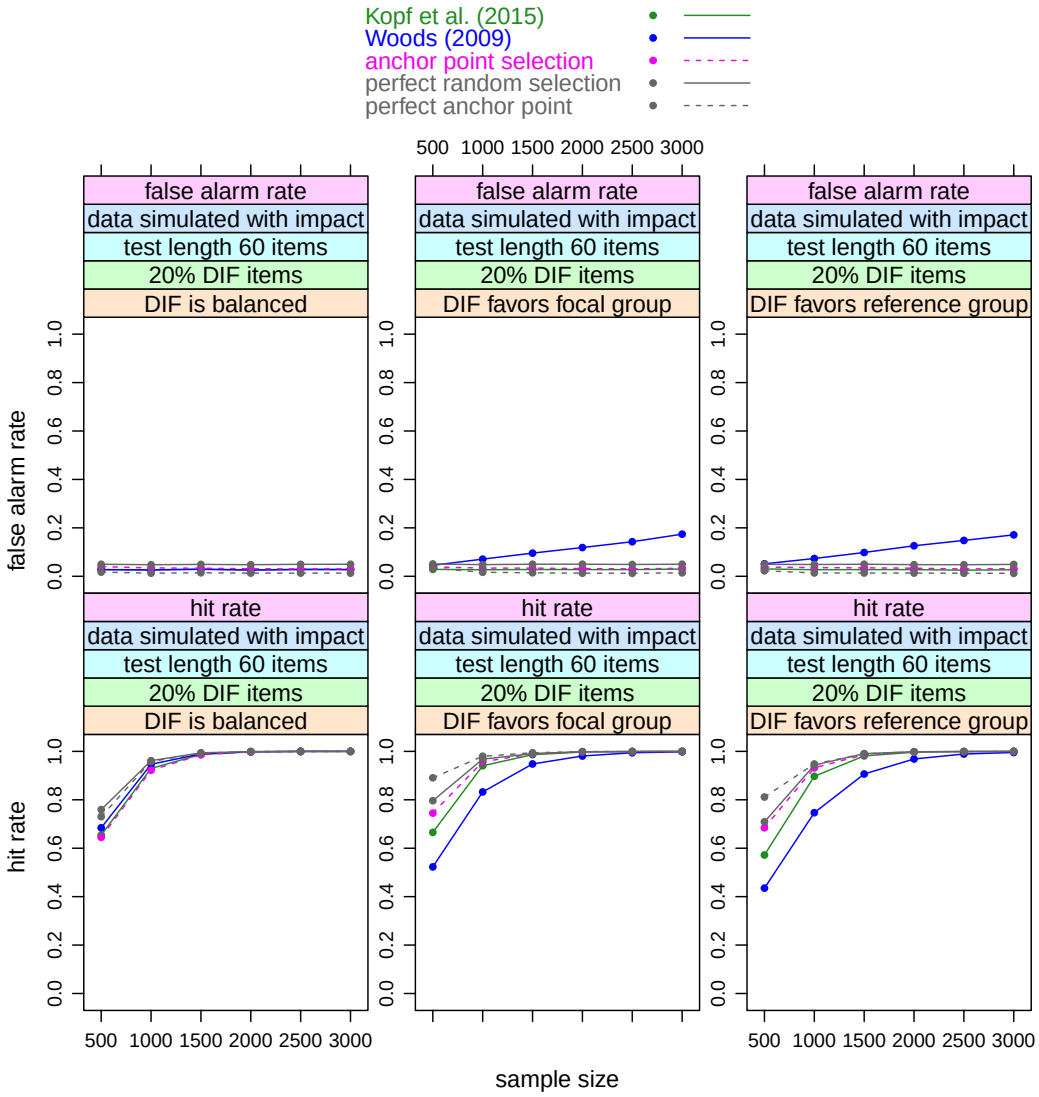

Figure 15: False alarm rate and hit rate for the simulation setting with a test length of 60 items and 20% DIF items.

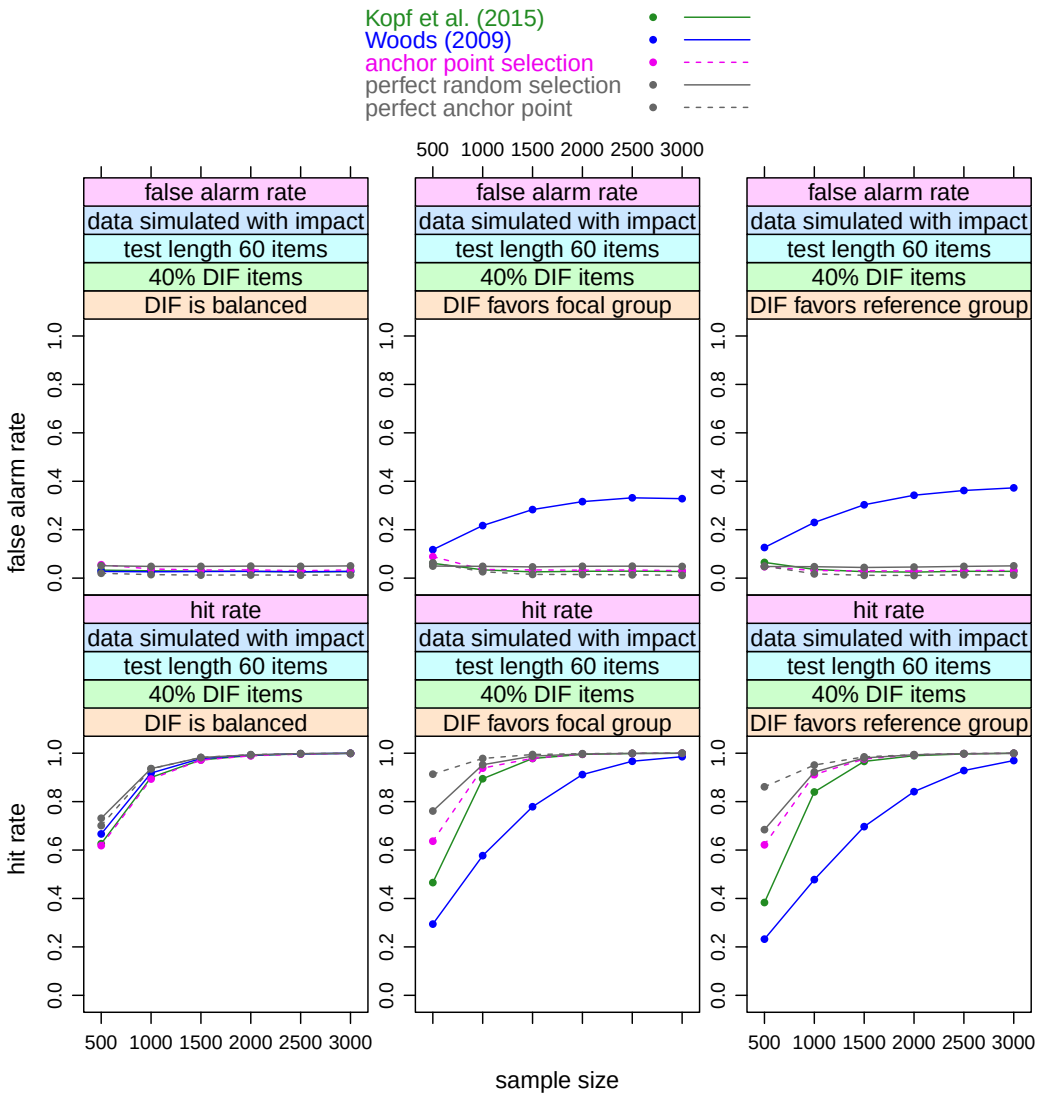

Figure 16: False alarm rate and hit rate for the simulation setting with a test length of 60 items and 40% DIF items.

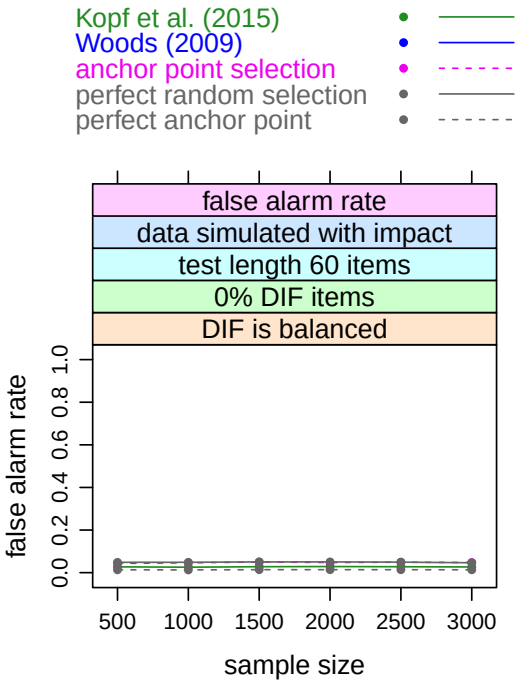

Figure 17: False alarm rate and hit rate for the simulation setting with a test length of 60 items and 0% DIF items.

Affiliation:

Carolin Strobl
Department of Psychology
Universität Zürich
Binzmühlestrasse 14, Box 27
CH-8050 Zürich, Switzerland
E-mail: carolin.strobl@uzh.ch

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://eeecon.uibk.ac.at/~zeileis/

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)