Universität Zürich[UZH]

DIF in multiple group scenarios: Optimization and generalization of anchoring and sampling methods

Thesis (cumulative thesis)
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by

Thorben Huelmann

Accepted in the fall semester 2019
on the recommendation of the doctoral committee composed of
Prof. Dr. Carolin Strobl (main supervisor)
Prof. Dr. Willibald Ruch

Zurich, 2019

DIF in multiple group scenarios: Optimization and generalization of anchoring and sampling methods

Thorben Huelmann

November 20, 2019

Contents

Chapter 1

Introduction

1.1 Scope of this work

In psychological and educational testing Item Response Theory (IRT) models are important tools to analyze test data. Advantages of IRT models, such as the Rasch model, are empirically testable assumptions. One of those assumptions is the assumption of measurement invariance. To check whether the assumption of measurement invariance is justified, items can be tested for Differential Item Functioning (DIF).
The topic of this thesis is on improving hit rates and false alarm rates in multiple group IRT DIF scenarios. This is achieved by investigating and developing anchoring and sampling methods for multiple group DIF scenarios. DIF can be described as a difference in performance on a specific item, that is not explained by the ability the item is intended to measure. For example, a math item that contains a lot of text could induce DIF as examinees need not only math skills to solve the item, but also reading skills. Students that lack these reading skills would be at a disadvantage. The concept of DIF is therefore closely linked to the idea of test fairness. Test fairness, on the other hand, is a requirement for valid comparisons. If, for example, in large scale assessments like PISA, NAEPS or TIMSS items were not controlled for DIF, inferences drawn from comparisons (e.g., between different countries or genders) could be heavily biased. Therefore it is very important to use reliable methods to control for DIF.
One of the biggest problems with the detection of DIF is that the true ability of test takers and the true item parameters are generally unknown. Therefore it is also impossible to detect DIF items without further assumptions. argue that items should not be generally classified as DIF or non-DIF items, but as DIF items relative to other items. But other assumptions are more common.

For example the assumption that the majority of items is DIF free. This assumption is justified by the belief that trained item writers produce more DIF free items than DIF items, but it is generally not empirically testable. However, this assumptions allows researchers to make an absolute classification of items into DIF and non-DIF items. This classification is often the purpose of DIF analysis, as DIF items can then be rewritten or deleted from the test.
Under the assumption that the majority of items is DIF-free, Kopf, Zeileis, and Strobl (2015a) compared several anchoring methods in two group scenarios. Anchoring is the process of aligning the item parameters of different groups. This is a necessity because of the scale indeterminacy of item parameters. It needs to be done with great caution, as an anchor - if chosen poorly - can induce artificial DIF and also result in deflated hit rates. Most anchoring methods rely on multiple testing and successively ranking the items for the anchor. In chapter 2, a direct approach and three aggregation rules are explored to extend various anchoring methods, that have originally been suggested only for the two-group case, to multiple groups. This chapter corresponds to the article *A Comparison of Aggregation Rules for Selecting Anchor Items in Multi-Group DIF Analysis* (Huelmann, Debelak, & Strobl, accepted for publication). In addition to extending anchoring methods to multiple groups, the anchoring method *Next Candidate* is investigated further, as simulations provided interesting insights why this method sometimes fails and how it could be improved. The simulation design was set up to compare different sample sizes and ratios, meaning different distributions of persons to groups. In this article we only investigated hit rates and false alarm rates in relation to the overall sample size, but it became obvious that the group size ratios also have an impact on hit rates and false alarm rates. We investigate this in the third chapter.
The third chapter is based on the manuscript *The effect of different ratios of group sizes in multiple group scenarios on the detection of DIF* (Huelmann, T., Debelak, R. and Strobl, C.). In this chapter, we develop a rule of thumb for practitioners to decide how many persons from each group should be sampled, when there are multiple groups and there is a limit on how many persons can be sampled overall. For example, there could be a financial limit that allows only 1000 test takers overall, coming from 3 different groups. The rule of thumb then answers the question of how many persons should be sampled from the first, the second and the third group. We also show that it is not only particularly hard, but also usually not possible to calculate these ratios with a closed form. Especially for practitioners the rule of thumb derived in this chapter is very valuable. A closed form to optimize these ratios would always rely on the exact knowledge of the DIF effect for every item

and every group. This does not seem to be a realistic expectation. The rule of thumb on the other hand can greatly improve hit rates and false alarm rates with a minimum of information on the DIF effects (in contrast to a naive approach, where all groups are of equal size). The rule of thumb was tested with a perfect anchor selection, meaning four items that are known to be DIF free were chosen as an anchor. In applied research this would not be possible, as it is generally unknown which items are DIF free. However, the focus of this manuscript was on optimal group ratios. We therefore did not use other anchoring methods as they might influence the analysis. But we do not want to underestimate the importance of anchoring. In the second chapter we already showed extensions of anchoring methods for two group scenarios to multiple group scenarios. In the fourth chapter we show how a different approach to anchoring can be extended to multiple groups.
Strobl, von Oertzen, Zeileis, and further Authors (2019) developed a method to align person parameters without specifying a priori a set of items as an anchor. Instead an anchor point is selected by means of optimizing an inequality criterion. In the fourth chapter we show that this idea can be extended to multiple group scenarios. We show that this method clearly outperforms the most common default setting for anchoring in statistical software (equal mean anchoring). Furthermore, this method shows hit rates and false alarm rates comparable to the aggregation rule combined with the anchoring method that showed the best results in the second chapter. Furthermore, with this method search paths can be plotted. These search paths can give useful additional information on the DIF structure and, combined with content knowledge, can even further lower false alarm rates and improve hit rates.

1.2 Contributing manuscripts

The first chapter of this thesis is already accepted for publication. The second and third chapter are based on unpublished manuscripts. These manuscripts were developed in cooperation with coauthors. The titles of the manuscripts are listed below together with a short description of the contributions from all authors.

- Huelmann, T., Debelak, R. and Strobl, C. (accepted): *A Comparison of Aggregation Rules for Selecting Anchor Items in Multi-Group DIF Analysis*. Journal of Educational Measurement.
  Thorben Huelmann studied the literature, suggested the *all rule*, designed and conducted the simulation studies, and drafted the manuscript.

Rudolf Debelak and Carolin Strobl contributed to the conception and presentation of the article.
Chapter 2 is based on this manuscript.
This manuscript addresses the topic of how anchoring methods for differential item functioning (DIF) analysis can be used in multi-group scenarios. The direct approach would be to combine anchoring methods developed for two-group scenarios with multi-group DIF-detection methods. Alternatively, multiple tests could be carried out. The results of these tests need to be aggregated to determine the anchor for the final DIF analysis. In this study, the direct approach and three aggregation rules are investigated. All approaches are combined with a variety of anchoring methods, such as the 'all other purified' and 'mean p-value threshold' methods, in two simulation studies based on the Rasch model. Our results indicate that the direct approach generally does not lead to more accurate or even to inferior results than the aggregation rules. The *min rule* overall shows the best trade-off between low false alarm rate and medium to high hit rate. However, it might be too sensitive when the number of groups is large. In this case, the *all rule* may be a good compromise. We also take a closer look at the anchor selection method 'next candidate', that performed rather poorly, and suggest possible improvements.

- Huelmann, T., Debelak, R. and Strobl, C.: *The effect of different ratios of group sizes in multiple group scenarios on the detection of DIF.*
  Thorben Huelmann studied the literature, investigated the possibility of a closed form solution, suggested the rule of thumb, designed and conducted the simulation studies, and drafted the manuscript. Rudolf Debelak and Carolin Strobl contributed to the conception and presentation of the article.
  Chapter 3 is based on this manuscript.
  The aim of this manuscript is to determine group ratios that are optimal for Differential Item Function (DIF) detection in Rasch models based on the generalized Lord's χ^2 test in multi group scenarios. In this study we first give an introduction into the theoretical deduction of the power in multi group DIF scenarios. We show that it is often not possible to derive the optimal group ratios analytically, because informations that would be necessary in practice, such as the exact DIF pattern, are often not available. Even when these informations are available, computation can be very demanding. We therefore introduce a rule of thumb that was derived by means of a an extensive simulation study. With this rule of thumb, practitioners can easily

determine a group ratio that results in comparably high hit rates in various scenarios.

- Huelmann, T. and Strobl, C.: *An extension of the anchor point selection method to multiple groups.*
  Thorben Huelmann studied the literature, designed and conducted the simulation studies and graphical presentation of the results, and drafted the manuscript. Carolin Strobl contributed to the conception and presentation of the article.
  Chapter 4 is based on this manuscript.
  The aim of this manuscript is to extend the anchor point selection method to multiple group scenarios. Anchor point selection is an aproach for DIF detection in Rasch models based on optimizing an inequality criterion. We show that this method can be easily extended to multiple group scenarios and clearly outperforms equal mean anchoring, which is often the default anchoring technique in DIF software packages. Furthermore we show that the performance of this extension is comparable to otherwise recommended anchoring methods for multiple groups. In addition, we show that through visual inspection of the search path further information can be gained, and if combined with content knowledge can further improve false alarm rates greatly.

Chapter 2

A Comparison of Aggregation Rules for Selecting Anchor Items in Multi-Group DIF Analysis

2.1 Introduction

Differential item functioning (DIF) is defined as a difference in item parameters across different subgroups despite controlling for the underlying ability (Angoff, 1993). This happens, for example, when students with a different first language struggle with mathematics items that contain a lot of text. We assume the underlying ability (mathematics) is equally distributed across students who are native speakers and students with a different first language, but students who are native speakers have an advantage in reading and therefore mathematics items that contain a lot of text are easier for these students than for those with a different first language. As these items are not intended to measure reading skills but to measure mathematical ability, they give an unfair advantage to a group. Therefore the problem of DIF is related to test fairness (e.g. Osterlind & Everson, 2009) and its relationship to multidimensionality has been pointed out by, e.g., Ackerman (1994). It should be noted that there are varying definitions of a fair test and being DIF-free is usually just one aspect of test fairness (Zieky, 2016).
One of the key parts of DIF analysis is the question how to control for the underlying ability. The problem is that we need to measure the ability so we can control for it. The ability on the other hand is measured by the test itself, which is under suspicion of containing DIF, leading to biased ability

estimates (Osterlind & Everson, 2009). One possible solution is to choose a subset of items from the test as an anchor to align the scales for the two groups. For this it is necessary to use an anchoring approach that can identify DIF-free items for the anchor, and many such approaches have been suggested for the two-group case, as will be discussed in detail below. The aim of this paper is now to broaden the scope of anchoring approaches for multi-group cases by means of extending well established anchor selection methods for two-group cases to multi-group cases.
To give an introduction into the necessary background, we will now explain and illustrate - first for the two-group scenario - why selecting anchor items with great diligence is important. Note that the model used in the example is the Rasch model (Rasch, 1960). We also use the Rasch model in the following simulation study. If an item in the anchor shows DIF the analysis can be flawed. In Figure 2.1 10 items are shown with their estimated difficulty for two groups. The first group is called reference group and is indicated by circles. The second group is called focal group and is indicated by red crosses. In this example, the first item was chosen to be the anchor (indicated by the square). Through the anchoring process we define the difficulty of the anchor item to be equal across groups. But it can be seen that the item difficulties for items 2 through 9 also align perfectly, despite being freely estimated. Only item 10 shows a difference in difficulty across the groups. We would now deduce that item 10 is a DIF item and items 1 through 9 are DIF-free. In Figure 2.2 the same data is used as in Figure 2.1. The only difference is that now item 10 was chosen to be the anchor. In this situation we would deduce that only item 10 is DIF-free and all other items are DIF items. From a purely analytical point of view it is not possible to determine which of the two diagrams is showing the correct DIF items.

For practical analysis therefore further assumptions need to be made. One that is often used in practice is to expect the majority of items to be DIF-free (e.g. Angoff, 1993; Bechger & Maris, 2015; Koretz & McCaffrey, 2005; Pohl, Stets, & Carstensen, 2017). This approach follows the logic that the majority of test items work as intended and only few items show DIF. While this is often a plausible assumption due to the high effort invested in the item construction, it cannot be tested empirically whether it is true. By this definition, picking the first item as an anchor would yield the correct response as it is shown in Figure 2.1. Then in the situation depicted in Figure 2.2 we would call the anchor contaminated. This means that the anchor contains DIF items. Contaminated anchors produce what can be called artificial DIF. This means that items that do not have any actual DIF are classified by the DIF analysis as DIF items. Also it can lead to false negatives, with

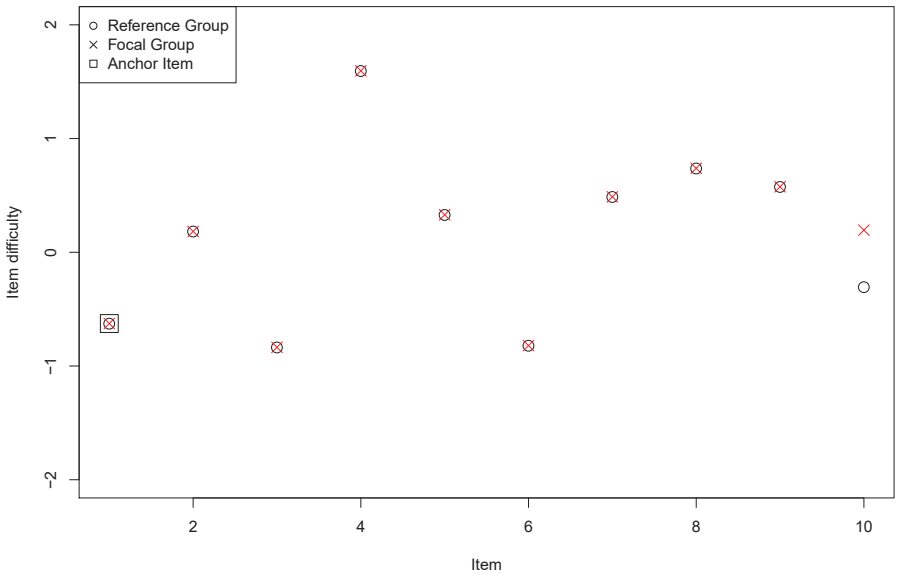

Figure 2.1: Estimated item difficulties for 10 items in two groups

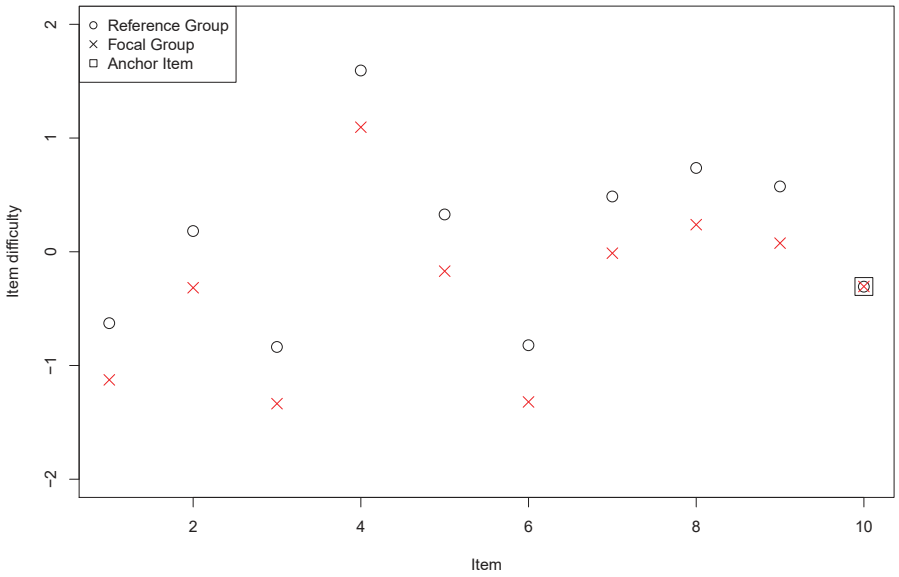

Figure 2.2: Estimated item difficulties for 10 items in two groups

items showing DIF being classified as DIF-free. Both principles can be seen in Figure 2.2. The contaminated anchor induced artificial DIF in the first 9 items while the actual DIF item number 10 was classified as DIF-free.
Unless there is prior knowledge about an item, it is not possible to choose an item for an anchor that is guaranteed to be DIF-free. Anchor selection methods try to statistically determine items that are most likely DIF-free. The most common anchor selection methods will be discussed in detail in the *Anchoring Methods* section. To give a first impression how the discussed anchor selection methods work, we will start with the number of significant tests (NST; Wang, 2004) method here as an example.
Many anchor selection methods work in a similar way: Every single item is tested for DIF, often even multiple times, and a ranking of items is deduced from the resulting test statistics or p-values. The ranking then determines the order in which the items are included in the final anchor for the final DIF test. The NST method for example tests every single item for DIF multiple times with every other item serving as a single anchor for one of these tests. This results in $k - 1$ p-values for every item, with k being the number of items overall.

In the NST method, the number of significant p-values is then counted for every tested item. When the tested item is truly DIF-free, it should ideally return a significant p-value when tested with a true DIF item as a single anchor. If, however, the item is tested with a truly DIF-free item as a single anchor, it is very likely to return a non-significant p-value. We use the term "ideally" here, because this is not a deterministic process but rather the desired outcome. Depending on the power, p-values might not be significant when DIF-free items are tested with DIF items as anchor. Also 5% of the p-values of tests for DIF-free items with DIF-free items as anchor will turn out to be significant as a type I error. But the idea holds that tests with DIF items as an anchor will more likely be significant than tests with DIF-free items as anchor, when the tested item is DIF-free itself. Under the assumption that the majority of items is DIF-free, this should result in a comparably small number of significant tests for DIF-free items. But if a true DIF item is tested, this process should be reversed: testing with a DIF-free item resulting in significant tests and non-significant tests for DIF items, resulting in a comparably higher number of significant tests. Therefore, after every item is tested with every other item as a single item anchor and the number of significant tests for every item is counted, items with a lower number of significant tests are more likely to be DIF-free and are thus preferable for the final anchor. Therefore the number of significant tests is used for ranking the items in the NST framework. If the aim is, e.g., to select the four best

anchor items as the final anchor, those with the lowest numbers of significant tests will be chosen first.

Up until now we only discussed DIF in a two-group scenario, where there is one reference group and one focal group. In reality often multiple group scenarios occur. For example when different nationalities as language groups are to be compared. There are several methods for DIF analysis that can deal with multiple groups and are used in practical research. But there is very little literature about anchoring in multiple group scenarios. In empirical research, the exact anchoring method is often not described and we assume that the software specific defaults are often used for anchoring. A common default is the equal mean anchoring class (for example in the difR package, see Magis, Béland, Tuerlinckx, & De Boeck, 2010). A detailed explanation of anchor classes follows in the *Anchoring Methods* section. The equal mean anchoring class works slightly different than the anchoring processes we explained so far. Instead of picking items that seem to be DIF-free, it is assumed that the test as a whole is unbiased and DIF effects cancel each other out. Under this assumption, the mean test result can be used to match persons across groups. The problem with this anchoring class is that this assumption is rather strong and in practice presumably only rarely met. Therefore the equal mean class should only be used when strong prior knowledge about the test is available. In many cases, other anchoring methods seem more advisable.

The aim of this paper is to broaden the scope of anchoring approaches that can be used in multi-group cases by means of extending anchor selection methods that are already well established for two-group cases to multi-group cases. To our knowledge this has not been done before. In the following, we will first focus on a scenario with three groups (one reference and two focal groups) for simplicity of the explanation. Later in the simulation study we will also cover a setting with more groups.

The most straightforward idea of extending anchoring approaches designed for two groups to multiple group settings would be to try to use the anchor selection methods in the same way they are used in the two-group scenario. This can be done by using a DIF detection method that gives the same number of test statistics or p-values as in the two-group scenario. This can be accomplished by using a DIF method that is an extension for multi-group cases. For example if the Wald Test is used in a two-group case for DIF detection, the generalized Lord's χ^2 test is an extension for this test to multi-group scenarios (Kim, Cohen, & Park, 1995). It has the same number of test statistics, as it will give for every item tested exactly one test statistic. With this approach, classic anchor selection methods can be applied without any need for aggregation over the different group comparisons. We call this the *direct*

approach. The *direct approach* has some limitations. The generalized Lord's χ^2 test tries to identify the items that might show DIF, but aggregates over all group comparisons. Therefore only one test statistic is produced for every item. If also the groups which show biased estimates should be identified, pairwise comparisons need to be done. This would lead to as many test statistics per item as there are pairwise comparisons. Now we receive one p-value for every possible anchor item as well as for all considered comparisons between two groups. These multiple p-values resulting from the considered pairwise comparisons need some form of aggregation before the classic anchor selection methods, like NST, can be used. These forms of aggregation are called aggregation rules in this article and are explained in detail in the *Anchoring Methods* section.

In this article the *direct approach* and three aggregation rules are compared in two simulation studies. The theoretical framework for this is given in the section *Anchoring Methods* and in the section *DIF Detection Methods*. The simulation design is described in the *Methods* section. The results of these simulations are presented in the *Results* section. In the *Discussion* section the results are discussed and recommendations for practitioners trying to implement an anchor selection method in their multi-group DIF analysis are given.

2.1.1 Anchoring Methods

Kopf et al. (2015a) introduced a taxonomy of anchoring approaches that divide the anchoring method into two parts: the anchor selection method and the anchor class. The anchor class defines of how many items the anchor should consist, while the anchor selection method defines how exactly the items are chosen for the anchor.

Anchor Class

Kopf et al. (2015a) differentiated between the all-other (e.g. Cohen, Kim, & Wollack, 1996), equal-mean (e.g. Wang, 2004), iterative forward (Kopf et al., 2015a), iterative backward (e.g. Candell & Drasgow, 1988; Drasgow, 1987; Hidalgo-Montesinos & Lopez-Pina, 2002), and constant anchor class (e.g. Shih & Wang, 2009; Wang, 2004). In practice the most common anchor classes are the all-other class and the equal-mean class (e.g. Berberoglu, 1995; Koretz & McCaffrey, 2005; Stubbe, 2011; Takala & Kaftandjieva, 2000). Both classes do not rely on a anchor selection, as the anchor is either set to all other items (all-other class), or the groups are matched according to

their mean score (equal-mean class). As we focus on varying anchor selection methods these classes were not included in the simulation study. The iterative forward and backward classes build an anchor in a step by step approach. Both rely on a stopping criterion. If the stopping criterion is a fixed number of items, for example four, the iterative forward class is identical to the constant anchor class. The iterative backward class cannot be used with every anchor selection method, as it starts with every item as an anchor. Therefore we do not consider these anchor classesin this study, but employ the constant anchor class.
The constant anchor class describes anchors with a fixed number of items in them, for example 4 items. Constant anchors can be used with every anchor selection method described in this article and can yield, depending on the anchor selection method, uncontaminated anchors. In the literature there is some debate about how many items should be used as an anchor. Shih and Wang (2009) and Wang (2004) recommend four items, Meade and Wright (2012) five items and Woods (2009) recommends between 10 and 20 percent of all items. All of these recommendations try to consider that the shorter an anchor is, the chance of false negatives rises. On the other hand the longer an anchor is, the chance for anchor contamination becomes bigger and a contaminated anchor can lead to false positives. In this study we used 4 items as the anchor length.

Anchor Selection Method

The anchor selection method describes how exactly the items for an anchor are chosen. For some anchor classes, an anchor selection method is not necessary. From the anchor classes discussed above, the all-other class, the equal-mean class, and the iterative backwards class do not rely on an anchor selection method, as they all start with an anchor consisting of every (other) item. The constant class, however, that will be used in the simulation studies, can be combined with a variety of selection methods.

Non-statistical: In certain situations it may be possible that for specific items it is already known that they are DIF-free. This could be the case for example if items are taken from different tests, some of which were already tested for DIF. Another possibility is to rely on expert knowledge. In these situations, the anchor selection can be conducted by simply relying on the prior knowledge. This non-statistical anchor selection will not be investigated further in this study, as it is assumed that there is no prior knowledge about the items containing DIF.

All-other (AO, Woods, 2009): The all-other selection works similar to the all-other anchor class. Every item is tested for DIF with all other items as an anchor. The items are then ranked according to their test statistic.

All-other-purified (AOP, Wang, Shih, & Sun, 2012): The all-other-purified selection starts like the all-other selection. After a first analysis the anchor is purified from items showing DIF. These steps are repeated until no item in the anchor shows signs of DIF.

Number-of-significant-tests (NST, Wang, 2004): The number-of-significant-tests selection in the two-group scenario starts with testing every item for DIF with every other item as a constant single anchor, resulting in k-1 test statistics for every item, where k is the number of items. For every item the number of significant tests is counted. The items are then ranked by this number.

Next-candidate (NC, Wang, 2004): The next-candidate selection picks, as an initial step, the first item of the anchor the same way it is done in the number-of-significant-tests selection. In the second step the next item is chosen with a DIF analysis taking the already picked item as an anchor. The item chosen is always the item with the lowest test statistic. The second step is repeated until the anchor has the desired length or other user specified criteria are met.

Mean-p-value (MP, Kopf, Zeileis, & Strobl, 2015b): For the mean-p-value-threshold selection every item is tested for DIF with every other possible single-item-anchor. For each item the mean p-value is computed and items are ranked according to their mean p-value.

Mean-test-statistic (MT, Shih & Wang, 2009): The mean-test-statistic selection works similar to the mean-p-value selection, except the items are not ranked by their mean p-value, but by their mean test statistic.

Mean-p-value-threshold (MPT, Kopf et al., 2015b): The mean-p-value-threshold selection starts like the mean-p-value selection. After the items are ranked the $\lceil k \cdot 0.5 \rceil$th p-value is chosen as threshold, where k stands for the number of items and $\lceil\rceil$ depicts the ceiling function. For every item the number of tests that exceed this threshold is counted. The items are ranked by this number.

Mean-test-statistic-threshold (MTT, Kopf et al., 2015b): The mean-test-statistic-threshold selection works just like the mean-p-value-threshold

selection, except that all calculations are done with the test statistic and not the p-value.

2.1.2 DIF Detection Methods

There are numerous methods for detecting DIF. The focus of this study is to define aggregation rules and give advice to practitioners which aggregation rule to use. The focus is therefore not on comparing different methods for DIF detection. A brief investigation was done with varying methods (Mantel-Haenszel and Logistic Regression, see e.g. Osterlind & Everson, 2009) to test for DIF during and after the final anchor has been chosen. For brevity, details are omitted.
The data suggested that there are no particular interaction effects between the detection method and the aggregation rule. This means that there is no evidence that one DIF detection method is to be preferred from the anchoring point of view. Advantages and disadvantages of DIF detection methods seem to be stable across the aggregation rules. Therefore only one detection method was used for the studies presented here in detail, because it allows us to combine the same method both with the *direct approach* and with the investigated aggregation rules: Lord's χ^2 uses a Wald-Type test to determine DIF. Kim et al. (1995) showed a generalization of Lord's χ^2 method for DIF-testing to work with multiple groups. In this study Lord's χ^2 was used for comparison of two groups and the generalization as introduced by Kim et al. (1995) was used for multi-group comparisons. Note that the generalization of Lord's χ^2 in its original implementation by Kim et al. (1995) does not compare focal groups to each other. Only the reference group is compared to every focal group. In the Appendix A.2 we also show why the generalized Lord's χ^2 test cannot compare the focal groups. Note, however, that other widely used DIF detection methods would not compare each focal group to each other either.

2.1.3 Aggregation Rules

The most straightforward way of dealing with multiple groups in an anchoring problem is to use DIF tests aggregating over multiple groups such as the generalized Lord's χ^2 test. Then the same anchor selection methods can be used as in the two-group scenario without any aggregation. However, as outlined above, this *direct approach* does not take into account comparisons between focal groups. An idea to solve this problem would be to conduct multiple DIF tests with pairwise comparisons. When this approach is used,

every item is tested with every other item as a single item anchor and also with every other group as a focal group, leading to $(k-1) \cdot g$ test statistics, with k being the number of items and g the number of focal groups. The described anchor selection methods can then no longer be used directly, as there is not the usual $k-1$ test statistics, but a matrix of $(k-1) \cdot g$ test statistics. In order to be able to use the full portfolio of anchor selection methods that are available for the two-group case, some kind of aggregation rule is needed. Kopf (2013) suggested the following possible aggregation rules:

Min rule: The first mentioned rule is to take the minimum of every row. We call this the *min rule*. This rule is similar to the minmax-strategy from decision theory. The minmax-strategy has the known issue of being pessimistic (Savage, 1951). Whether this is a problem when it is used as an aggregation rule remains to be inspected.

Mean rule: The second mentioned rule is to take the mean of the rows (the *mean rule*). This approach is supposed to be less pessimistic than the *min rule*, but it is unknown in how far outliers will influence the outcome. The *min* and the *mean rule* are both visualized in Table 2.1.

All rule: In addition to the *min* and *mean rule* suggested by Kopf (2013), we suggest and investigate the *all rule*. The *all rule* ignores the matrix structure and treats the matrix just like a vector. This is visualized in Table 2.2. Note that the reported p-values are the same as in Table 2.1, but are organized in a vector instead of a matrix. In some cases this aggregation rule will deliver the same results as the *mean rule* (if the anchor selection is based upon a mean, it does not matter, whether first the mean was taken row wise and then again, or if the overall mean was used from the beginning). In other cases this aggregation rule should be less prone to outliers than the *mean rule* and should be less pessimistic than the *min rule*.

Weighted mean rule: Kopf (2013) also mentioned a third rule, which is not considered in this study. It is based on a weighted mean. For example the values could be weighted with the group size. In this first study, all focal groups have the same size. Therefore the weighted mean would be the same as a simple mean, as long as the reference group also has the same size as the focal groups and would only give comparisons with the reference group a higher value when the reference group is bigger than the single focal groups. Therefore it is not considered in this first study of aggregation rules.

Table 2.1: Visualisation of the min and mean rule. The table shows p-values for testing item 1 for DIF with each of the other three items of a four item test serving once as the anchor item for every comparison.

	Ref:Foc1	Ref:Foc2	mean rule		min rule	
Item	p-value	p-value	mean p-value	significant	min. p-value	significant
2	0.700	0.020	0.360	0	0.020	1
3	0.002	0.040	0.021	1	0.002	1
4	0.500	0.300	0.400	0	0.300	0
sum				1		2

In Table 2.1 the *min* and *mean rule* are shown combined with the number-of-significant-tests anchor selection method. In this example an item is analyzed from a four item test. There is one reference group and two focal groups. We can now calculate that there are $(k-1) \cdot g = 3 \cdot 2 = 6$ p-values for this item. These p-values are shown in the left part of Table 2.1. In the middle and the right part of the table, the mean and minimum p-values across the focal groups for the three items used as anchor items are shown. It can be seen that the NST calculated with the *mean rule* (1) is lower than the NST calculated with the *min rule* (2). This is due to the fact that the minimum will alway be less than or equal to the mean. Therefore there are situations where the minimum is significant but the mean is not. The *min rule* can therefore differentiate between items that show a significant p-value in at least one comparison and items that do not show a significant p-value in any comparison. This is still a pessimistic approach in the sense that all items showing DIF are equal, with no regard to how many comparisons showed significant DIF. In comparison, the *all rule* is less pessimistic in the sense that it differentiates between items showing different degrees of DIF. In Table 2.2 the *all rule* is visualized. The NST calculated with the *all rule* (3) is higher than the NST calculated with the *min rule* in this example. This is due to the fact, that the *min rule* does not take into respect how many pairwise comparisons are made. When calculated with the *min rule* or the *mean rule*, the maximum NST is $k-1$ (with k being the number of items), while the maximum NST calculated with the *all rule* is $k-1$ times the number of pairwise comparisons.

In contrast to the *direct approach*, all aggregation rules could easily implement comparisons between focal groups. In order to keep the analysis of the

Table 2.2: Visualisation of the all rule. The table shows p-values for testing item 1 for DIF with each of the other three items of a four item test serving once as the anchor item for every comparison.

Comparison	Item	p-value	significant
	2	0.700	0
Ref:Foc1	3	0.002	1
	4	0.500	0
	2	0.020	1
Ref:Foc2	3	0.040	1
	4	0.300	0
sum			3

direct approach and the aggregation rules comparable, however, these comparisons were not done with the aggregation rules either in this first study. The aim of this paper is to find out whether there is an effect of the aggregation rule on the DIF analysis and furthermore give recommendations to practitioners. Therefore two simulation studies were conducted.

2.2 Method

In the simulation study the following factors were varied: The number of groups, the overall sample size, the group sizes, the ratio of reference group size to focal group size, the anchor selection method, the direction of DIF, and the aggregation rule. The number of groups was either three or six and the group sizes range from 500 to 3600. The number of groups, the sample sizes and the group sizes could not be varied independently from each other. The sample sizes were held constant across the different numbers of groups and ratios of focal to reference group sizes. This study design was chosen to ensure the same overall sample sizes across the three group and six group scenario because we expected the overall sample size would be the most influential factor. Therefore the ratios could not be kept equal between the three and six group scenarios but we tried to keep them comparable. As the anchor class, a constant anchor of four items was used.
For pairwise comparisons within the anchor selection step Wald Tests were used. The generalized Lord's χ^2 test (Kim et al., 1995) was used as the final DIF test, employing the previously identified anchor. The *direct approach* and the three aggregation rules introduced before (*min rule*, *mean rule* and

all rule) were used in every simulation.
Data was simulated using the Rasch model (Rasch, 1960). The Rasch model was chosen because it is widely known and has easy to test properties for DIF detection (Fischer & Molenaar, 1995). Every respondents ability was simulated as standard normally distributed N(0,1). 40 items were simulated. The simulated item difficulties of the items were taken from Wang et al. (2012). The percentage of DIF items was set to 45 percent, resulting in 18 items showing DIF. The items showing DIF were chosen randomly. This was done to prevent any effects that the difficulty of the item might have on the DIF detection. DIF was induced by adding 0.6 to the difficulty of the randomly chosen DIF items. The value 0.6 was chosen because according to Chang, Huang, and Tsai (2015) it resembles a moderate to high amount of DIF. In a baseline condition, no DIF at all was simulated. In the other two conditions the direction of DIF was varied. In the first of these two conditions all DIF items put all focal groups at the same disadvantage. In the second of these two condition the direction of DIF could vary between the focal groups, resulting in items favoring some groups and putting others at a disadvantage. This condition is not to be confused with a completely balanced design, where DIF effects cancel each other out across items as well as across groups. In the 3 group scenario only for 6 of the 18 DIF items the effects cancel each other out across the groups and only for one reference group the DIF effects across items cancel each other out. This design was chosen as it resembles a more realistic scenario, then the scenario where the reference group is always at an advantage. A completely balanced design was not included because it would only make estimation easier, as DIF effects cancel each other out. Furthermore it is a quite unrealistic assumption that effects cancel each other out.

2.2.1 Three Group Scenario

In the first study there was one reference group and two focal groups. The reference group was always chosen to be the same size or bigger then the focal groups, while the reference groups were of equal size. In the three group scenario 1200, 600 and 300 participants were simulated for the reference group. The focal groups were simulated representing different ratios between the reference group and the focal groups. In the 1:1 ratio, focal groups were the same size as the reference group. Because there were two focal groups, the number of persons not in the reference group was twice the number of persons in the reference group for this ratio. In the 1:2 ratio the reference group was twice as big as each focal group and in the 1:3 ratio the reference group was three times as big as each focal group, resulting in an equal amount

of persons in and out of the reference group (1:2 ratio) or of more persons in the reference group than in the other groups combined (1:3 ratio).

2.2.2 Six Group Scenario

The second simulation used 6 groups in total (1 reference group, 5 focal groups). The overall sample sizes were kept just like in the first simulation. As it is not possible to keep both the overall sample size and the ratios the same in the two simulations, the ratios were allowed to change in the six group scenario. The ratios used in the six group scenario were 1:1, 1:1.67 and 1:2.14. These ratios were chosen to keep ratios similar to the ratios used in the three group scenario, while assuring the same overall sample size.

2.2.3 Computational Details

All simulations were done with 500 iterations and carried out using the R software version 3.5.1 (R Core Team, 2019). For data generation the eRm package (Mair & Hatzinger, 2007) and for DIF analysis the difR package (Magis et al., 2010) was used. Anchor selection methods were taken and adapted from the psychotools package (Zeileis, Strobl, Wickelmaier, Komboz, & Kopf, 2018).

2.3 Results

For every condition the hit rates, meaning the percentage of correctly identified DIF items, and the false alarm rates, meaning the percentage of items being flagged as DIF without beeing simulated as a DIF item, were reported. The results were similar throughout all simulation conditions. In some cases the *mean* and the *all rule* led to identical results. In the following figures the *mean rule* overprints the *all rule* in these cases. This is why in some figures the line for the *all rule* is not visible. The figures also include a *perfect* selection as a baseline condition. *Perfect* selection in this case means that the anchor items were chosen randomly from the pool of items simulated to be DIF-free.

2.3.1 "No DIF" Scenario

In Figure 2.3, false alarm rates are shown for the three group scenario without any DIF. For an optimal result, the false alarm rates should not deviate substantially from the nominal α-level of 5%. An observed false alarm rate above 5 % suggests an inflated type I error rate, while a false alarm rate below 5% would result in a too conservative test.
All aggregation rules approximately meet the α-level of 5% ($\pm$ 1-2%) in most conditions of this scenario. Only in combination with the NC method it is notable that the *min* and *mean rules* show an inflated type I error rate of 11.7% and the *all rule* of 7.0%.

2.3.2 "DIF Always Favors Reference Group" Scenario

In Figure 2.4, false alarm rates for the three group scenario with all DIF items putting the focal groups at a disadvantage are shown. From this scenario, the *min rule* seems to be the most favorable overall. Although the *min rule* shows α inflations in all conditions in this scenario, in most conditions these inflations are only small or at least converge towards 5%. Only combined with the AO or the AOP anchor selection method, the false alarm rates seem to rise with the sample size. The *all rule* and the *mean rule* also show rising false alarm rates in these conditions, but their false alarm rates are consistently higher than the false alarm rates for the *min rule*. The *all rule* and the *mean rule* are very similar, not only in these two condition, but in all conditions in this scenario. In comparison with the *min rule* they show similar or slightly less inflated α rates in combination with the MP, MPT, MT and MTT anchor selection, but substantially higher false alarm rates in combination with NST, NC, AO and AOP anchor selection. The *direct approach* shows similar results to the *min rule* when it is combined with the MP, MPT, NST, and even slightly less inflated false alarm rates in combination with the NC and AOP anchor selection. However, the false alarm rates for the *direct approach* in combination with the MT and MTT anchor selection is highly inflated and seems to be rising with the sample size.
The *direct approach* as well as all three aggregation rules show inversely u-shaped false alarm rates in combination with the NST and NC anchor selection. A similar pattern was already reported by Kopf et al. (2015a) for two-group cases.

In Figure 2.5, hit rates for the three group scenario with all DIF items putting the focal groups at a disadvantage are shown. It can be seen that ev-

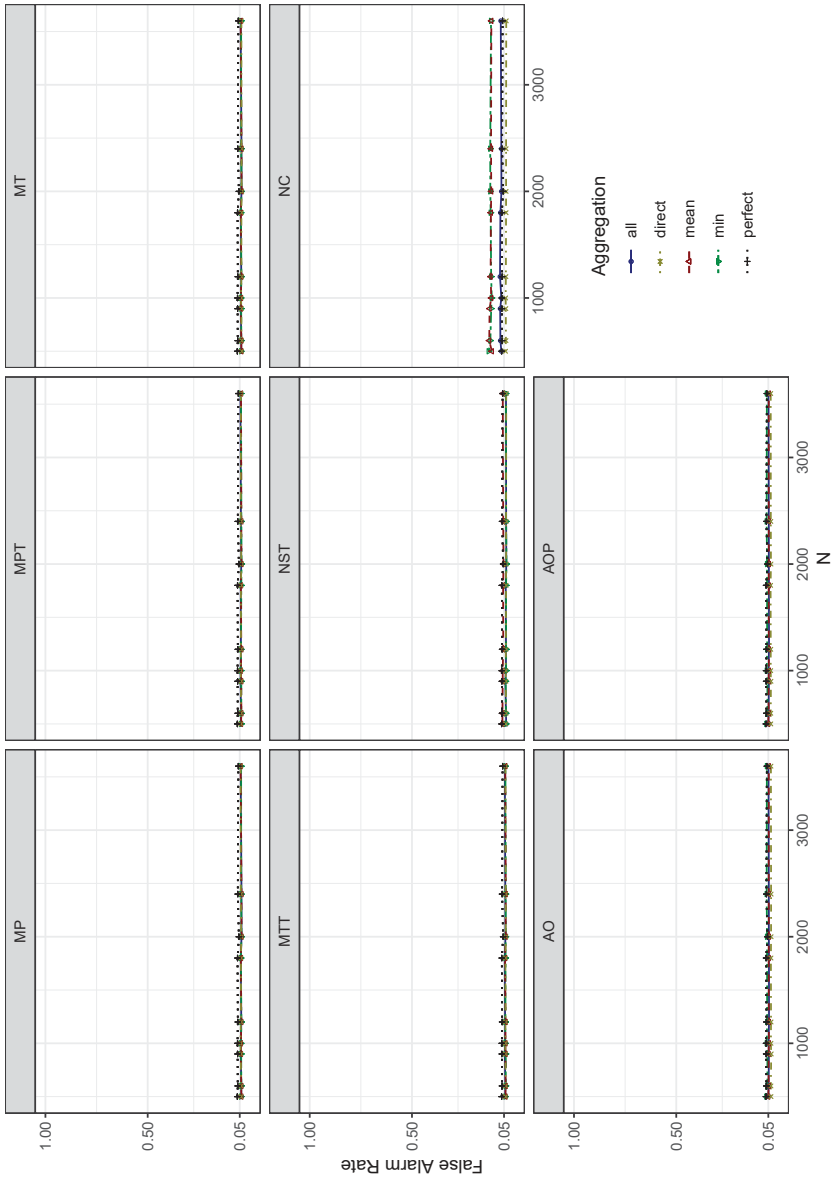

Figure 2.3: False alarm rates in the three group scenario without any DIF

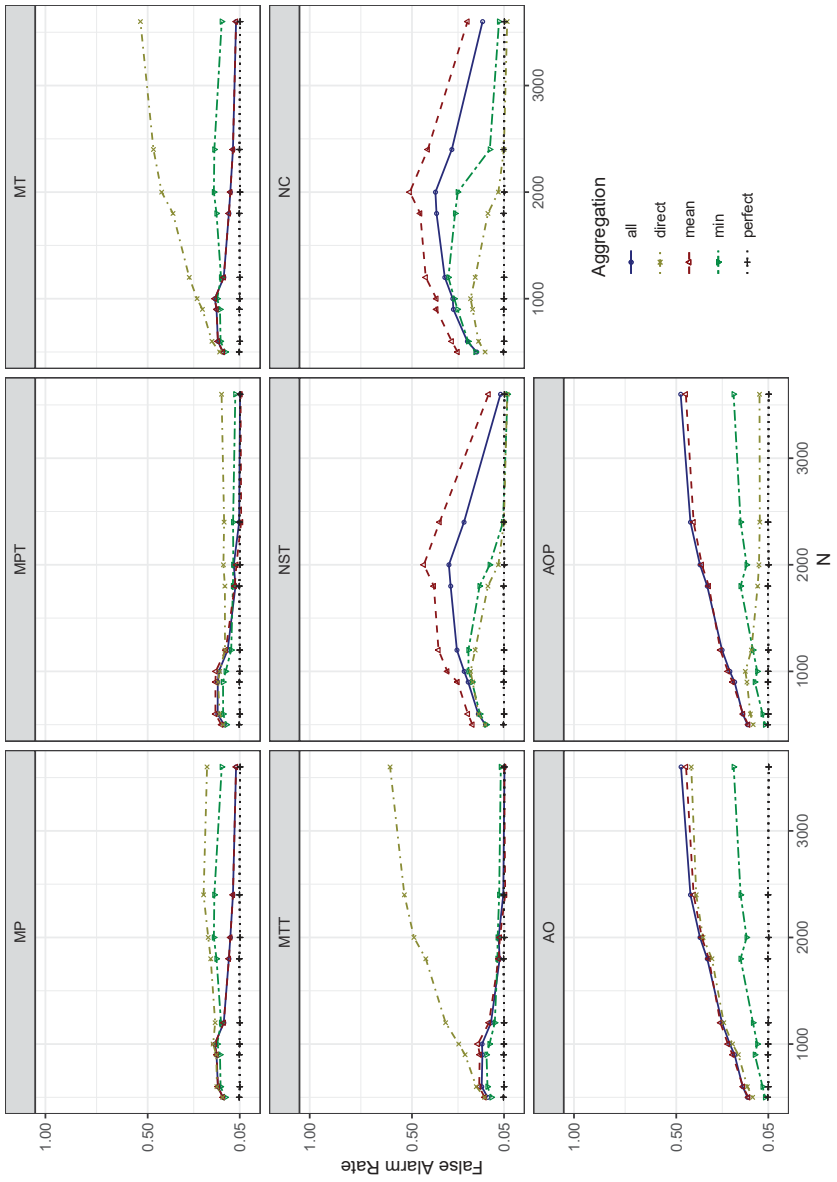

Figure 2.4: False alarm rates in the three group scenario with all DIF items favoring the reference group

ery aggregation rule under every anchor selection method seems to converge towards hit rates close to one as the sample size increases. The rules themselves do not seem to differ strongly in this scenario. The only exceptions are the *min rule* outperforming the other rules with the AO or AOP anchor selection method and the *direct approach* performing substantially worse than other rules in combination with the MT and MTT anchor selection methods.

2.3.3 "DIF Favors Reference Group or Focal Groups" Scenario

In Figure 2.6, false alarm rates for the three group scenario with some DIF items favoring a focal group and others putting all focal groups at a disadvantage are shown. The *min* and *all rule* perform similarly, except for AO and AOP. In combination with AO and AOP the *all rule* shows highly inflated false alarm rates. Both, the *min* and the *all rule*, show again the inverse u-shape with the NST anchor selection method and the NC anchor selection method. The *min rule* shows false alarm rates that are very close to 5% in nearly all conditions and outperforms the other rules. The only exception in this scenario happens in combination with the NC anchor selection. In this condition the *min rule* shows an inflated false alarm rate and is outperformed by the *direct approach*. The *mean rule* does not perform well under some conditions. Especially under the NST and NC selection, where false alarm rates increase with sample size and reach 100%. The *direct approach* also shows high false alarm rates that seem to rise with the total sample size under MT and MTT.

In Figure 2.7, hit rates for the three group scenario with some DIF items favoring a focal group and others putting all focal groups at a disadvantage are shown. The patterns are similar to what was seen in Figure 2.5, with the *min rule* outperforming the other rules in some combinations. Only the difference between the *min* and the *all rule* becomes less substantial, while the *mean rule* performs particularly badly with NST and NC. The results for the *direct approach* are also similar to what was already seen in Figure 2.5. Again, the *direct approach* performs similarly to the other rules except for the MT and MTT selection where it performs worse than the three aggregation rules.

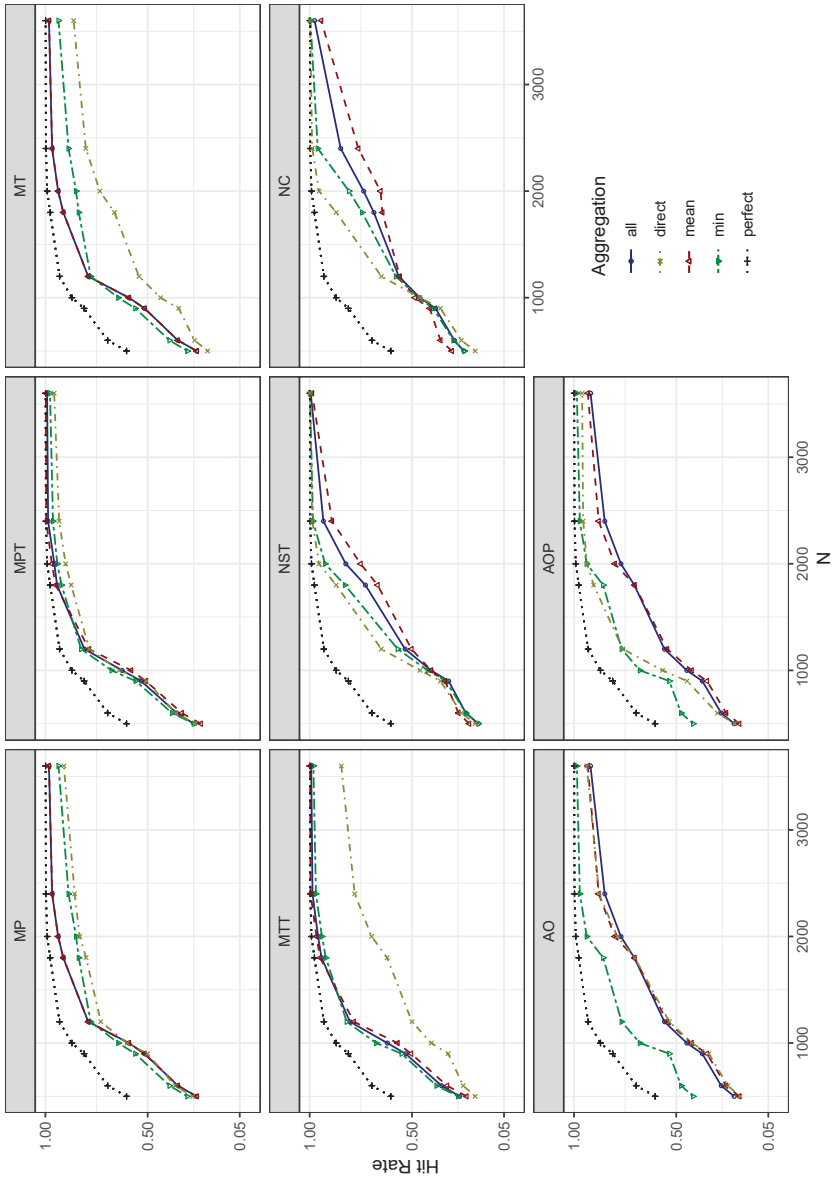

Figure 2.5: Hit rates in the three group scenario with all DIF items favoring the reference group

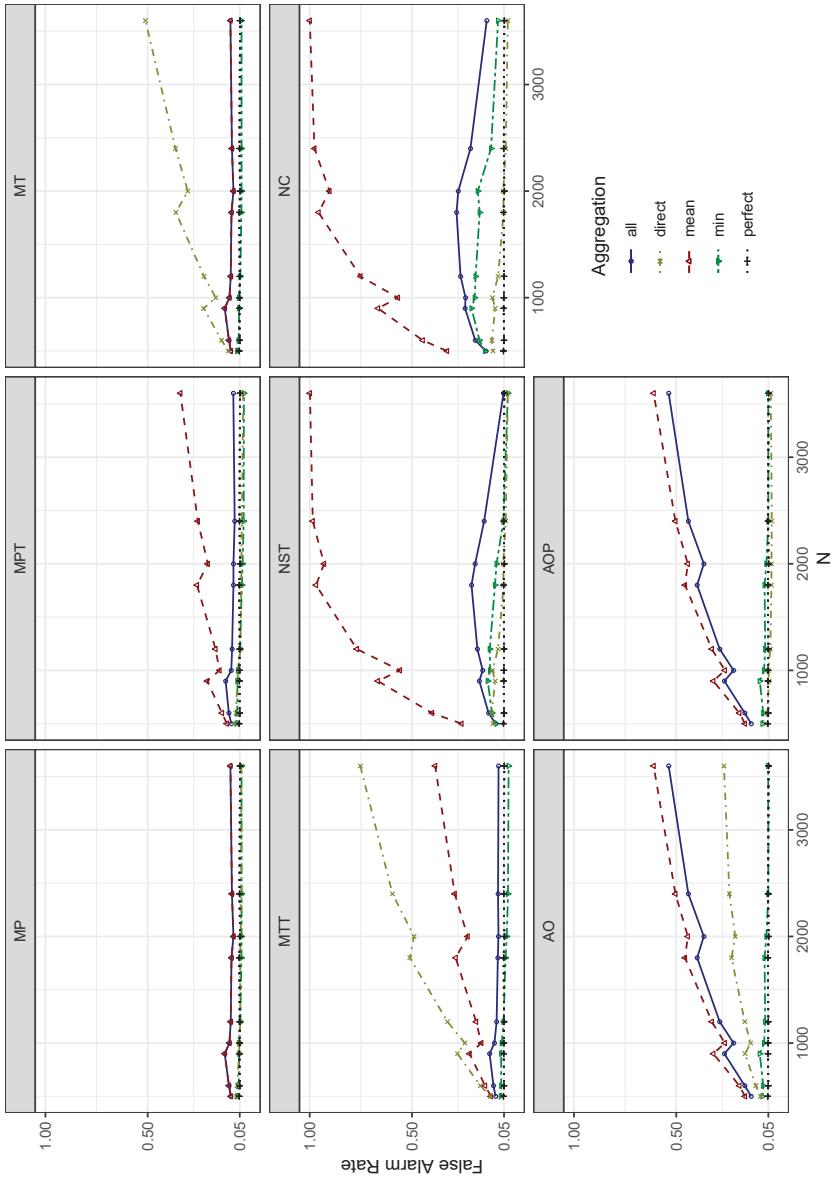

Figure 2.6: False alarm rates in the three group scenario with half of the DIF items favoring the reference group

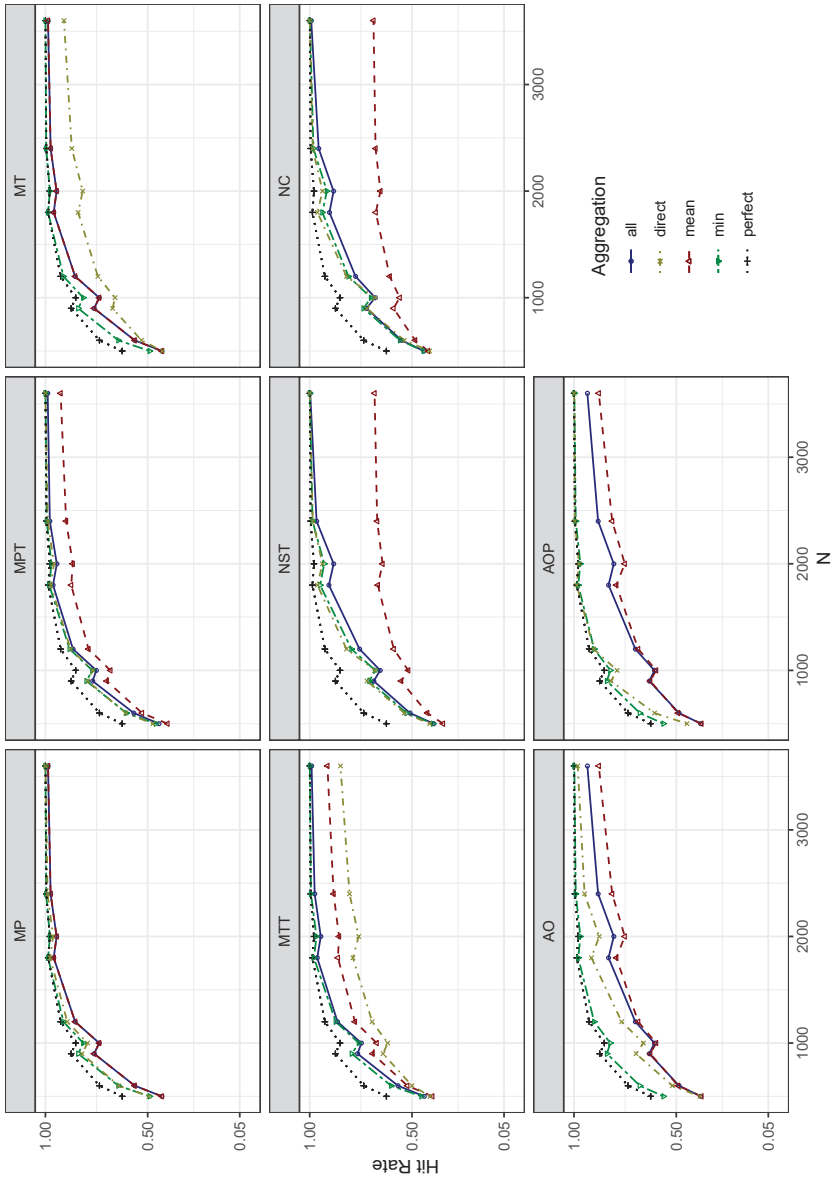

Figure 2.7: Hit rates in the three group scenario with half of the DIF items favoring the reference group

2.3.4 False Alarm Rates and Hit Rates in the Six Group Scenario

False alarm rates and hit rates in the six group case were similar to the three group case. Overall the insights gained from the three group case also remain valid in the six group case. For brevity, the plots A.1 to A.5 are therefore displayed in an online appendix. The main difference in results is that in the six group scenario false alarm rates and hit rates converge slower, meaning bigger sample sizes are necessary to achieve the same power.

2.3.5 Further Exploration of the Next Candidate Method

The NC anchor selection was performing particularly poorly in the presented results. Similar findings were shown in Kopf et al. (2015a). We decided to take a closer look at this anchor selection method, as the idea seems promising but fails to deliver good results. We therefore developed an additional method to assess the quality of an anchor selection method and used this method on the already simulated data. The best anchor selection method is always a perfect selection of every DIF-free item. This is an upper bound of quality, although it is rarely reached. On the other hand, the worst anchor selection should be a random choice between all items, including those containing DIF. Any method for anchor selection should at least outperform picking items at random. Therefore picking items completely at random is labeled as worst case here. If items are picked at random until the first DIF item is picked, the probability for this happening on the first pick, the second pick and so forth can be described by the means of a geometric distribution. For the geometric distribution, an expectancy value can be calculated. Also for the tested anchor selection methods and aggregation rules similar values can be calculated. The simulations yielded an order in which items should be picked for the anchor. With these orderings it can be estimated how likely it is for a certain anchor selection method and aggregation rule that the first, second, or up to the twenty-third item in this setup was the first DIF item. These calculated values can then be easily interpreted. For example in the worst case of picking randomly, the expectancy value is 2.16 in the setting of the three group scenario. If a 4 item anchor with randomly picked items is chosen, it would be very likely to have at least one DIF item. So in general if a fixed number of items is chosen to be the anchor, this number should be well below the calculated value as this value marks the point were it becomes more likely to include at least one DIF item in the anchor. We do not have the room to discuss here all the combinations of sample size, number of

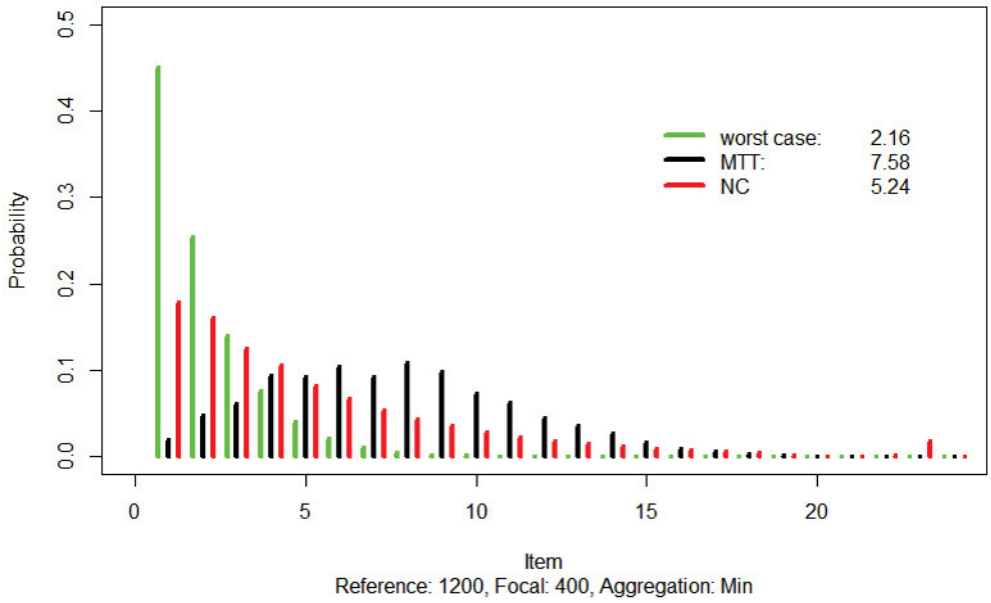

Figure 2.8: Probability of picking the first DIF item

groups, anchor selection and aggregation rule. Therefore only an illustrative example is shown in Figure 2.8.

Picking at random is, as it was designed, the worst case in the illustration. The probability of picking at least one DIF item within the first four items is about 95%, so it is nearly guaranteed to pick a DIF item and contaminate the anchor. The expectancy value for the NC is 5.24. This is higher than 4, the anchor length used in this study. But it should be noted again that these are expectancy values and not hard cut-off points. Being close to this value means there can still be a rather high chance for anchor contamination. The MTT selection performed the best in this set up. The expectancy value is 7.58. In this setup it seems rather safe to chose an anchor of length 4.
These reported differences were also expected to be seen. Picking items at random was designed to be the worst case and it did perform the worst. In the simulations also MTT outperforms NC and this can also be seen here. But this analysis can help to find reasons why the NC does not perform well and also directs further research to improve the NC anchor selection. In Figure 2.8 it can be seen that for a DIF item to be picked up as a first item, the probability is rather high. The NC anchor selection is an iterative

approach, which means it is particular important to get the first item right, or otherwise highly contaminated anchors can be yielded. On the other hand in Figure 2.8 there is a peak at the twenty-third item for the NC anchor selection. This means, the NC anchor selection method substantially more often than other methods picked the twenty two DIF-free items first. This pattern can also be seen in other settings (not shown for brevity). This is especially remarkable as no other selection method yielded comparable peaks at the last items to be picked up. This shows that it has potential to yield longer uncontaminated anchors than the other anchor selection methods, but only if the starting items are well chosen.

2.4 Empirical Application: General Knowledge Quiz

To illustrate further the effect of the different aggregation rules, we include a practical example. We are using the SPISA dataset from the psychotree R package (Trepte & Verbeet, 2010). This dataset contains a subsample from the general knowledge quiz 'Studentenpisa' conducted online by the German weekly news magazine SPIEGEL. It contains the quiz results from 1075 Bavarian university students to 45 questions. The questions can also be found in Appendix A.3. Each question is assigned to one of the following topics: Politics, history, economics, culture or natural sciences. The dataset furthermore contains some sociodemographic data. Here, for illustration, we only use the SPON variable. This variable measures the frequency of students accessing the SPIEGEL online (SPON) magazine. The variable is divided into 7 categories. Some categories are only sparsely populated, which would lead to poor item parameter estimates and test results. Therefore, we recoded the variable into three categories: Accessing SPON never, between once a month and up to three times a week, and more frequently than 3 times a week.
To illustrate the effect of the aggregation rule on the result of the analysis, we now model the responses of the test takers to the 'Studentenpisa' items via the Rasch model and test all items for DIF. Since this analysis is carried out for illustration purposes, we do not check any further assumptions of the model here. We expect that items from the field of politics, economy, and culture might be easier for students who often access SPON, as the magazine generally focuses on news from these topics. We analyze the data with the MPT anchor selection method, that performed well in our simulation studies, in combination with the *direct approach* and the three aggregation rules.

The results show that there is a great overlap in which four items were chosen as anchor items by the different aggregation approaches. For example, all approaches chose items 12 and 35 as an anchor. In Appendix A.3 we summarize which items were picked as anchors (indicated by the letter A in Tables A.1-A.5) by which approach. From a content perspective, none of these anchors can be assumed to be superior to the others, as they all make use of items we would suspect to be DIF free with respect to SPON usage as they do not rely on daily news.
In Appendix A.3 we also summarize which items were identified as DIF items by which approach (indicated by the letter D in Tables A.1-A.5). The *direct approach* and all aggregation rules identified the same 17 items to be DIF items. Furthermore, all approaches except the *mean rule* identified item 8 as a DIF item. The *min rule* identified two additional items as DIF items that no other approach identified. The *mean rule* identified 3 items as DIF items that no other approach identified. It is noteworthy, that the *direct approach* and the *all rule* identified the exact same items as DIF items.
The relatively high number of DIF items is what we were expecting based on the different SPON usage. The *min rule* showed higher hit rates than the other methods in the simulations. Therefore we assume that the method detected more actual DIF items. The *mean rule*, on the other hand, showed higher false alarm rates in the simulations. This could mean that it also flagged some items falsely as DIF items in this example.
For many items identified by all methods there are plausible reasons why they are supposed to be DIF items. For example item 1: 'Who determines the rules of action in German politics according to the constitution?' (the federal chancellor). In September 2009 an election for the German Bundestag was coming up. We therefore assume that also more general facts about the election, like the role of the federal chancellor, would be covered in the daily reporting. It should be mentioned, however, that also some items that are rather hard to explain based on their contents, in particular from the natural science field, were identified by all methods as DIF items.
For the two items identified only by the *min rule* there are also plausible reasons. For example, item 17: 'Which of the following countries is not a member of the EU?' (accompanied by a selection of countries). The correct answer is Croatia. At the time of the quiz, Croatia was in negotiations with the EU and hoped to complete accession talks within 2009. News outlets frequently reported about this fact, this is why frequent readers of SPON could have an advantage in this question.
On the other hand, the three items identified only by the *mean rule* are rather hard to explain. For example item 43: 'Which kind of bird is this' (accompanied by a picture of a black bird). This item is supposed to be

harder for frequent readers of SPON, which is not particularly plausible.

2.5 Discussion

In this study we investigate the important issue of anchoring in multi-group scenarios. Anchoring is an essential part of DIF analysis (e.g. Kopf et al., 2015b; Wang et al., 2012; Woods, 2009) and therefore helps to create unbiased tests, which again are a necessity for fair testing (Ree, 1993). Our simulation study gives valuable advice on how to translate anchor selection methods from two-group scenarios to multi-group scenarios. We based our work on the findings of Kopf (2013) and our study is generally in accordance with these findings. Often a two-group approach is not appropriate, for example, when the grouping variable is nationality or language. Therefore this research gives helpful advice for applied researchers for a variety of plausible multi-group scenarios, as we are not aware of any prior study analyzing aggregation rules.

We examined a *direct approach* and three aggregation rules for anchor selection in multi-group scenarios, namely the *min*, *mean*, and *all rule*. The *mean rule* was clearly performing worst and is generally not advisable. This finding can be explained by how the *mean rule* works. Depending on the anchor selection, it takes a mean of p-values or test statistics. The problem becomes especially clear on p-values. The idea of a mean over p-values would be that a high mean speaks for a low probability of the alternative hypothesis. But this is generally not the case. Only if the alternative hypothesis is true, lower p-values are to be expected. Under the null-hypothesis p-values are uniformly distributed across 0-1 and carry no actual information. A mean would treat these p-values as meaningful. There are cases when this does not matter, as the effects cancel each other out and appear as simple noise, for example with the MP selection. Here the mean over a lot of p-values is taken and if the null hypothesis is true, the p-values are just random noise with a mean of 0.5. In this case the *mean rule* performs adequately, but then it is also equivalent to the *all rule*. In other cases, when the number of p-values is small, the noise does not cancel out and the mean becomes more or less random. This is for example the case with the NST and NC selection. Future research may investigate more appropriate approaches for aggregating p-values and test statistics, potentially adopted from meta analysis.

The *direct approach* performs well with many anchor selection methods but performs substantially worse than every other rule with the MT and MTT anchor selection. Until now, we do not have a conclusive explanation for this behavior and further research is needed. But with this uncertainty the *direct*

approach is also generally not advisable.
When the aim is to find an aggregation rule that does not show a substantially inflated type I error rate in any condition and has a high or medium hit rate over all conditions, the *min rule* can be recommended. However, there are also several conditions where the *all rule* shows that it can achieve a higher hit rate in combination with certain anchor selection methods, such as the MP anchor selection method.
Furthermore it should be noted that it is also a philosophical and practical question how DIF is supposed to be treated. The *all rule* does not lose any information and can differentiate between an item that discriminates only few focal groups or all focal groups. The *min rule* on the other hand cannot distinguish and treats all cases the same as soon as only one group shows DIF. This can become a problem when the number of groups increases, as the *min rule* may be too sensitive then. The advice to the researcher here is to determine how DIF should be treated. If the goal is a perfectly DIF-free test while accepting the risk of many items being flagged as DIF, then the *min rule* is to be preferred. If the idea is to have a test that is free of DIF for most groups, with the advantage of less items been flagged, the *all rule* is preferable (the items not being flagged by the *all rule* as opposed to the *min rule* may be actual DIF items, but may only affect few people).
Furthermore we also looked at the anchor selection process and determined why the NC anchor selection is outperformed by other methods. As an iterative method it highly emphasizes on the first selected items, so these items should be picked especially carefully. But often the first item picked by the NC anchor selection method is contaminated. Remember that in the first step of the NC anchor selection, the first item is picked using the NST anchor selection method, which is not performing well on average. Craig (2017) showed that changing the method of finding the first item does not yield substantially better results and other methods still outperform the NC anchor selection. Further research could determine whether choosing more than one item as a starting point or even using multiple starting points and average an anchor out of these could improve the NC anchor selection.
Finally we added a practical example from a general knowledge quiz to illustrate the importance of thoughtful anchoring. We showed that different choices of the aggregation rule can lead to different results. The *min rule* seemed to be performing better then the other aggregation rules. It flagged more items as DIF items that were plausible. This is in accordance with the high hit rates we saw in the simulation study. The *direct approach* and the *all rule* performed at an acceptable level. Items were generally plausible although in comparison with the *min rule* some DIF items might not be detected. This again reflects our findings from the simulation study. The

mean rule performed substantially worse than the other aggregation rules. Items flagged by the *mean rule* were either also found by the other methods or hard to interpret. This is most likely due to the high false alarm rates this approach already showed in the simulations. Therefore the *mean rule* is generally not advisable.
Overall, the findings of this study highlight the importance of using suitable methods for DIF analysis, and provide guidance in this endeavor. Another important problem in practice may be sample size restrictions. As opposed to large scale assessments, parts of educational research can only rely on smaller samples and thus may suffer from insufficient power. Insufficient power has a direct effect on the hit rate of DIF tests. So in many research areas it may be of interest to optimize the choice of anchoring methods with regard to hit rates and choosing an appropriate aggregation rule is one means to achieve this.

Chapter 3

The effect of different ratios of group sizes in multiple group scenarios on the detection of DIF

3.1 Introduction

Determining the necessary sample size for a study can be a complex topic. Especially when the sample should contain multiple groups. Time and financial restrictions can limit the overall sample size and the question remains how many test takers should be tested from every group. In this study we develop guidelines for optimal group sizes when the aim is to detect Differential Item Functioning (DIF) in multiple group scenarios with a fixed sample size. We are using the term ratio to describe the distribution of persons to groups. This is an important topic, as it has, to our knowledge, not been investigated in a multi group scenario before.
While the focus of Huelmann et al. (accepted for publication) was on aggregation rules in multi group DIF scenarios, the results showed that slight changes on the overall sample size that go along with strong changes in the group ratios can have drastic effects on the hit rates (i.e. the power to detect true DIF items).
This implies that the ratios have an important effect on the hit rates and we investigate here how to optimize the ratios with regard to hit rates.
In the literature little is known about effects of the group ratios on hit rates for DIF. In general, a high number of participants boosts the precision of parameter estimation. Therefore even tiny DIF effects can be statistically

detected in very large samples. Unfortunately, previous studies measuring effects of the group ratios in two group scenarios often confound the group ratio with the overall sample size. For example Awour (2008) and also Kilmen (2016) investigated two group scenarios. A reference group size of 1000 was simulated and the focal group size varied between 250, 500 and 1000, simulating ratios of 4:1, 2:1 and 1:1. The reported hit rates were highest in the 1:1 ratio and lowest in the 4:1 ratio. It can be criticized, however, that the effect of the ratio is confounded with the overall sample size. The overall sample size in their studies was 1250, 1500 and 2000. Of course the hit rates are expected to be higher with higher overall sample sizes.
In our study we therefore investigate fixed overall sample sizes and vary only the ratio of persons within and out of the reference group. Moreover, we do not examine two group scenarios, as solutions for two group scenarios are rather trivial to optimize, but expand on multi group scenarios. First we analyze the theoretical background. We show that the group ratios influence the hit rates of DIF tests, but also that it is not feasible to develop closed forms to calculate optimal group ratios. A closed form is difficult to derive and will rely on information the practitioner will usually not have, like the exact structure of DIF. We therefore determine a rule of thumb that gives acceptable hit rates in many conditions and relies on assumptions that are easier to make for researchers.

3.1.1 Theoretical Approach

In this study we concentrate on the generalized Lords χ^2 test for assessing DIF. Furthermore we concentrate on the Rasch model as a data generating process. The test statistic of the generalized Lords χ^2 test Q_j for item j is calculated as

$$Q_j = (\mathbf{C}\mathbf{v}_j - \mathbf{C}\xi_j)^t(\mathbf{C}\Sigma_j\mathbf{C}^t)^{-1}(\mathbf{C}\mathbf{v}_j - \mathbf{C}\xi_j) \tag{3.1}$$

where $\mathbf{v}_j$ is the estimated parameter vector for item j and all K groups and ξ_j is the corresponding hypothesized parameter vector. The t marks the transpose of a matrix. Σ_j is the covariance matrix of $\mathbf{v}_j$. Note, that the variance does depend on the group sizes, as it is the variance of the estimated parameters. It is also noteworthy that this formulation would also be valid for IRT models other then the Rasch model. $\mathbf{C}$ is called the contrast matrix. The contrast matrix defines which comparisons are made. A common contrast matrix C in a scenario with 2 focal groups and one reference group would be

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \tag{3.2}$$

To see which comparisons are made, the contrast matrix is simply multiplied with the parameter vector:

$$\mathbf{C}\xi_j = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \xi_{1j} \\ \xi_{2j} \\ \xi_{3j} \end{pmatrix} = \begin{pmatrix} \xi_{1j} - \xi_{2j} \\ \xi_{1j} - \xi_{3j} \end{pmatrix} \tag{3.3}$$

In this case, the reference group (first entry) would be compared to the focal group 1 (second entry) and also to the focal group 2 (third entry).
An alternative contrast matrix like

$$C = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{3.4}$$

would lead to the reference group being compared to focal group 1 and focal group 1 being compared to focal group 2. A first intuition now leads to the (as we will soon show wrong) assumption that the contrast matrix is the reason why the group ratios have an influence on the hit rates, as it is the only obvious part of the test statistic that gives the groups different emphasis, because they appear more or less often in comparisons. However, the test statistic is actually independent of the choice of the contrast matrix (at least under very broad assumptions).
The null hypothesis that all parameters are equal in all groups can be formulated as:

$$H_0 : \mathbf{C}\xi_j = \mathbf{0} \tag{3.5}$$

Introducing this restriction in equation 3.1 gives us:

$$Q_j = (\mathbf{C}\mathbf{v}_j)^t (\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^t)^{-1}(\mathbf{C}\mathbf{v}_j) \tag{3.6}$$

Q_j then follows a χ^2_p distribution where p denotes the number of rows of $\mathbf{C}$. As row vectors of contrast matrices are linearly independent by definition it is also the rank of $\mathbf{C}$. If we define now $\mathbf{C_1}$ and $\mathbf{C_2}$ as two contrast matrices with rank p, it can be shown that a matrix $\mathbf{B}$ exist with $\mathbf{C_1} = \mathbf{BC_2}$ where $\mathbf{B}$ is non singular (Johnson & Wichern, 1992). Then we can show that $\mathbf{B}$ cancels out:

$$\begin{aligned} Q_j &= (\mathbf{C}_1\mathbf{v}_j)^t(\mathbf{C}_1\mathbf{\Sigma}_j\mathbf{C}_1^t)^{-1}(\mathbf{C}_1\mathbf{v}_j) \\ &= (\mathbf{B}\mathbf{C}_2\mathbf{v}_j)^t(\mathbf{B}\mathbf{C}_2\mathbf{\Sigma}_j\mathbf{C}_2^t\mathbf{B}')^{-1}(\mathbf{B}\mathbf{C}_2\mathbf{v}_j) \\ &= (\mathbf{C}_2\mathbf{v}_j)^t\mathbf{B}^t(\mathbf{B}^t)^{-1}(\mathbf{C}_2\mathbf{\Sigma}_j\mathbf{C}_2^t)^{-1}(\mathbf{B}^t)^{-1}\mathbf{B}^t(\mathbf{C}_2\mathbf{v}_j) \\ &= (\mathbf{C}_2\mathbf{v}_j)^t(\mathbf{C}_2\mathbf{\Sigma}_j\mathbf{C}_2^t)^{-1}(\mathbf{C}_2\mathbf{v}_j) \end{aligned}$$

So we see that the test statistic is independent from the choice of the contrast matrix.

To determine the power of a test, we need to investigate the test statistic under the alternative hypothesis. Under the alternative hypothesis Q_j is distributed as follows:

$$Q_j = (\mathbf{C}\mathbf{v}_j)'(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}')^{-1}(\mathbf{C}\mathbf{v}_j) \sim \chi^2_{p,\lambda} \tag{3.7}$$

with λ being the non centrality parameter, defined as the sum of the squared true parameter differences. Note, that this quantity reflects what in simpler tests corresponds to the effect size, but it is much more complex and harder to quantify, as discussed below. To maximize the power, we need to maximize the expectation of Q_j. The item parameters are fixed and leave no room for optimization. The variance of the estimated item parameters depends on the actual item parameter and the group sizes. The group sizes, or as we usually describe it, the ratio, can be influenced by the researcher and thus provide a way to optimize studies with regard to power.
Unfortunately it is particularly hard to maximize this function. From an analytical point of view, the function is rather hard to derive, as the factors heavily depend on each other. Note that even analytically determining the variance as a function of group size is not a trivial task (Thissen & Wainer, 1982). In addition to these calculations being quite demanding, the function also relies on informations the researcher usually will not have available. These informations include, for example, the exact item parameters not only for the reference group, but for all groups. This is essentially the same as knowing the item parameters for the reference group and the size of the DIF for every group and every item parameter, which would serve as the effect size in a multi group DIF analysis. While it is true that for power calculations usually educated guesses are made regarding the effect size, for DIF analysis this is particularly difficult to do. In a classical power analysis only one effect has to be estimated before the data is drawn. For example in a study to test a new drug, only the effect of medication has to be guessed. In a power analysis for DIF in a multiple group scenario not only one, but multiple estimations regarding effect sizes have to be made. For example in a test with 40 items and 6 groups $40 \times (6 - 1) = 200$ estimates would have to be made before the data is drawn. Also, the estimations would have to be made on the scale of the item parameter, but practitioners are usually not trained in thinking in these scales. For example, for a single item it is easier to make a guess like "group A has a 20% higher chance of answering correctly", than a guess like "group A's item parameter is 0.34 higher" (both in comparison to a reference group). We therefore do not think a closed form would be useful and it would also be hard to determine. Therefore we tried instead to develop a rule of thumb that relies on information that we expect to be actually available to scientists in realistic settings.

3.2 Method

We developed the rule of thumb iteratively. We first started with a naive rule of thumb and simulated various DIF-structures. We saw certain situations where the naive rule of thumb failed and therefore derived a new rule of thumb from these scenarios, that succeed in these simulations. We then constructed more DIF-structures where the new rule of thumb might fail. But also in these scenarios the rule of thumb produced acceptable to very good hit rates, but left room for improvement in certain other scenarios. From this knowledge we derive suggestions for practitioners.

3.2.1 Simulation Design

To develop a rule of thumb we conducted multiple simulation studies. We varied the size and direction of DIF effects (which will be described in detail later), the number of groups (three and six) and the overall sample size (1000, 3000 and 5000). For brevity only the results for the six group scenarios are shown in the following as the results for the three group scenarios were very similar to the results from the six group scenarios. Furthermore, some DIF structures cannot be implemented in the three group scenario and therefore the results are omitted in this study. Data was simulated according to the Rasch model. Fourty items were simulated and 18 items from these were chosen randomly to be DIF items. Item difficulties for the 40 generated items were taken from Wang et al. (2012) and can be found in the appendix B.1. Datasets were analyzed with the generalized Lords χ^2 test. Four random non-DIF items were chosen as anchor items to align the parameters. In practice it is generally unknown which items DIF free are and make good anchor items, but the focus of this study is to develop a rule of thumb. Therefore we chose this anchoring approach as a baseline to compare different rules of thumb without the burden of anchoring, which can be a rather complex topic (Kopf et al., 2015a). Group ratios were either determined by rules of thumb, which are explained later, or at random. For random assignments a fixed total sample size N was randomly distributed across the groups with the restriction that every group consists of at least 100 persons. These random assignments were used to compare their performance to those of the rules of thumb. For every scenario 100 random assignments were drawn.

3.2.2 Computational Details

For the simulations we used 500 iterations for every assignment of the 100 random group ratios and 100 iterations for the group ratios determined by

the rule of thumb. All computations were done in R (R Core Team, 2019). Furthermore we used the R packages difR (Magis et al., 2010), psychotools (Zeileis et al., 2018), and ltm (Rizopoulos, 2006).

3.2.3 Deduction of a rule of thumb

As we showed in the *Theoretical Approach* section, the ratio of group sizes does have an effect on the hit rates. Therefore it is advisable to optimize group ratios. As we also showed, a closed form is not easy to derive. Even if this closed form would be derived, the necessary information is not easy for practitioners to find. Every closed form will rely on an accurate estimate of the effect size before the actual data is collected. In the case of multi group DIF, this means that for every group and every item practitioners would have to make exact guesses of the size of the effect. This is especially hard, as it has to be estimated on the same scale as the IRT parameters. But when thinking about DIF, practitioners usually do not think in the scales of IRT parameters. A feasible rule of thumb should therefore not rely on exact guesses of all DIF effects.

3.2.4 Naive rule of thumb

A first naive approach for a rule of thumb would be to set all group sizes equal. While this is a reasonable idea when no information about the DIF structure is at hand, it is easy to construct cases where this rule of thumb does not give satisfying solutions. We therefore developed a new rule of thumb by means of investigating several simulation settings.

3.2.5 New rule of thumb

The first round of simulations showed that the naive rule of thumb without any assumptions on the DIF structure fails to deliver good results under certain conditions. We therefore developed a new rule of thumb based on assumptions researchers might be able to make in practice. The least amount of information on DIF structures is needed when assumptions are made on a nominal scale. By this we mean that the researcher does not need to quantify the exact amount and structure of the DIF, but only make a statement whether or not DIF is expected to be present between two groups and whether or not it is equal to other DIF effects. Furthermore we showed in the theoretical approach that the test statistic is independent of the choice of the contrast matrix. Therefore, the new rule of thumb will not rely on the contrast matrix or, equivalently, on the choice of the reference group.

The idea of this new rule of thumb is that for certain items groups can behave similarly. Take, for example, an item that shows DIF between native and non-native English speakers. It is probably safe to assume that the DIF effects for Austrians and Germans are comparable, as in both countries German is the most prominent language and the school systems are comparable with respect to training in English. The idea of the new rule of thumb is that groups can "borrow" information from similar groups in DIF situations. Remember that the test statistic does not try to identify which group shows DIF, but only whether or not there is DIF between any groups for an item. If we now think of a test where every item is composed in such a way that the Austrian and German group have roughly the same DIF effect, we assume that one group can "borrow" 100% of information from the other group's test takers for estimating the DIF effect. This means that the Austrian group, for example, can rely on the Austrian group plus the German group. We introduce here the terminology of actual groups and the consortium of a group. The actual group is the group as defined by a certain criterion. In our example this would be the nationality. So, for example, the Austrian test takers are an actual group. The consortium of a group on the other hand is the aggregation of groups that a certain group can rely on for estimating the DIF effect. In our example the consortium of the Austrian group is the combination of the Austrian and the German group.
The next idea for constructing the new rule of thumb is that every group should have the same number of participants in their respective consortium. This restriction reflects the believe that an optimal group ratio has the same amount of information on every group. If there is not the same amount of information on every group at least one group will be underrepresented and therefore item parameters for this group could be measured with a lack of precision. Assume our sample consists of three actual groups in total, for example American, Austrian and German test takers. The American group can rely on no one else in this example and therefore the consortium of Americans consists only of the American group. If every consortium should be of equal size of participants, we can deduct the following restrictions:

$$\begin{aligned} USA_c = AUS_c = GER_c \\ N = USA + AUS + GER \end{aligned}$$

with USA, AUS and GER representing the actual groups and USA_c representing the consortium of the American group (and the others accordingly). Furthermore we can conclude these restriction concerning the relationship

between the actual groups and the consortia of the groups:

$$USA_c = USA \tag{3.8}$$
$$AUS_c = AUS + GER \tag{3.9}$$
$$GER_c = AUS + GER \tag{3.10}$$

From equation (9) and (10) we can conclude:

$$GER_c = AUS_c$$

And therefore also:

$$N = USA + AUS_c$$

If we now want to maximize every groups consortium at the same time, we can deduct:

$$USA = N/2$$
$$AUS_c = N/2$$

If in a second step we also maximize the number of persons in every individual group, we arrive at the solution $AUS = GER = N/4$. Note that mathematically other solutions would also be possible here. Setting, for example, the German group to zero and dividing the overall N between USA and Austria would also satisfy the formula. But our simulations showed that dividing equally between groups that behave in a similar way delivers better hit rates. This is most likely due to a trade-off between the precision of estimation and the size of the bigger group. If one group is very big and therefore the other group is nearly empty, there is no precision of estimation in the small group. But if both groups are of equal size, in both groups a certain level of precision can be guaranteed. This means in a three group scenario where two groups act essentially the same, we would suggest half of the total N to be in the outstanding group and the two groups acting the same taking equal amounts of the other half. With a N of 400, for example, we suggest 200 American, 100 German and 100 Austrian test takers. Note, that this is also the optimal solution suggested in our first simulation.
But groups acting essentially the same might not be a realistic assumption. Therefore we expanded the idea of "essentially the same", to "partially essentially the same". In keeping with the analogy, a group where 50% of the DIF items are essentially the same as in another group could also only 'borrow' 50% of the other groups' members. Therefore the consortium of that group

would only consist of the actual group and 50% of the group it can borrow from. This could be the case if we, for example, substitute the German group with a group from the Netherlands in our example. We assume here that Austrian and Dutch test takers have some similar DIF because of a similar language, but there could also be Austrian or Dutch test taker specific DIF items. If the study only consists of Americans, Austrians and Dutch, and we further assume that 50% of the DIF items are equal for Austrian and Dutch test takers, we then calculate the consortium of Austrians as the actual group of Austrians plus 50% of the Dutch group.
If now all percentages of how much a group can borrow from another are organized in a matrix, it is a symmetric $g \times g$ matrix we will call P, where g is the number of groups. In the example of the American, Austrian and Dutch group, the matrix P would be:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}$$

The row sums are then 1, 1.5 and 1.5. In a study where all actual groups consisted of only one person, these row sums would give the respective sizes of the consortia. We have assumed that an optimal solution is when every consortium has the same size. Thus, calculating group ratios becomes a simple arithmetic problem. In matrix notation this can be written as:

$$N \frac{P^{-1}\mathbb{1}}{\mathbb{1}^t P^{-1}\mathbb{1}} \tag{3.11}$$

with N the overall sample size, $\mathbb{1}$ a g-dimensional vector consisting of 1's and P a $g \times g$ matrix where the entry p_{ij} depicts how many of all DIF items from group i and j are essentially the same. In the example with the American, Austrian, and Dutch group, the inverse of the matrix P is:

$$P^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$$

Used with equation 3.11 we get

$$N \times \frac{\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}{\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{3} & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}} = N \times \frac{\begin{pmatrix} 1 \\ \frac{2}{3} \\ \frac{2}{3} \end{pmatrix}}{\frac{7}{3}} = N \times \begin{pmatrix} \frac{3}{7} \\ \frac{2}{7} \\ \frac{2}{7} \end{pmatrix}$$

For a total N of 700, for example, this rule would then suggest 300 ($= 700 \times \frac{3}{7}$) persons for the American group and 200 ($= 700 \times \frac{2}{7}$) persons for the Austrian as well as the Dutch group.
From a technical point of view, note that if vectors of P are linearly dependent, P is not invertible. This is the case in the example with the American, Austrian and German group, where every focal group acts the same. But then one of the linearly dependent vectors can be crossed out of the matrix and after the calculation is done, the result for the not-crossed out group is divided equally across the crossed out groups. In the example we could cross out the German group, calculate then and divide the result for the not-crossed out Austrian group between the German group and the Austrian Group. R code for the calculation is provided in the appendix.
Take, for example, the hypothesized DIF structure portrayed in Table 3.1.

We can now calculate the matrix P by counting how many DIF items are expected to be equal between two groups. The percentages are given in Table 3.2.

In this example focal group 1 and focal group 4 have the same DIF effects. For the calculation we can therefore cross out one of the groups and divide the result for the other group equally across both groups. We can now use formula 3.11 on the P matrix. Group ratios are then calculated for example for 1000 participants as given in Table 3.3.

See that there is no ratio calculated for focal group 4. This is due to fact that we crossed out focal group 4 before the calculation. To get an estimate for the optimal group size of focal group 4, we simply divide the estimated group size of focal group 1 between the two groups: $263.93/2 = 131.965$. Remember that focal group 1 and 4 had the exact same expected DIF effects, and that is why we crossed out group 4 in the first place.
See that for the calculation we did not rely on the exact DIF structure but only on the matrix P. So a different DIF structure - for example a DIF structure where all effects equal to 0.2 are replaced by -0.5 - that leads to the same matrix P would give the same results for the optimal group ratios. Note that this rule of thumb is also applicable in the two group scenario, but the result is trivial, as it will always be a 50/50 split.
Further, an issue with this new rule of thumb could be mathematically optimal solutions, that are not reasonable or even possible in the actual application. For example, the new rule of thumb can give very low group sizes or even groups of size 0 in cases where the ratio of a consortium to its group is relatively big in comparison to other groups. For example when a group A can rely for 50% of the items on information of group B and for the other 50% on information of group C. The matrix P can then be written as:

Table 3.1: Example DIF structure

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.40	0.40	0.30	0.40
0.00	0.30	0.40	0.40	0.30	0.20
0.00	0.30	0.40	0.40	0.30	0.30
0.00	0.30	0.40	0.20	0.30	0.30
0.00	0.30	0.30	0.20	0.30	0.20

Table 3.2: Example of a matrix P

	Ref	Foc 1	Foc 2	Foc 3	Foc 4	Foc 5
Ref	1	0	0	0	0	0
Foc1	0	1	0.2	0	1	0.4
Foc2	0	0.2	1	0.8	0.2	0.2
Foc3	0	0	0.8	1	0	0.4
Foc4	0	1	0.2	0	1	0.4
Foc5	0	0.4	0.2	0.4	0.4	1

Table 3.3: Example ratios for 1000 participants

Ref	Foc 1	Foc 2	Foc 3	Foc 5
325.510	263.93	73.310	219.94	117.300

$$P = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}.$$

Using equation 3.11 we can then derive the solution of

$$N \times \begin{pmatrix} 0 \\ 0.5 \\ 0.5 \end{pmatrix}.$$

Note, that group A will then always be suggested to be 0 and the observations are split between group B and C. This might be inconvenient in a study if there are further research questions related to group A. In these cases we suggest to define a minimal size for every group and only distribute the remaining observations through the new rule of thumb. In extreme cases the calculated optimal groups sizes could even produce negative values. This occurs when the ratio of a consortium to a group is very much larger than the ratio of consortium to group in the other groups. Negative values did not occur in the investigated simulations, even though they cover a wide range of scenarios. Therefore we do not expect this to occur frequently in practical simulations.
When the ratio of the consortium of a group to its actual group is very big in comparison to the other groups, it means that a group "borrows" information from many other groups. The logic behind the new rule of thumb then

concludes that other groups can explain reasonably well the DIF effects in this group and the group itself is therefore not needed or even disruptive.

3.3 Simulation Studies

In the following section we will illustrate the performance of the naive and the new rule of thumb in various different simulated settings.

3.3.1 Simulation 1

Method

This simulation was used to derive the new rule of thumb. In the first simulation we simulated 6 groups. The reference group is DIF free by definition and the focal groups all exhibit the exact same DIF effects compared to the reference group. We call this DIF structure the norm-DIF structure.

Results

The naive rule of thumb (equal group sizes) in comparison to random group ratios is shown in Figure 3.1. It can be easily seen that the achieved hit rate is not optimal.

After inspecting the naive rule of thumb in this scenario, we investigated the one group ratio from the random assignments that was performing the best in regard to hit rates. This was to set the reference group as half of the overall sample size and divide the rest equally across the other groups. This insight was used to derive the new rule of thumb explained above. Hit rates for the new rule of thumb were added in Figure 3.1. We already showed with the theoretical approach that the reference group in Lord's χ^2 tests does not have a special role compared to the focal groups. We therefore have to conclude that a rule of thumb should not be based on the choice of the reference group. The reference group in the first simulation was special in contrast to other groups as it was the only group not showing DIF, and even more importantly, all other groups showed the exact same DIF structure. We therefore conclude, that for a rule of thumb to give good results in various scenarios, it has to consider the DIF structure. The new rule of thumb gave the optimal result in this scenario and we therefore investigated it further.

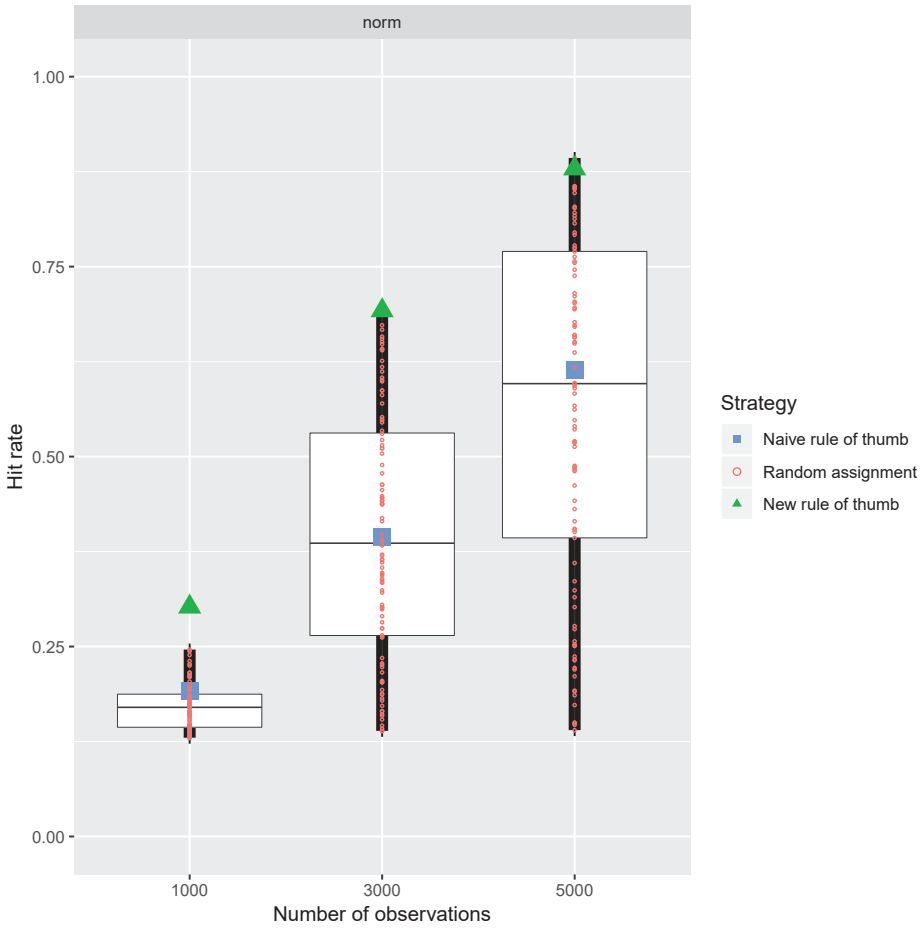

Figure 3.1: Hit rates in the six group scenario under the "norm" DIF structure

3.3.2 Simulation 2

Method

To see whether the rule of thumb derived from the first simulation would give satisfying results in other settings too, we simulated more DIF-structures. The second simulated DIF-structure is called var. In this scenario, two groups showed no DIF at all, two groups showed positive DIF on all 18 DIF items, and two groups showed negative DIF on all 18 DIF items.

Results

With the new rule of thumb good ratios were found in this scenario. In Figure 3.2; hit rates are given under the var DIF-structure.

3.3.3 Simulations 3 to 8

Method

For brevity we do not discuss all simulation settings here in detail, but merely give an overview. We constructed diverse DIF-structures to test whether the new rule of thumb holds in different scenarios. In the next six scenarios we investigated, the rule of thumb gave adequate hit rates. We named these scenarios chaos, bal, groups, incdec, equal and alter. In the chaos scenario all 18 DIF items had a random value between -0.6 and 0.6 added. In the bal scenario DIF was balanced within each focal group by either adding or subtracting 0.3 from the difficulty parameter. In the groups scenario, two focal groups had a DIF of size 0.3 and three focal groups had a DIF size 0.6. In the incdec scenario the size of DIF was increasing from 0.3 in focal group 1 to 0.7 in focal group 6, while the number of items affected by DIF was decreasing from 18 items in focal group 1 to 6 items in focal group 6. In the equal scenario DIF was either set to 0.3 or to 0.6. The number of DIF items varied between 18 and 9 items. The sum of DIF within each focal group was set to 5.4. So, for example, there were 18 DIF items for the first focal group with a constant DIF effect of 0.3 added (always compared to the reference group). On the other hand there were only 9 DIF items for the fifth focal group, but the DIF effect for these 9 items was set to 0.6. An overview of the DIF structures is given in the appendix.

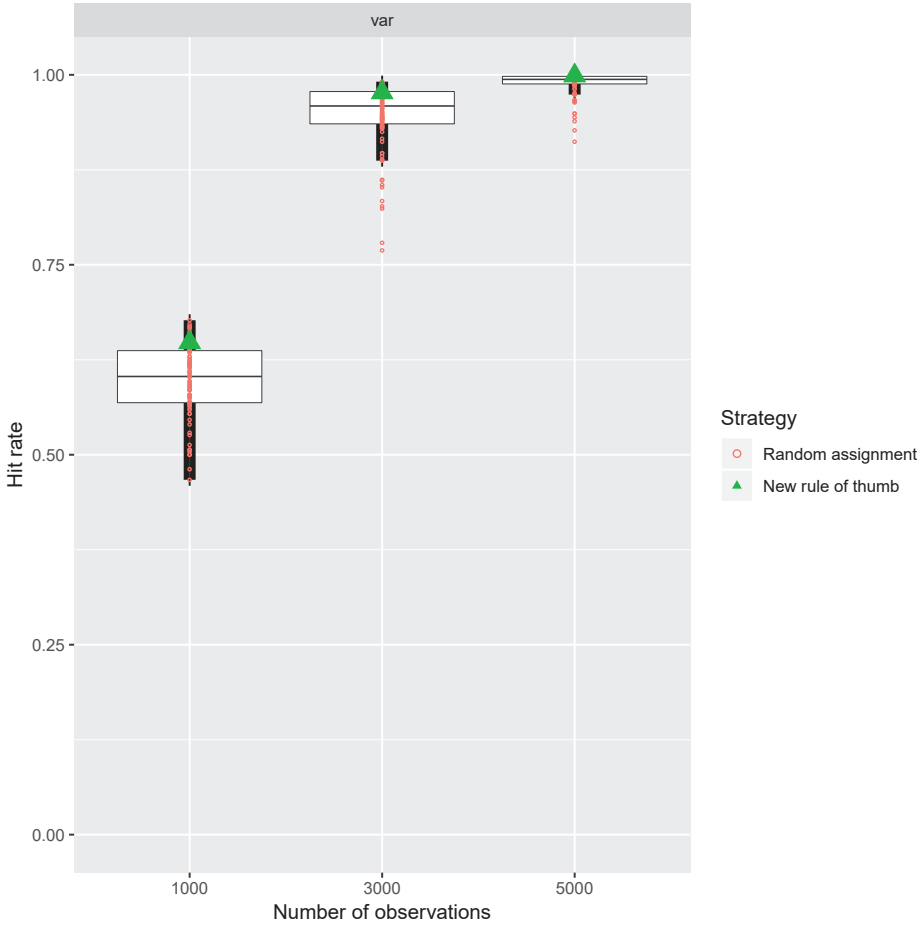

Figure 3.2: Hit rates in the six group scenario under the "var" DIF structure

Results

Overall in all these simulations the new rule of thumb gave satisfying hit rates. We define a satisfying hit rate as hit rates above the median of hit rates from random assignments. Especially in the groups, incdec, equal and alter scenario hit rates where consistently well above average. All hit rates can be inspected in Figure B.2 in the appendix.
In the chaos and bal scenario hit rates with the rule of thumb where not satisfying in small samples of one thousand observations. Keep in mind that with six groups overall, this means group sizes can become rather small. However, with the overall N rising in the bal scenario hit rates are again well above average for the rule of thumb. Only in the chaos scenario the hit rates for the rule of thumb are only slightly above average for the sample sizes bigger than 1000. We evaluate this performance as still satisfying, as in this scenario the rule of thumb was used without correct information and still delivered above average hit rates. Hit rates for the chaos and the bal scenario can be inspected in Figure B.1 in the appendix.

3.3.4 Simulation 9

This simulation gives an outlook on a scenario where additional information could be used to derive an extension of our rule of thumb.

Method

In the inc scenario we simulated a small increase of DIF from group to group. Therefore the first focal group had a DIF of 0.1 for every DIF item. For the following focal groups DIF was increased by 0.1 until in focal group 5 a DIF of 0.5 was reached.

Results

In this scenario reported hit rates were slightly above average. The results of our simulation can be seen in Figure 3.3.

We evaluate this as adequate behavior, as the resulting hit rates are still better than most random assignments. However, rules of thumb relying on more information than our rule of thumb could give better hit rates in this scenario. For example, if the overall sample size was split between the reference group and the focal group 5 (the group with the biggest DIF) and all other groups are left empty, better hit rates could be reported. Note, however, that a rule of thumb based on this principle would always rely on the information of which group has the biggest DIF.

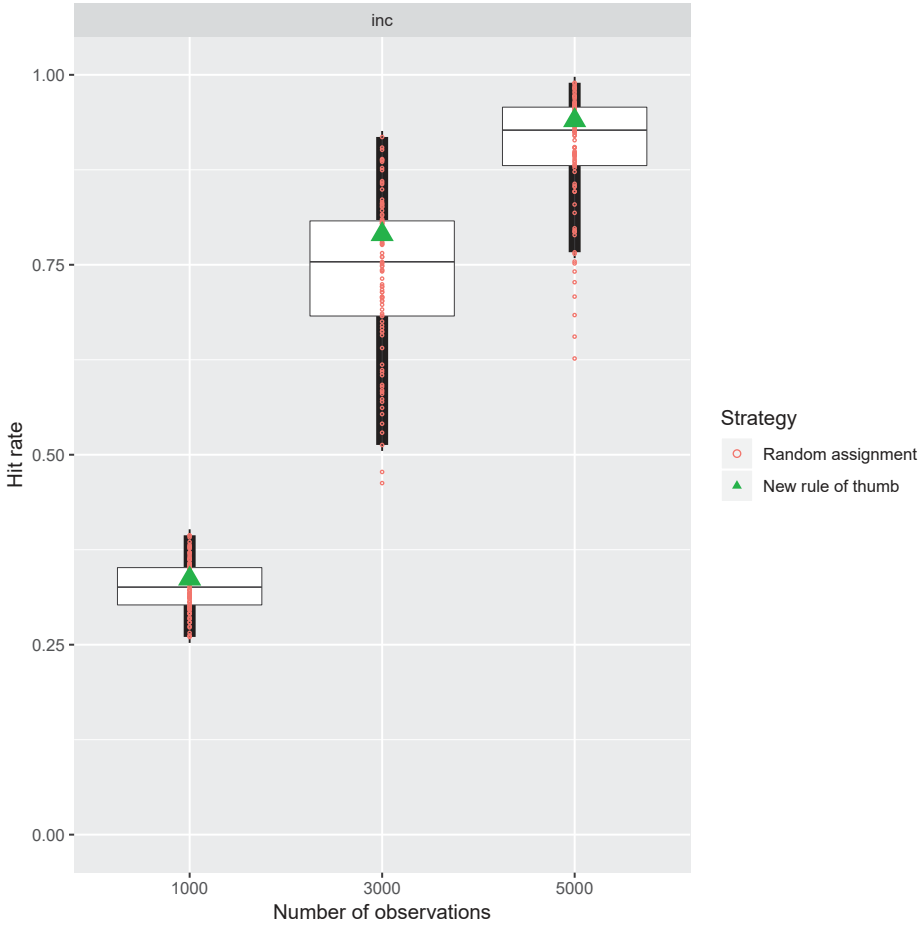

Figure 3.3: Hit rates in the six group scenario under the "inc" DIF structure

3.4 Discussion

In this study we developed a rule of thumb to determine group ratios in multi group DIF scenarios. We started with a theoretical approach and showed that a closed form is hard to determine as well as not feasible in practice. We then developed a rule of thumb in a two-step approach. In a first step we used a naive rule of thumb, setting all groups to be equal. If it is not possible to make an educated guess about the DIF structure, this is a viable strategy. Note that the naive rule will give an optimal group ratio only in few scenarios, but will also only fail completely in rather rare cases.
If an educated guess on the DIF structure on a nominal scale can be made, we suggest a new rule of thumb that relies on the similarity of the groups. A guess on a nominal scale here means that researchers can make assumptions whether DIF effects are present and if they are equal to each other. Researchers do not need to make assumptions about the order, direction or size of the effects. This approach yielded better hit rates than the naive approach and in many scenarios also good to nearly optimal group ratios. This approach did show some systematically lower hit rates in certain scenarios with smaller sample sizes, but in comparison to random assignments of groups these hit rates were still acceptable. Therefore we still recommend this rule of thumb if no assumptions on the DIF structure on an ordinal scale can be made.
Assumptions on an ordinal scale in this context would mean that assumptions on the direction and order of the DIF effects can be made, but not necessarily on the actual size. If assumptions on an ordinal scale can be made, we recommend to look for dominating group comparisons, meaning the DIF between two groups is always bigger than between any other groups. We simulated this DIF structure in the 9th simulation with the inc DIF structure. In this specific case, a group ratio where the dominating groups are maximized and all other are not filled, or only with a minimum number, would possibly show better results than the rule of thumb. A scenario where this DIF structure could occur, is, for example, when we add a Chinese group to the example with the American, Austrian and Dutch group. We still assume there is DIF between each group. But as we also still assume that the language is the only reason for DIF in this test, we could conclude that the DIF effect between the Chinese group and the American group is always the biggest, as Dutch and German are more closely related to English than Chinese. Further research is needed to generalize this idea into another rule of thumb specialized for scenarios with dominating group comparisons.
In all but one simulation we used the true DIF structure for computing the optimal group sizes. In applied research this DIF structure will not be avail-

able. Note, however, that the exact same problem also occurs in classical power analysis, where assumptions about the true effects need to be made and the actual power may be lower if the effect size is misspecified. To test for robustness of our method we also included a chaotic DIF structure. In this scenario, the true DIF structure was not known, but only the expectancy values of the DIF effects. The hit rates in the chaotic scenario where still acceptable and therefore we conclude that our rule of thumb is robust to slight misspecification of the DIF structure.
Another issue with our rule of thumb is the problem of empty groups. In some cases the rule of thumb advises to leave certain groups empty. As long as the DIF analysis is the only analysis on the dataset, this is also advisable to maximize hit rates. However, if further analysis are planned for the dataset it might be not optimal to leave groups empty. In this case we would advise to give a minimal number of participants to the empty group which can be taken equally from all other groups. Furthermore, it is possible that optimal group ratios are estimated to be negative. This did not occur in our simulation, but is theoretically possible. Further research is needed to define rules for these special scenarios.
No rule of thumb delivered consistently the optimal group ratio. Often some random group ratios performed slightly better. This leads to the last suggestion. If assumptions can be made on a metric scale, a power simulation with random group ratios can determine close to optimal hit rates. While simulations can be time consuming, especially when the overall sample size or the number of groups is high, they are still comparably easy to do in comparison to analytically determining an optimal ratio.

Chapter 4

An extension of the anchor point selection method to multiple groups

4.1 Introduction

Differential item functioning (DIF) is an important issue for test developers. It is closely linked to test fairness and identifying DIF items can lead to improved test quality (Osterlind & Everson, 2009). The IRT framework has testable assumptions and is therefore suitable for DIF detection, as DIF can be formalized as a violation of these assumptions. In this study we will focus on the Rasch model (Rasch, 1960). Strobl et al. (2019, note this manuscript is still in preperation so that "anchor point selection" might not be the final title of the paper) proposed a method to test for DIF in two group scenarios, that is based on optimizing an inequality criterion. In this study we will expand the idea to multi group scenarios.
In the IRT framework the classic approach of DIF analysis is to align item parameters in some way. This alignment is usually called anchoring and often done by choosing a set of items as so called anchor items, either based on expert knowledge or by means of statistical approaches (see Kopf et al., 2015a, for a comparison of anchoring methods).
Choosing anchor items is not a trivial task and there is a no overall agreement in the literature on the process of this choosing. A researcher is confronted here with two main problems: Which items are optimal and how many items should be chosen. An actual DIF item in the anchor may corrupt the whole analysis and produce artificial DIF. The anchor length, on the other hand, is a trade-off problem. The longer an anchor is, the better the resulting

alignment. But this only works as long as the anchor stays uncontaminated, meaning there are no DIF items in the anchor. But with an increasing anchor length the risk of accepting a DIF item into the anchor rises.
The process of anchoring relies on multiple assumptions, depending on the anchoring method. One assumption often made is the assumption that the majority of items is DIF free. While this assumption is viable in general, as trained authors of items might produce more DIF free items then DIF items, it still is a strong assumption and cannot be tested. Bechger and Maris (2015) even argue that DIF could not be seen as an absolute size, but only in reference to an item. The idea is that for a single item it is impossible to know whether this item is fair in the sense of being DIF free or not. It is only possible to say whether an item is DIF free or not if there is at least one other item that can be used to align the item parameters. But as it is also impossible to say with absolute certainty that this second item is DIF free, the only conclusion possible is that the first item is DIF free or not when the second item is used as a reference item (or, as we usually call it in this study, an anchor item). We conclude therefore that anchoring is a complex topic and can have a big impact on the outcome.
In contrast to classical anchoring, where a subset of items is fixed as the anchor and which relies on the assumption that the majority of items is DIF-free, Strobl et al. (2019) proposed a method that follows a different rationale. The so called anchor point selection method does not rely on a subset of items specified in advance. Instead, item parameters for the reference group are estimated freely.
The term freely might be somewhat confusing as there are still restrictions on the estimation. One restriction often made is that the average difficulty of the items is set to zero. This is done to set the metric for the estimated parameters. Equivalently the first parameter could also be set to zero, which is another commonly used method to fix the parameter metric. While the anchor point selection method does not depend on a certain method for fixing the metric, it is probably easiest to imagine the procedure with the first item beeing fixed to zero. Therefore the reference group is freely estimated with the first item parameter fixed to zero.
For the focal group the first item is then also fixed to a number, but not necessarily to zero. Instead it is fixed to a value determined by an optimization algorithm based on an inequality criterion. For every single item parameter this algorithm checks the differences between the focal group and the reference group and tries to find the point where the majority of items have only a small difference and a minority of items have very big differences. See that this is a very different approach from classical optimizations like, for example, the least squares estimation for regression models, where it is the

goal to avoid big differences between the true and estimated values. As the criterion for this optimization Strobl et al. (2019) use the Gini index (e.g. Ceriani & Verme, 2012). The Gini index is an index known from poverty research. It is set out to measure inequality. For example it can be used to measure the inequality of wealth in nations. Haiti has a very big Gini index as there is great inequality in wealth in Haiti, which means few people own great wealth and many people are rather poor. In contrast, Norway has a rather low Gini index, as wealth is distributed evenly across the population (at least in comparison to Haiti). While Strobl et al. (2019) suggested the Gini index as the optimization criterion, others would be possible. For example Asparouhov and Muthén (2014) described a very similar method where the optimization criterion was not the Gini index but the sum of the square root of the absolute values of distances. While this criterion behaves very similarly, Asparouhov and Muthén (2014) did not motivate it as an inequality measure.
Strobl et al. (2019) used a Wald test (see, e.g. Glas & Verhelst, 1995) to compare two groups after an optimal point for anchoring was found. In this paper we would like to extend the idea of anchor point selection from a two group scenario to a multi group scenario.

4.2 Anchor point selection

Strobl et al. (2019) developed a method for aligning item parameters, that follows a different rationale than classical anchoring in IRT DIF analysis, where a set of anchor items is chosen based either on expert knowledge or on statistical procedures that typically assume that the majority of items is DIF-free. As stated above, in any type of anchoring, a restriction is needed for estimating the item parameters. In a DIF analysis this restriction has to be the same for all groups and is what we call anchoring. A possible, but not necessarily useful restriction would be to set the item parameter for one arbitrarily chosen item, e.g. the first item, to zero not only for the initial estimation, but to keep it like this also for the DIF testing. The problem in a DIF analysis is that the item parameters of the anchor items will not be estimated freely but set equal across all groups. If an anchor item happens to be a DIF item, a test would never show this, as the item parameters are equal by definition. Furthermore, as all other items are aligned according to this item, item parameter estimates will be shifted and DIF free items can test positive for DIF while actual DIF items could no longer be recognizable as DIF items. Alternatively not only one item, but the mean of a set of items can be set to a certain value to solve the problem of scale indeterminacy. For

the estimation of item parameters within one group this restriction can be set arbitrarily, as one restriction can easily be obtained by another. If, for example, the first item was set to zero as a restriction, by simply shifting the item parameters and recalculating the variance-covariance-matrix, another restriction, like setting the second item to zero, can be obtained. But for a DIF analysis all restrictions define certain items to be DIF free and can, if chosen poorly, induce artificial DIF by the restriction of setting certain items to be equal and shifting other items accordingly.
The approach of Strobl et al. (2019) now does not specify a certain anchor a priori, but optimizes over all possible anchors. The optimization criterion then is crucial to the analysis. Strobl et al. (2019) suggested the Gini index. As mentioned before, the Gini index is a measure of inequality. The idea behind the use of the Gini index is that a high amount of inequality very well reflects the commonly used concept of DIF, namely that few items have great DIF and the majority has no or little DIF.
The easiest way to do the optimization is to cycle through every possible restriction via a grid search. While this procedure is rather time consuming, it adds the benefit of a search path that can then be visualized and interpreted as shown below. With this search path researchers are able to identify easily not only the global maximum of the Gini index but also local maxima. This is very useful as sometimes content-wise the global maximum might not be the best choice, for example, when the restriction is violated that only few items show DIF, but the majority shows DIF. As long as some items interlock perfectly, this will show up as a local maximum. The researcher then can decide which solution is most appropriate. Strobl et al. (2019) showed that the grid search is not completely needed as you only need to cycle through the possible locations of the extreme points that can be derived mathematically. The optimal point found by the Gini index is called the anchor point here. With this anchor point the scales are aligned and Strobl et al. (2019) use a Wald test to compare the item parameters (see, e.g. Glas & Verhelst, 1995). When testing with the optimal anchor point, not all item parameters will necessarily have a variance. This is again due to the scale indeterminacy. As certain parameters are not estimated but set to a certain value, these parameters do not have a variance. But without a variance, parameters cannot be tested for equality. There are different possibilities to deal with this isssue. The most simple way is to leave items that do not have a variance out of the analysis. If an item parameter is set equal across all groups, independently of the variance it could not show DIF. Therefore leaving this item out is a viable strategy, as it would never test positive for DIF, even if it was a true DIF item. But for the anchor point approach, usually the parameters and variances are estimated in a first step and in a second step

only the item parameters are shifted. This shifting is the process we explained before. In order to avoid any problems with the variance of the item that was used in the initial restriction, we use quasi-variances, as introduced by Firth (2003).
Firth (2003) motivates quasi-variances with an example where ship types are compared in terms of the rate in which wave damage incidents occur. The example is adapted from McCullagh and Nelder (1989). Five ship types (A, B, C, D and E) are compared and ship type A is used as a reference category. Therefore the effect of the ship type A is defined as 0 and there is no standard error for this ship type. As long as only inference about other ship types in comparison to ship type A is supposed to be drawn, this is also completely sufficient (For example, whether type C is more prone to wave damage than ship type A). But more complex contrasts are not necessarily possible and especially it is not possible to draw inference on the effect of ship type A on wave damage itself. Quasi-variances are a way to calculate a standard error for these effects and make these inferences possible. As this inference is exactly what we are looking for, we decided to use quasi-variances here. Just like in the example with the ship type A, one item is used for the restriction - similar to a reference category - in every group. Therefore, as already mentioned, there is no standard errors for this item. But with the quasi-variances even this item can be compared across groups to test for DIF.

4.3 Anchor point selection in multiple group scenarios

To expand the idea of anchor point selection to a multi group scenario, we will first take a look at the performed DIF test. Strobl et al. (2019) implemented the Wald test (see, e.g. Glas & Verhelst, 1995) to test for DIF after the parameter alignment. We therefore use the generalized Lords χ^2 test (Kim et al., 1995) to test for DIF, as it can be seen as an extension of the Wald test for multi group scenarios. The choice of the test statistic is important for the development of the extension of the anchor point selection, as it defines which comparisons between the groups are made. The test statistic Q_j for the generalized Lords χ^2 test is defined as

$$Q_j = (\mathbf{C}\mathbf{v}_j - \mathbf{C}\xi_j)^t(\mathbf{C}\Sigma_j\mathbf{C}^t)^{-1}(\mathbf{C}\mathbf{v}_j - \mathbf{C}\xi_j) \tag{4.1}$$

where $\mathbf{v}_j$ is the estimated parameter vector for item j and all K groups and ξ_j is the corresponding hypothesized parameter vector. The t marks the transpose of a matrix. Σ_j is the covariance matrix of $\mathbf{v}_j$. $\mathbf{C}$ is called

the contrast matrix. The contrast matrix defines which comparisons are made. The comparisons made by the most common contrast matrix are that the reference group is compared to each focal group. The choice of the reference group is arbitrary, as the test statistic is independent of the contrast matrix. But as only the differences between the reference group and the focal groups have an influence on the test statistic and not the differences between the focal groups, for the multi group alignment we also only optimize for differences between each focal group and the reference group.
The optimal anchor point c is then defined as a (k-1)-dimensional vector, where k is the number of groups. See that this definition holds also true in the two group case where c is a one-dimensional vector. Each entry of c then describes the optimal anchor point for a focal group in relation to the specified reference group.
While in a two group scenario a grid search for the optimal anchor point was feasible, this is not the case in general, as the number of dimensions can become very big. Therefore we did not try to find the optimal anchor point with a grid search, but with the method suggested by Strobl et al. (2019). This method does not only provide the globally optimal anchor point, but all points where the derivative of the optimization criterion is zero or not continuous. From this finite set of points, the point maximizing the optimization criterion is then chosen as the optimal anchor point. Note that this is not necessarily the optimal - or the only optimal - solution from a content point of view. Therefore it might be interesting for researchers to not simply take the globally optimal point selected by this algorithm, but to visually inspect all local optima over the range of the possible points suggested by the method of Strobl et al. (2019).
To illustrate why it might be interesting to not simply look at the globally optimal value of c but also check for other solutions, we introduce a small toy example. Imagine a test for mathematics ability administered to Americans, Dutch and Germans. The Test consists of 20 items. The first 8 of these items are exercises containing text and the test is generally administered in English. The first eight items therefore measure not only the mathematics ability, but also English language ability. Note, that the groups should have different mean English language abilities. We set the American group as the reference group and therefore to be DIF free by definition. Furthermore we expect a moderate mean difference on the first 8 items for the Dutch group and a high mean difference for the German group (always compared to the American group). We explain these differences with differences in the exposure to the English language. For example, in the Netherlands movies are often not dubbed but only subtexts are added, while in Germany usually all movies are dubbed to German.

We can then plot the shifting parameter c against the Gini index for the Dutch group (focal group 1) and the German group (focal group 2) in two separate plots. An example plot is shown in Figure 4.1. In both plots there are two peaks. If we use the shifting parameter c according to the highest peaks we can plot the item parameters like we did in Figure 4.2. See that the first 8 items appear to be shifted and the last 12 items interlock. This is the most reasonable result in this example because mathematics ability is the primary dimension that we try to measure and language is only a nuisance dimension.
But the other peak might also be interesting for researchers. We plotted the item parameters with the shift according to the second peak in Figure 4.3. See how in this case the first 8 item parameters interlock. This would mean that these items define the dimension we are trying to measure, which would automatically induce DIF in the 12 other items.

From a mathematical point of view, both solutions are possible due to the scale indeterminacy. Most anchoring methods make the additional assumption that the majority of items is DIF free, meaning that the majority of items measure the main dimension the test is supposed to measure. In this example this is mathematics ability and would lead us to interpret the first 8 items as showing language DIF, which corresponds to the first peak. However, this also illustrates that if more than one peak appears in the path of the Gini index this indicates that a group of items measures another dimension, in this case the language ability. Whether this is a nuisance dimension and the items should be modified or excluded because they show DIF, or whether this is a second interesting dimension and this additional dimension should be modeled by means of a multidimensional IRT model, depends on the research question.

4.4 Method

To see whether the extension to multiple groups works adequately, a number of simulations were carried out. Every simulation was iterated 500 times and hit rates and false alarm rates were reported. Item parameters were taken from Wang et al. (2012). In each simulation 20 items were simulated. Item parameters are reported in the appendix B.1. Hit rates and false alarm rates were compared to a DIF analysis with a "perfect" anchor of four items. This means four items were randomly chosen as an anchor that are known to be DIF free. It is important to note that this would not be possible in an applied study as in practice it is unknown which items are DIF free and which are not, but in this simulation study the "perfect" anchor is used as an upper

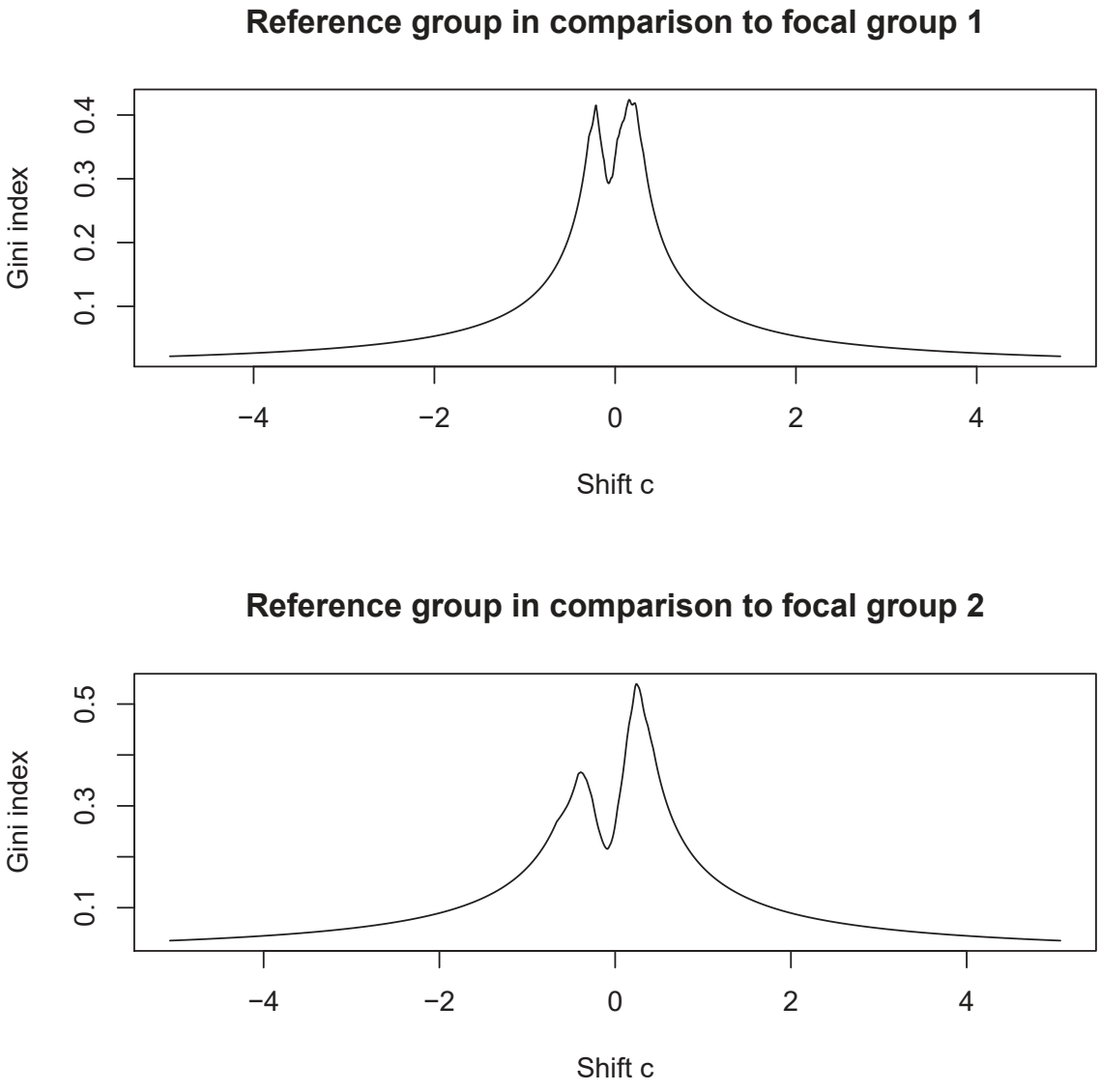

Figure 4.1: Search paths for the two focal groups from the toy example.

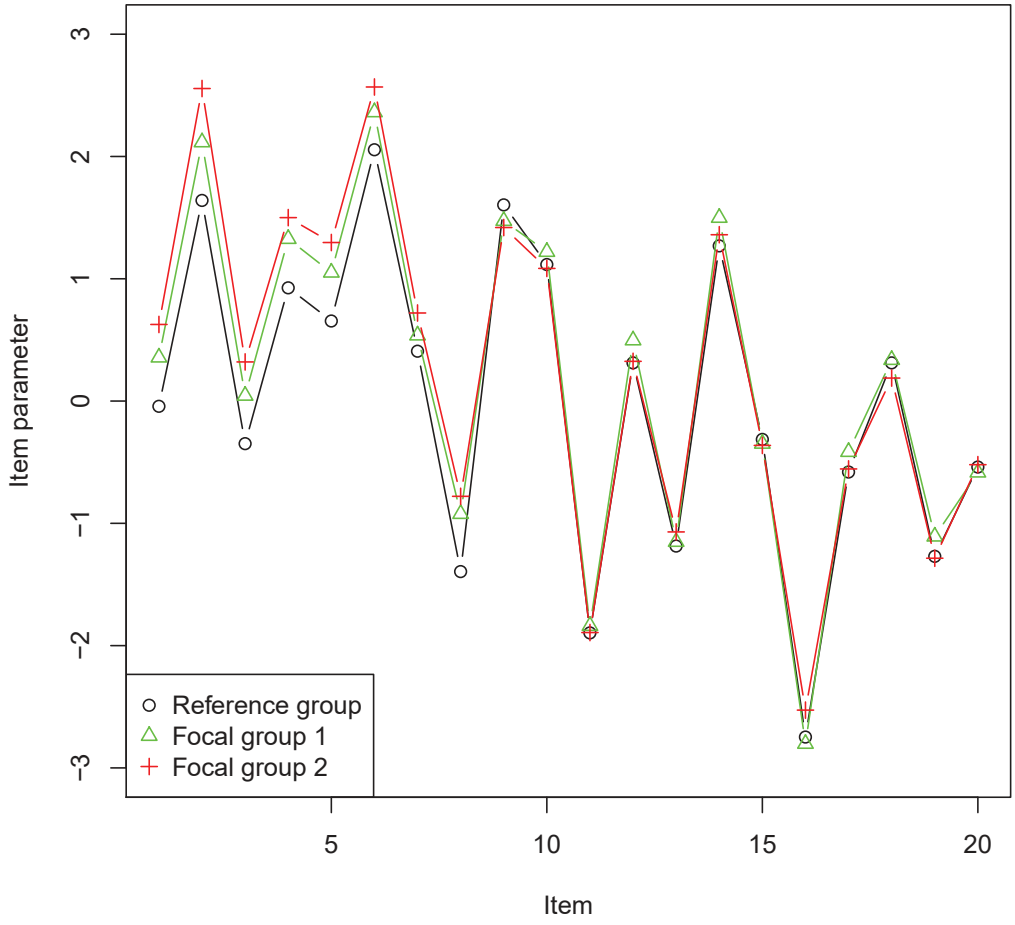

Figure 4.2: Item parameters according to the highest peak (global maximum).

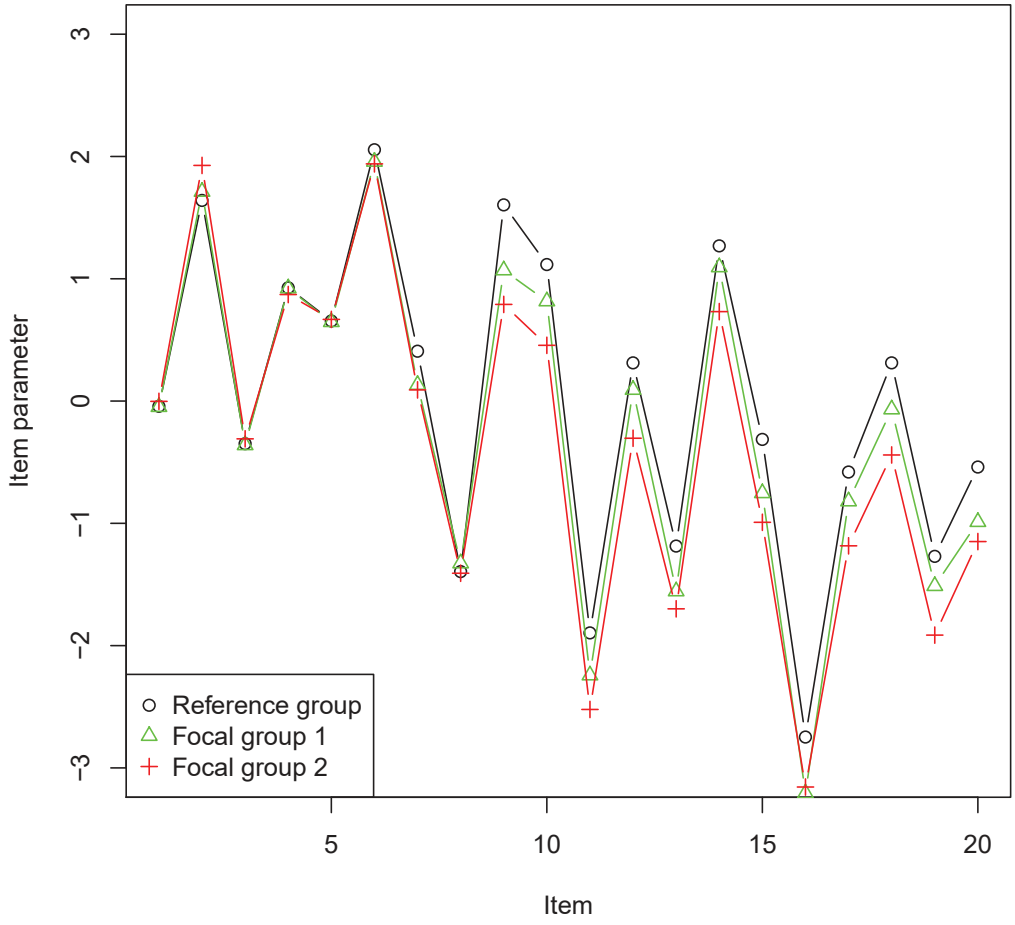

Figure 4.3: Item parameters according to the second highest peak (local maximum).

limit of performance.
In addition hit rates and false alarm rates were calculated for a DIF test with an equal mean anchor. An equal mean anchor describes the method of setting the mean of the estimated parameters of the groups equal. While equal mean anchoring is generally not advisable in scenarios with unbalanced DIF, it is often the default for software programs. This is why it is used here as a comparison method. Furthermore we compared our extension to classic anchoring methods. Kopf et al. (2015a) identified the *mean-p-value-threshold* (MPT) method to be advisable in two group scenarios. Huelmann et al. (accepted for publication) identified an aggregation rules to apply this anchoring method to a multi group scenarios, the so called *min-rule*. We therefore also implemented the MPT anchor with the *min rule* as a comparison.
Factors varied in the simulations were the percentage of DIF items (0.1, 0.3 or 0.4), the number of groups (3 or 6), the size of each group (100, 500 and 1000), and the DIF pattern. The DIF pattern was either balanced, meaning DIF effects in favor and against the reference group canceled each other out, or always favoring the reference group. Furthermore the size of the DIF effect was either set to a constant 0.6 or drawn from a normal distribution with an expectancy value of 0.6.
All simulations were done in R (R Core Team, 2019), using the packages psychotools (Zeileis et al., 2018) and quantreg (Koenker, 2008). Furthermore the code for the anchor point selection was taken from Strobl et al. (2019) and adapted towards multi group scenarios.

4.5 Results

A graphical overview of false alarm rates and hit rates of all simulations can be found in Figure 4.4 and 4.5. Even for the perfect anchor method slightly inflated false alarm rates were reported in some scenarios, especially in scenarios with a high DIF percentage. These inflated false alarm rates may be due to random fluctuation. But overall the inflated false alarm rates for the perfect anchor method are about 6% and can still be viewed as acceptable.
For the multiple anchor point selection a similar pattern emerged throughout all simulations. In small datasets the false alarm rates were inflated up to about 30%. But with increasing sample size the false alarm rates went down to an acceptable level slightly above the perfect selection. Only in few cases false alarm rates of 10 % were still reported (e.g. in the scenario with 1000 persons in each group). These findings are further investigated below.
The equal mean anchoring often performed conservative in simulations with a balanced DIF pattern in regard to the false alarm rates. False alarm rates

of about 3% were reported. However, in unbalanced designs the equal mean anchoring performed very poorly, with false alarm rates even rising along with the overall sample size. Note that inflated false alarm rates were to be expected for the equal mean anchoring method in these simulations, as its assumption of balanced DIF is violated. Since, however, one does not generally know a priori whether such an unbalanced DIF scenario is present, the equal mean method is not recommended.
Hit rates were comparable between the perfect selection and the anchor point selection. In some cases the anchor point selection was even outperforming the perfect anchor selection. However, keep in mind that the false alarm rates were also slightly higher for the anchor point selection and the very high hit rates are partially due to this. The equal mean anchoring performed consistently worse than the other two methods.
Hit rates and false alarm rates for the MPT anchoring method with the *min rule* were comparable to the hit rates and false alarm rates from the anchor point selection. The hit rates were usually a little bit lower for this anchoring method than for the anchor selection, but the false alarm rates were also slightly lower.
We further investigated single iterations of the simulation for the anchor point selection with inflated false alarm rates to explain these. As mentioned above, inflated false alarm rates occurred in small data sets. In Figure 4 you can see that false alarm rates were particularly high in the setting with six groups, a constant DIF effect always favoring the reference group, and a high DIF percentage of 0.4. We chose the 100th iteration of the simulation for our illustration as it seemed to be a "typical" iteration (meaning it had a high false alarm rate of 41.7% and a hit rate of 75%). We investigated the search paths for this particular iteration. The search paths can be found in Figure 4.6. In nearly all comparisons two clearly distinct peaks can be easily identified, the first occurring in all slightly before zero and the second slightly after zero.
The location of the first peak in the search paths corresponds to the anchor point where the first eight items interlock, and the location of the second peak corresponds to the point where the last twelve items interlock. Just like in our toy example described earlier, in this setting with a constant DIF effect always favoring the reference group and a high DIF percentage of 0.4, the first eight items all have the same amount of DIF in the same direction and could be considered as a subset of the test that measures a different aspect or dimension (such as verbal ability).
The two solutions, the first eight items interlocking and the last twelve showing DIF, or vice versa the last twelve items interlocking and the first eight items showing DIF, are mathematically equivalent due to the scale indeter-

minancy. In a practical research situation, the two-peaked form of the search paths would be very informative and content considertaions could guide the decision whether and which part of the items should be considered as DIF items or whether the test should be analyzed by means of a two-dimensional model.

The simulation design, however, was motivated by previous simulation studies for anchoring methods that rely on the assumption that the majority of items is DIF free. In this design, the majority of the last twelve items is defined to be DIF free, and the minority of the first eight items is defined to have DIF. Therefore, only when this solution (corresponding to the second peak in the search path) is found by the anchoring method the results will show a low false alarm rate, as explained below. If, however, the other possible solution (corresponding to the first peak in the search path) is selected, where the first eight items interlock and the last twelve do not, this results in a high false alarm rate because those items defined as DIF free will be labeled as having DIF and vice versa.

To illustrate how this affects the simulation results, we consider Figures 4.7 through 4.9. Figure 4.7 shows the item parameters for the comparison of focal group 4 to the reference group. Figure 4.6 showed that for focal group 4 the highest peak is clearly in the location where the last twelve items largely interlock and the first eight show DIF, just like DIF was defined in the simulation design. If we look at a focal group whose highest peak was in the other position, for example focal group 1, we see a different pattern. As illustrated in Figure 4.8, when the highest peak is used, the item parameters for the comparison of focal group 1 to the reference group largely interlock for the first eight items and show DIF for the last twelve items. This is opposite to how DIF was defined in the simulation design and will lead to a high false alarm rate, because the final DIF test will label an item as a DIF item as soon as it shows DIF in any group-comparison.
If, however, by visual inspection and content considerations - like it could be done in practical research - the other peak is picked, we see in Figure 4.9 that now also for focal group 1 the last twelve items largely interlock and the first eight show DIF, which matches the simulation design. In the case presented here, simply by visually identifying the positions of the second peaks for all focal groups and analyzing the data with the corresponding anchor point, we could bring the false alarm rate down to 8.3% while the hit rate remained at 75%.

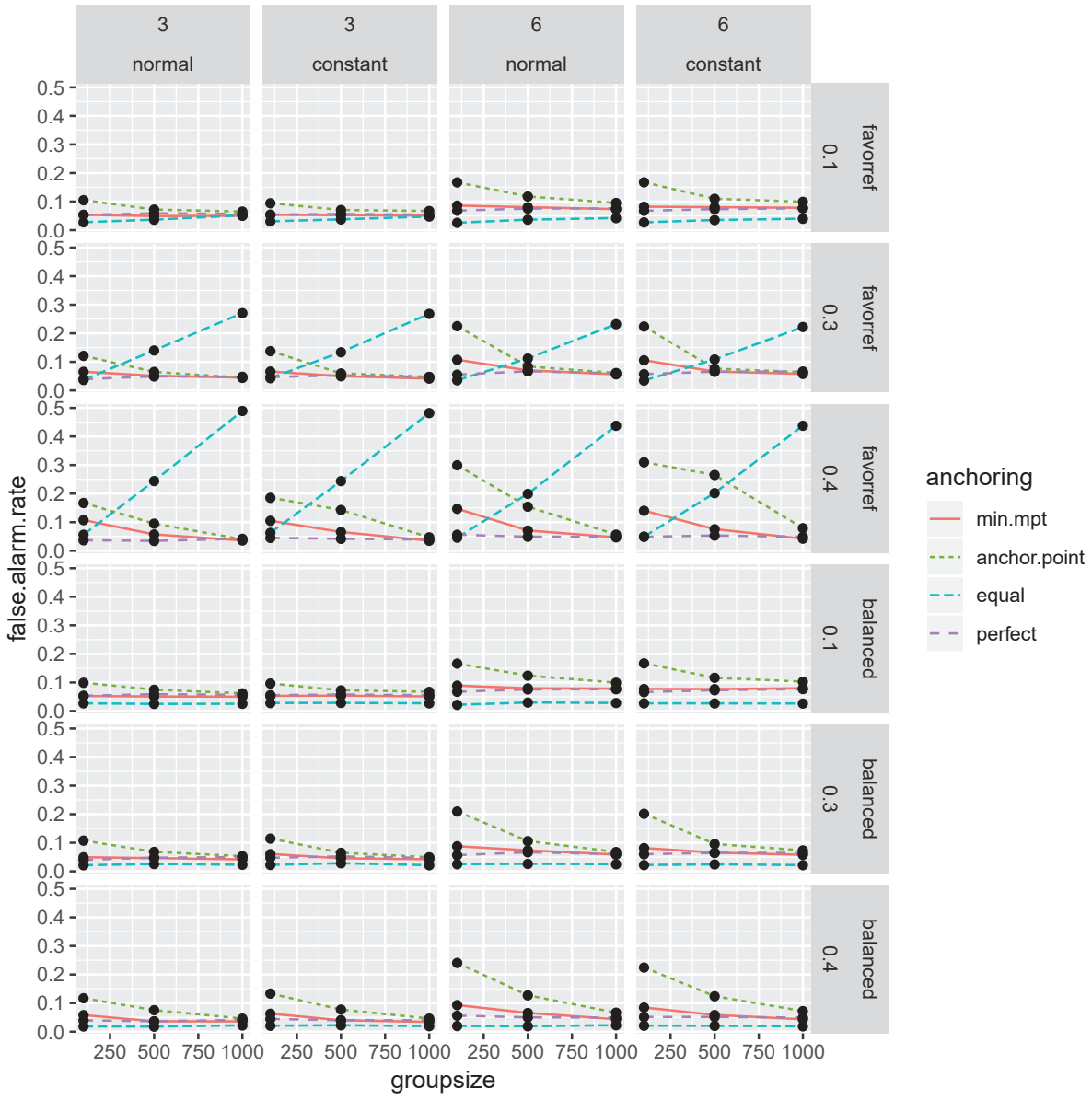

Figure 4.4: Overview of false alarm rates.

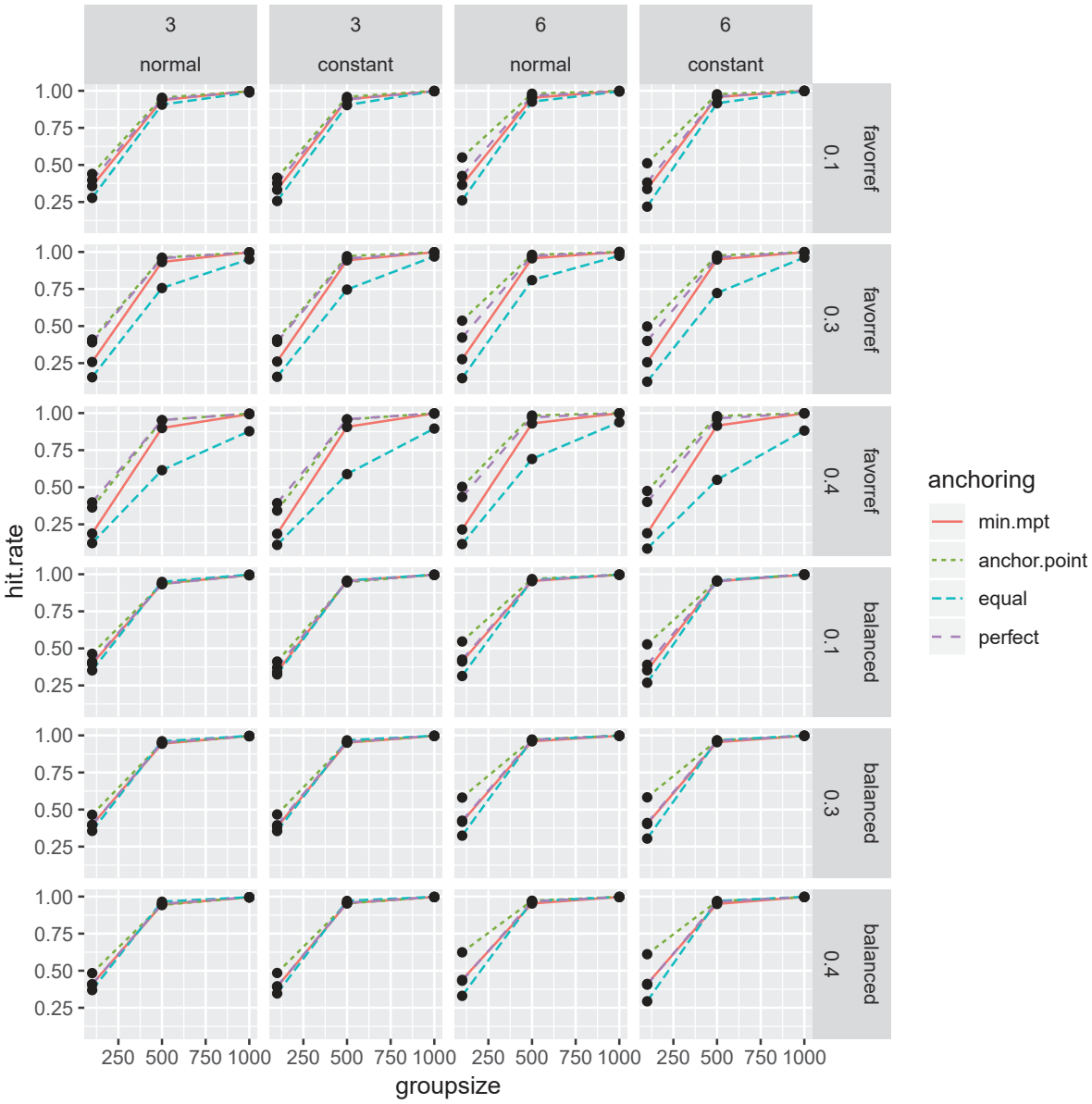

Figure 4.5: Overview of hit rates.

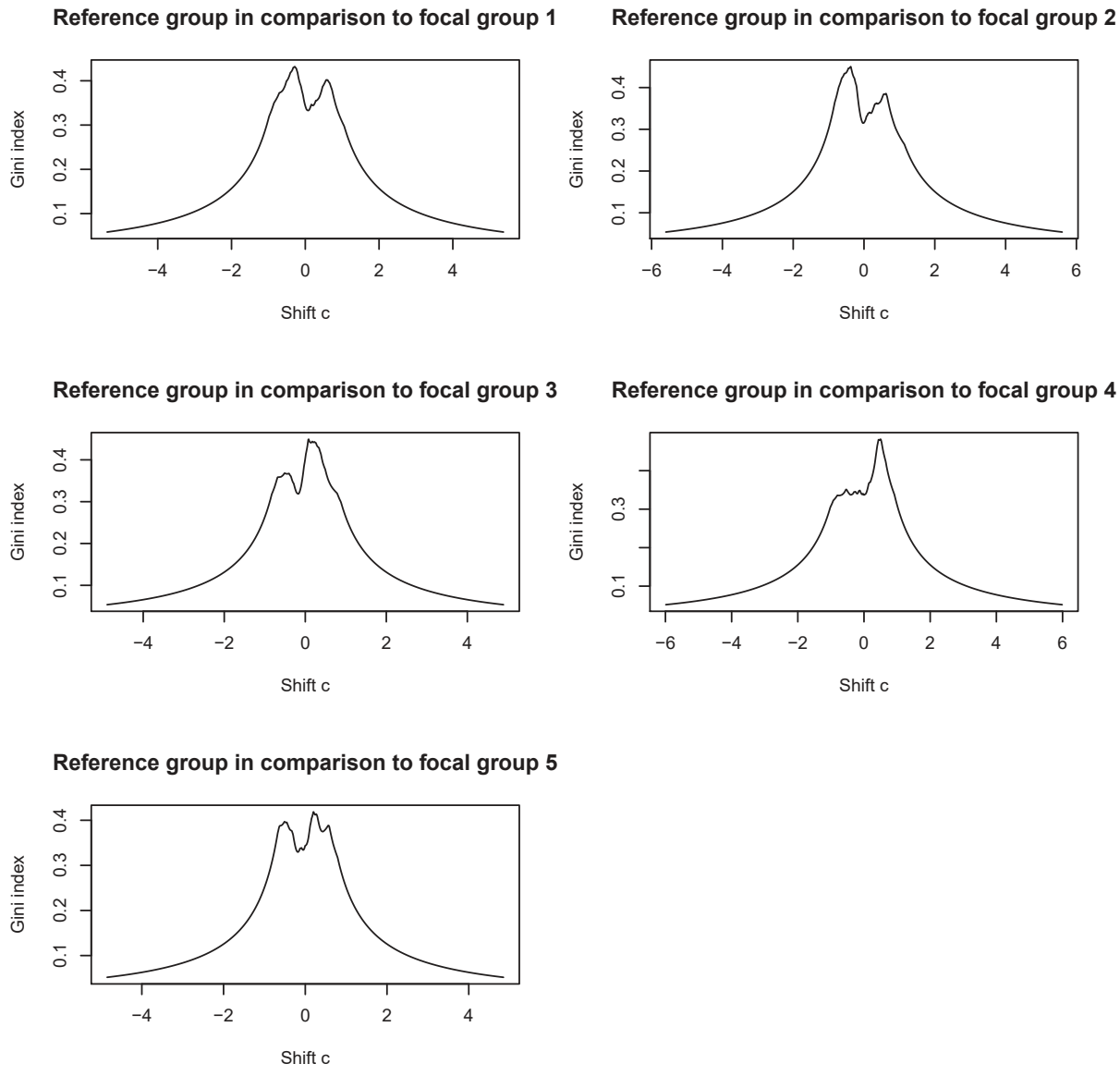

Figure 4.6: Search path for a single iteration from the setting with six groups, a constant DIF effect always favoring the reference group, and a high DIF percentage of 0.4.

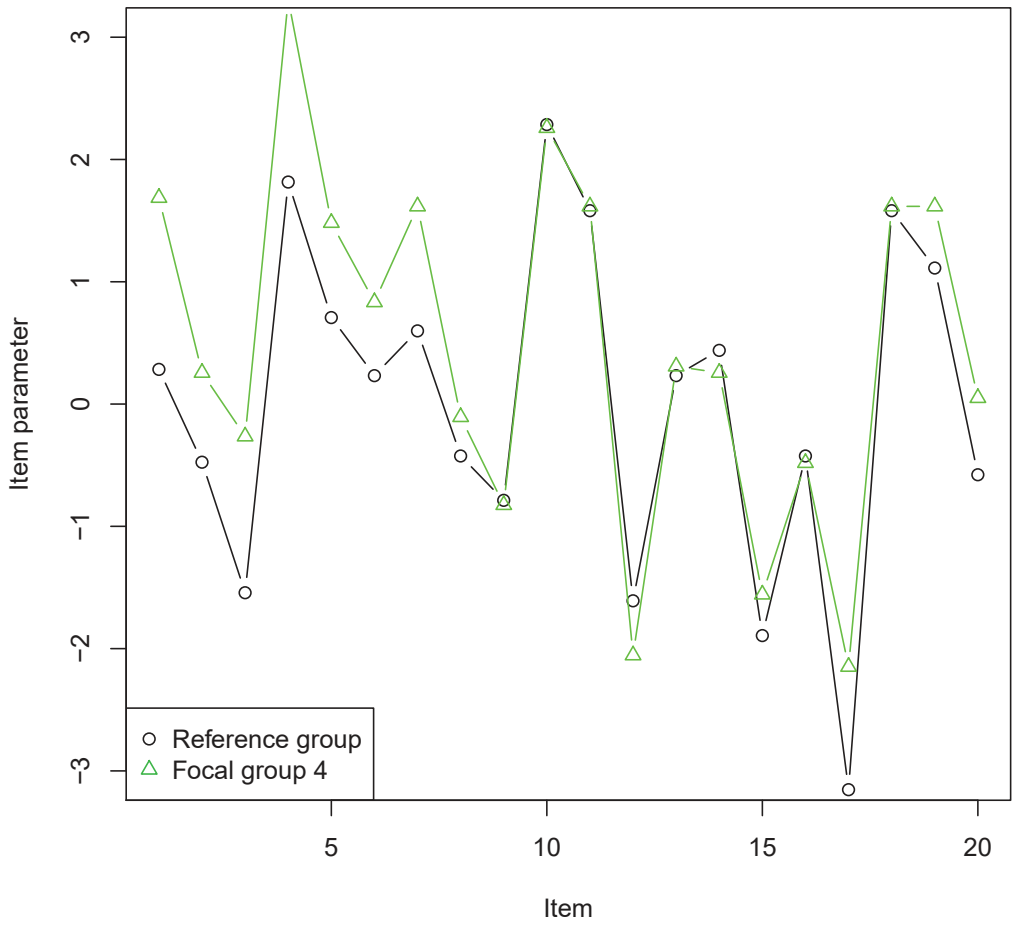

Figure 4.7: Item parameters for focal group 4 according to the highest peak (global maximum).

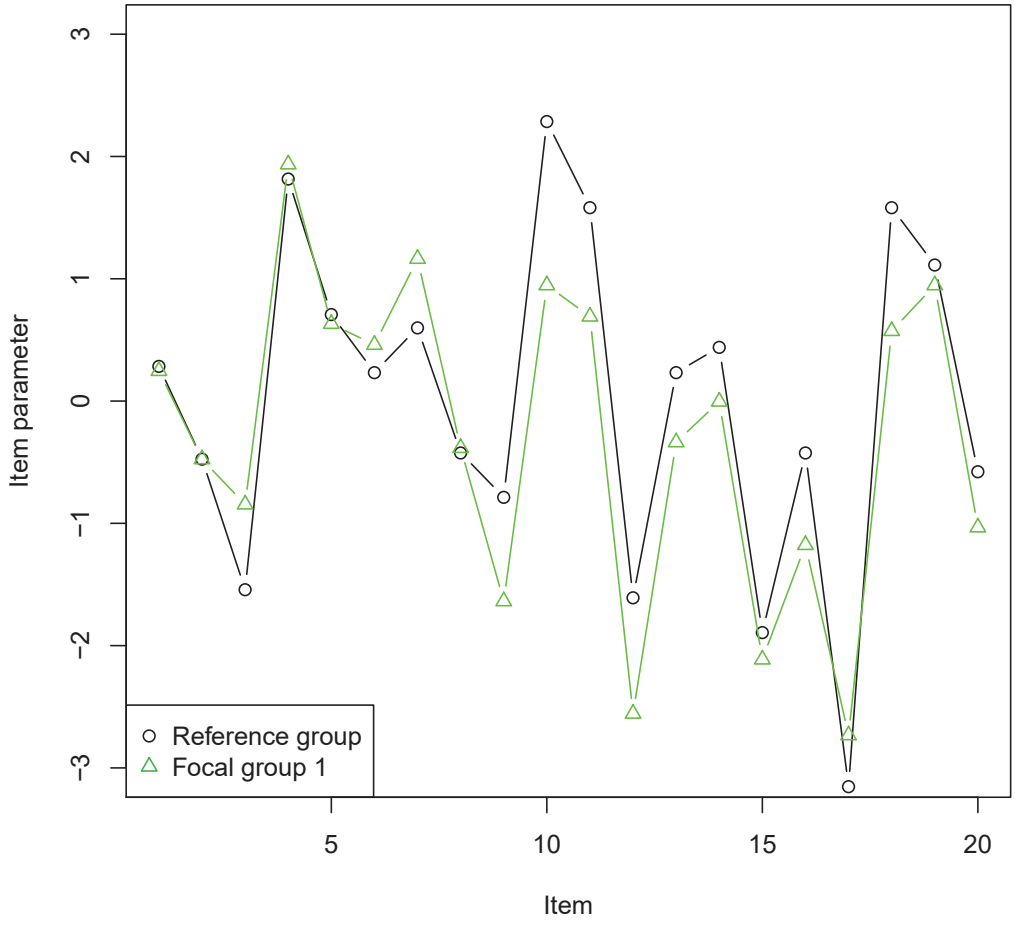

Figure 4.8: Item parameters for focal group 1 according to the highest peak (global maximum).

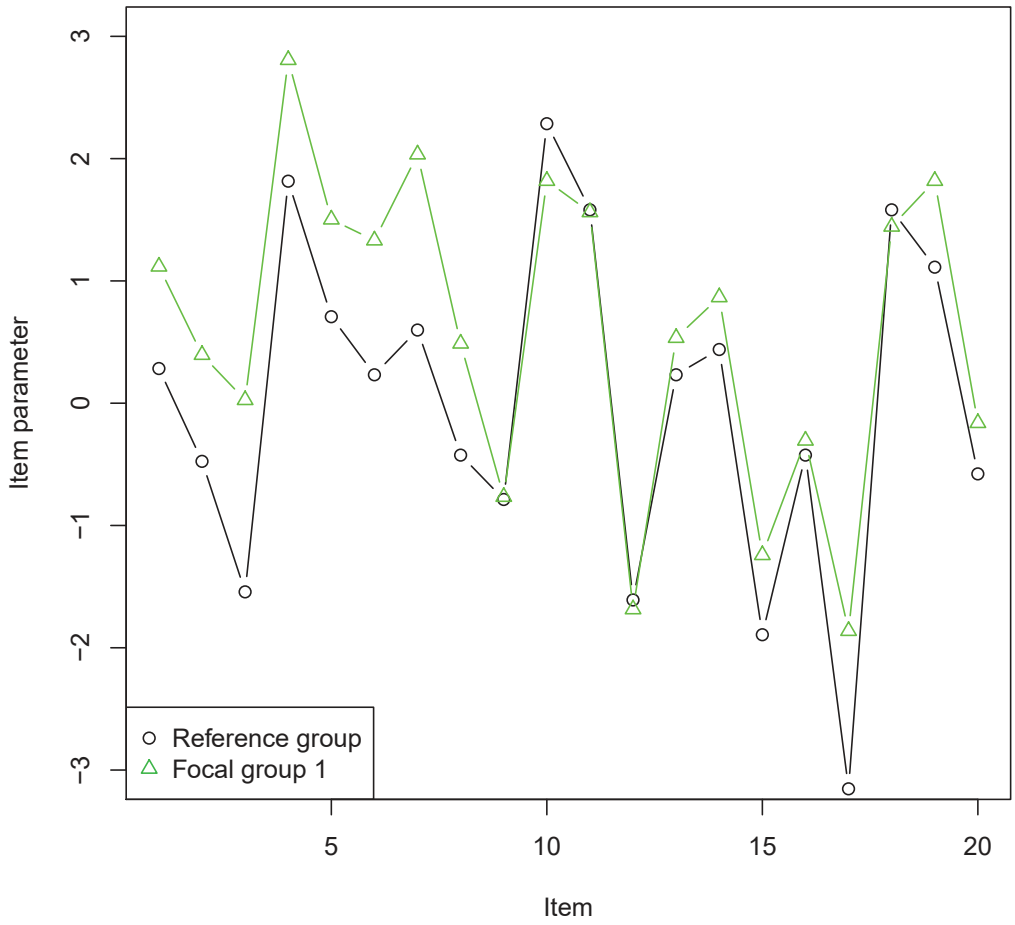

Figure 4.9: Item parameters for focal group 1 according to a visual inspection.

4.6 Discussion

DIF detection is an important issue as it relates to fairness of tests. As a prerequisite of DIF detection, some form of anchoring is needed. Especially in multi group scenarios this anchoring process can be quite demanding. Strobl et al. (2019) suggested a method for two group scenarios that does not rely on an a priori defined anchor, but is looking for one or more optimal anchor points based on an inequality measure. In this study we expanded this idea to multi group scenarios. We compared this expansion to an equal mean anchoring method. Equal mean anchoring is the standard default in widely used statistical software for DIF detection. Our results showed that overall our extension outperforms the equal mean anchoring. While the equal mean anchoring performed well in balanced scenarios, it failed completely in unbalanced scenarios, as false alarm rates were even rising with rising sample size. Therefore equal mean anchoring is not recommended. The anchor point selection on the other hand performed consistently well in all simulations. Only in small datasets it showed highly inflated false alarm rates, but with rising sample size the false alarm rates decreased.
Moreover, we could show that with an informed choice of the anchor point through a simple visual inspection of the search path, false alarm rates could be decreased immensely for the anchor point selection. In future studies we will investigate how cases like the one in our toy example, where the fact that there are multiple peaks in the search path is very informative from a content point of view, can be better reflected in the simulation results.
The MPT anchoring method also performed well in all scenarios. Hit rates tended to be slightly lower than hit rates from the anchor point selection, but the false alarm rates were also lower than those reported without inspecting the search path.
In this study we only optimized the anchor points in a way where all focal groups were only compared to the reference group and no comparison between the focal groups has taken place. We did this because the test statistic used here does not take these differences into regard anyway, since the usual contrast matrix defines the comparisons made. Furthermore, the test statistic is independent of the choice of the contrast matrix. Therefore a different contrast matrix that would take different differences into account would result in the exact same test statistic (at least under very broad assumptions). In future research it could be mathematically investigated whether by optimizing only in regard to the reference group the differences between the focal groups are already being optimized.

Chapter 5

Discussion

In this chapter the most important results of this thesis will be summarized. Furthermore the limits of this work are discussed and future research questions are presented.

5.1 Summary of the most important results

The identification of DIF items is a necessity for fair testing. The scope of this work in general is therefore to give researchers new methodological tools to more accurately ensure that items are DIF free. We specialized on multiple group scenarios because up to now there is only little methodological research, despite the fact that multiple groups are common, for example in large scale assessments like PISA.
In the second chapter we focused on the development of aggregation rules for multiple group DIF analysis. These aggregation rules are necessary when anchoring methods developed for two group scenarios shall be applied to multiple group scenarios. We compared three aggregation rules: the *min-*, the *mean-*, and the *all-rule*. Furthermore we investigated a direct approach. We showed that the *min-rule* is generally advisable. We furthermore showed that the *all-rule* can produce similar results and that the decision between these rules is also a philosophical question on how DIF should be treated. If no amount of DIF is acceptable, the *min-rule* is more advisable. But the *min-rule* can be too sensitive when the number of groups is high. The *all-rule* is less sensitive in these situations, but this comes at the risk of overlooking DIF that only affects single groups. The *mean-rule* is generally not advisable. While using a mean to aggregate seems like the most obvious idea, in the case of anchoring methods usually this mean is taken over p-values, which is generally not advisable. The direct approach showed similar results

to the *all-* and *min-rule* in many situations. But it also critically failed in some situations. We did not have a conclusive explanation for this behaviour and would therefore also not recommend the direct approach in practice.
The third chapter is focused on the development of a rule of thumb to determine optimal group ratios in multiple group DIF scenarios. We developed our rule of thumb in an iterative way. We first used a naive rule of thumb, which was to sample equal amounts of persons from each group. We could show that this procedure does not fail, but leaves room for optimization. We then developed a rule of thumb that relies on a minimum amount of information on the DIF effect. In the tested scenarios, this rule of thumb always delivered acceptable and in many scenarios very good results. We can therefore generally recommend to use this new rule of thumb as it usually outperforms the naive rule of thumb. We also showed simulations, where the new rule of thumb did not fail, but left room for optimization. New rules of thumb could be developed specifically for these scenarios.
In the fourth chapter we focused on the extension to multiple groups of the anchor point selection method. Anchor point selection is an approach suggested by Strobl et al. (2019) for two group scenarios that does not rely on an a priori defined anchor, but is looking for one or more anchor points based on an inequality measure. We could show that our extension clearly outperforms the usual default anchoring technique in statistical software (equal mean anchoring). Furthermore we compared the extended anchor point selection to an anchoring method that is generally advisable according to Kopf et al. (2015a) combined with the *min-rule* as an aggregation rule. We already showed in chapter 2 that this aggregation rule is the most advisable. Hit rates and false alarm rates were comparable between these two methods. The anchor point selection usually showed slightly higher hit rates, but at the cost of slightly higher false alarm rates. We furthermore showed that by using their content knowledge researchers can greatly improve false alarm rates for the anchor point selection by inspecting the search path for the optimal anchoring point.

5.2 Limits of this work

The generalization and optimization methods for multiple group DIF scenarios discussed in this thesis open up new possibilities for researchers. But due to the fact that these methods are the first of their kind, some limitations may apply. First of all, it is not possible to test every possible DIF effect in a simulation study. We therefore carefully chose DIF effects that cover a wide array of realistic situations. But of course there will always be specific situ-

ations we did not cover in this work. Furthermore we only analyzed uniform DIF in the Rasch Model. It was not possible for us to include, for example, non-uniform DIF and higher-order IRT models like the 2PL model, without reducing the simulation designs in other ways. Therefore we concentrated on uniform DIF in the Rasch model, as the Rasch model is widely used. For similar reasons we also concentrated on the Lord's χ^2 test as a test for DIF. The topic of this thesis is on anchoring and sampling methods and not on test statistics. We compared different test statistics in a small pre-study for the article presented in chapter 2, but it seemed like there were little to no interactions between the tested methods and the choice of the test statistic (meaning an anchoring method that worked well with one test worked also well with other tests). We therefore believe that the reported effects of anchoring and sampling methods also hold true when other test statistics are used. But this still needs to be confirmed by extensive research.
In the second chapter we focused on aggregation rules. We investigated three aggregation rules. As there are infinite possibilities of aggregation rules, we had to limit this study on a finite set. We chose the three rules as they seemed to be the most obvious choices. But there may be other aggregation rules that work adequately that we did not consider. Furthermore, we made the decision to translate anchoring methods to be applicable in multiple group scenarios. We therefore made the decision that we wanted to define the same set of anchor items for all comparisons between the reference group and the focal groups. Due to time and space limitations we could not investigate anchoring techniques that use a specific set of items as anchor items for every single comparison. In the fourth chapter we extended the anchor point selection to multiple groups. In a way, this can be seen as specific sets of anchor items for every comparison. But the methods used were too different to make direct comparisons.
In the third chapter we developed a rule of thumb to determine group ratios that give good hit rates and false alarm rates in various situations. Again due to time and space limitations only a finite set of situations could be investigated. Therefore specific situations where the rule of thumb fails might not be covered.
In the fourth chapter we extended the anchor point selection to multiple groups. We used the Lord's χ^2 test to detect DIF. This test uses a contrast matrix to define comparisons. Usually the contrast matrix is defined in a way to compare the reference group to every focal group. We therefore also only tried to optimize these comparisons. If another contrast matrix would be used, other comparisons would be made and therefore other comparisons should be optimized. We chose to optimize only for this contrast matrix, as there a numerous possible contrast matrices, but usually only the con-

trast matrix comparing the reference group to every single focal group is used. Furthermore, as we explain in more detail above, the resulting test statistic is independent of the contrast matrix. Therefore we believe, that the test statistic is also independent of the choice of which comparisons are optimized, as long as this choice corresponds to a contrast matrix that involves all groups. More research is needed to substantiate mathematically this belief.

5.3 Future research questions

As already mentioned in the section on limits of this work, we are only able to cover a finite set of DIF effects and more generally also a finite set of simulation settings. We tried to cover situations that seemed realistic to us. In future research we would like to focus on further simulation settings that are more specialized. Furthermore we also would like to generalize our findings to more complex models than the Rasch model. Also we would like to address non-uniform DIF and polytomous items in future research.
In the second chapter we took a closer look at the anchoring method *next candidate*. The method was not performing as well as we expected, but we could show that this method still has great potential. The method produces an anchor in an iterative way. We showed that the anchor selection based on this method is often good, but when in one of the first iterations a DIF item is picked as an anchor item, the method fails completely. Future research could investigate how this method could be improved. This could be done by choosing the first items more carefully. Another way would be to calculate multiple anchors based on subsets of the data and then merge the anchors. This "bootstrapped" fashion of anchoring should stabilize the choice of anchor items.
In the third chapter we developed a rule of thumb that works with a minimal amount of information. If more information is available, other rules of thumb might be more appropriate. For example when ordinal information is available on the DIF effect (meaning the order of DIF effects is known, but not necessarily their exact size), a rule that focuses only on the biggest differences could lead to better results. Therefore more research is needed. We only briefly mentioned the possibility of simulations to determine optimal group ratios in the third chapter. These simulations usually rely on a lot of information. As we focused on methods that would be applicable with minimal information, a closer look at simulations is not covered in the chapter. But future research might show the potential of simulations for these

situations.
The fourth chapter is concerned with the extension of the anchor point selection to multiple groups. We only optimized this extension for the comparison of the reference group to every focal group. But technically, all other comparisons, that relate to the same null hypothesis would give the same test statistic (at least with Lord's χ^2 test). For example, when the reference group is compared to the first focal group, the first focal group to the second, and so forth, the test statistic would be identical to the test statistic where the reference group is compared to every focal group. This is due to the independence of contrast matrix and test statistic. We believe, therefore, that while we only optimized for the comparison between reference group and every focal group, we do not need to consider other comparisons as these other comparisons would lead to the same test statistic. Future research is needed to mathematically prove this belief.
Overall this thesis has to be seen as a first step in generalizing common anchoring methods and optimizing sampling methods for multiple group scenarios. Therefore it also opens up numerous future research questions and we could only present a small sample of these here.

Bibliography

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255–278.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. doi:10.1080/10705511.2014.919210. eprint: https://doi.org/10.1080/10705511.2014.919210

Awour, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based monte carlo study with SIBTEST and mantel-haenszel procedures.*

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340.

Berberoglu, G. (1995). Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation*, *21*, 439–456.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253–260.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: Extracts from variabilità e mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, *10*(3), 421–443. Retrieved from https://EconPapers.repec.org/RePEc:kap:jecinq:v:10:y:2012:i:3:p:421-443

Chang, Y.-W., Huang, W.-K., & Tsai, R.-C. (2015). DIF detection using multiple-group categorical CFA with minimum free baseline approach. *Journal of Educational Measurement*, *52*(2), 181–199.

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.

Craig, B. (2017). *The empirical selection of anchor items using a multistage approach* (Doctoral dissertation, University of South Florida).

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19–29.

Firth, D. (2003). Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology*, *33*(1), 1–18. doi:10.1111/j.0081-1750.2003.t01-1-00125.x. eprint: https://doi.org/10.1111/j.0081-1750.2003.t01-1-00125.x

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments and applications*. New York: Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In *Rasch models: Foundations, recent developments and applications* (pp. 69–96). New York: Springer.

Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement*, *62*(1), 32–44.

Huelmann, T., Debelak, R., & Strobl, C. (accepted for publication). A comparison of aggregation rules for selecting anchor items in multi group DIF analysis. *Journal of Educational Measurement*.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kilmen, S. (2016). Effect of DIF magnitudes, focal group sample size, and DIF ratio on the performance of SIBTEST. *International Journal of Social Sciences and Education*, *6*(1), 91–98.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261–276.

Koenker, R. (2008). *quantreg: Quantile regression*. R package version 5.41.

Kopf, J. (2013). *Model-based recursive partitioning meets item response theory: New statistical methods for the detection of differential item functioning and appropriate anchor selection* (Doctoral dissertation, University of Munich).

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

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

Koretz, D. M., & McCaffrey, D. F. (2005). *Using IRT DIF methods to evaluate the validity of score gains*. National Center for Research on Eval-

uation, Standards, Student Testing (CREST); Center for the Study of Evaluation (CSE); Graduate School of Education, and Information Studies University of California.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9).

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second Edition). London: Chapman and Hall.

Meade, A. W., & Wright, N. A. (2012). Solving the measuremant invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Second Edition). Thousand Oaks, California: SAGE.

Pohl, S., Stets, E., & Carstensen, C. H. (2017). *Cluster-based anchor item identification and selection* (NEPS Working Paper No. 68). Bamberg: Leibniz Institut for Educational Trajectories, National Educational Panel.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Ree, M. J. (1993). Foreword: Differential Item Functioning (DIF): A perspective from the air force human resources laboratory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. xi–xii). Hillsdale: Lawrence Erlbaum Associates.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05/

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67.

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., von Oertzen, T., Zeileis, A., & further Authors. (2019). *Manuscript under preperation*.

Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation*, *17*(6), 465–481.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412.

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

Wang, W.-C. (2004). Effects of anchor item methods on the detection of Differential Item Functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of Differential Item Functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.

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

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2018). *psychotools: Infrastructure for psychometric modeling*. R package version 0.5-0. Retrieved from https://CRAN.R-project.org/package=psychotools

Zieky, M. J. (2016). Fairness in test design and development. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assesment and measurement* (pp. 9–32). New York: Routledge.

Appendix A

Supplementary material: Chapter 2

A.1 Hit Rates and False Alarm Rates in the Six Group Case

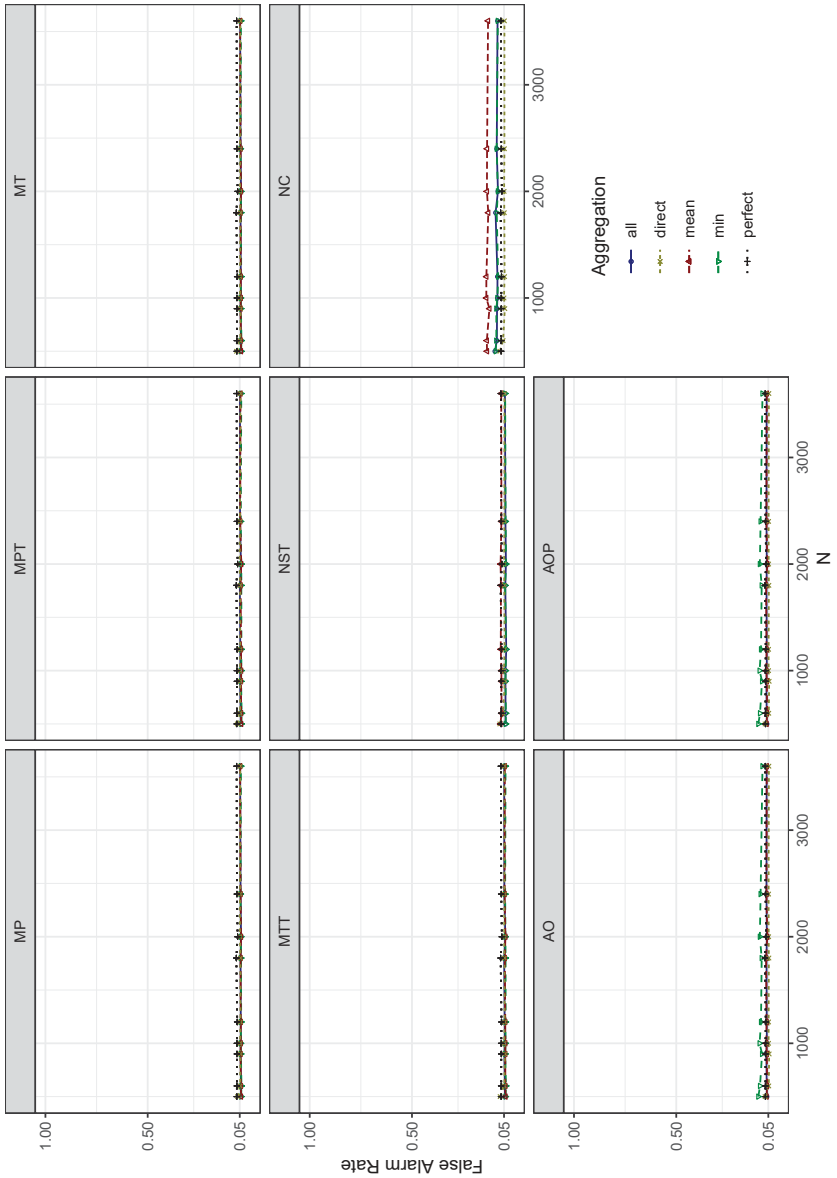

Figure A.1: False alarm rates in the six group sceanrio without any DIF

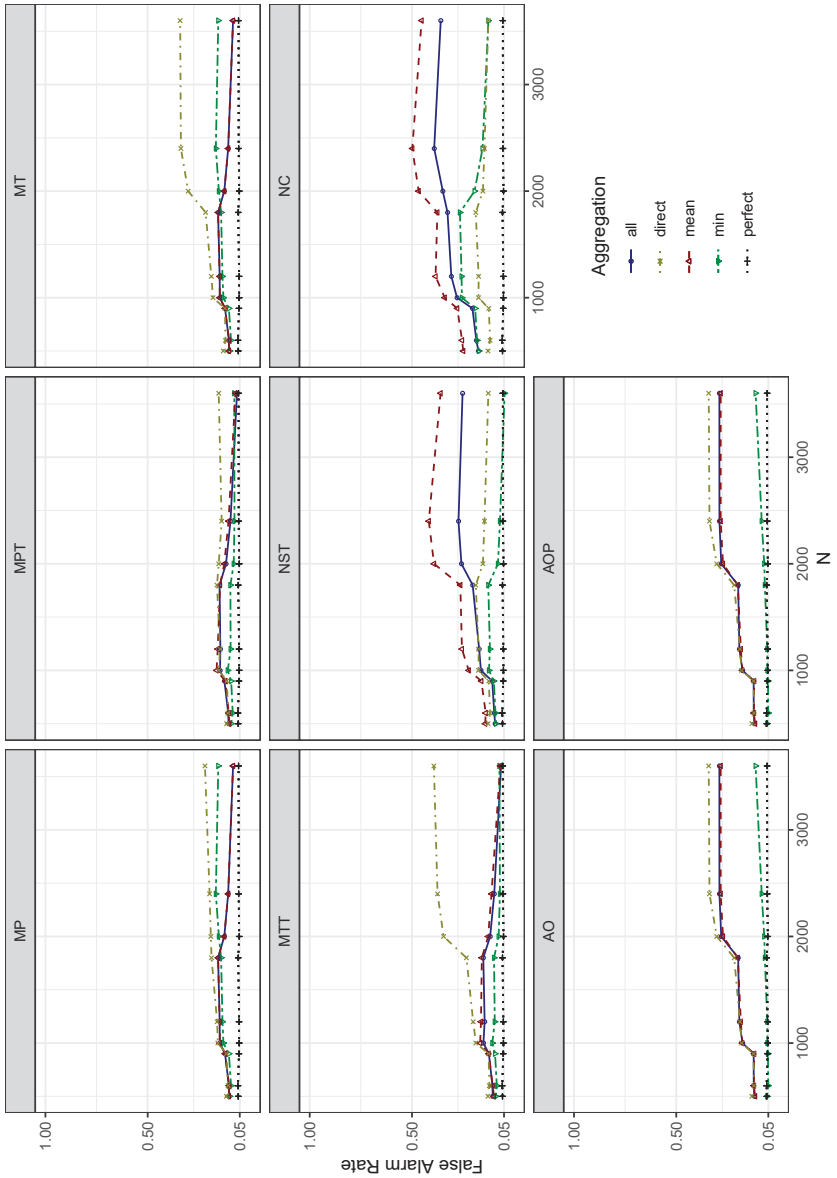

Figure A.2: False alarm rates in the six group sceanrio with all DIF items favoring the reference group

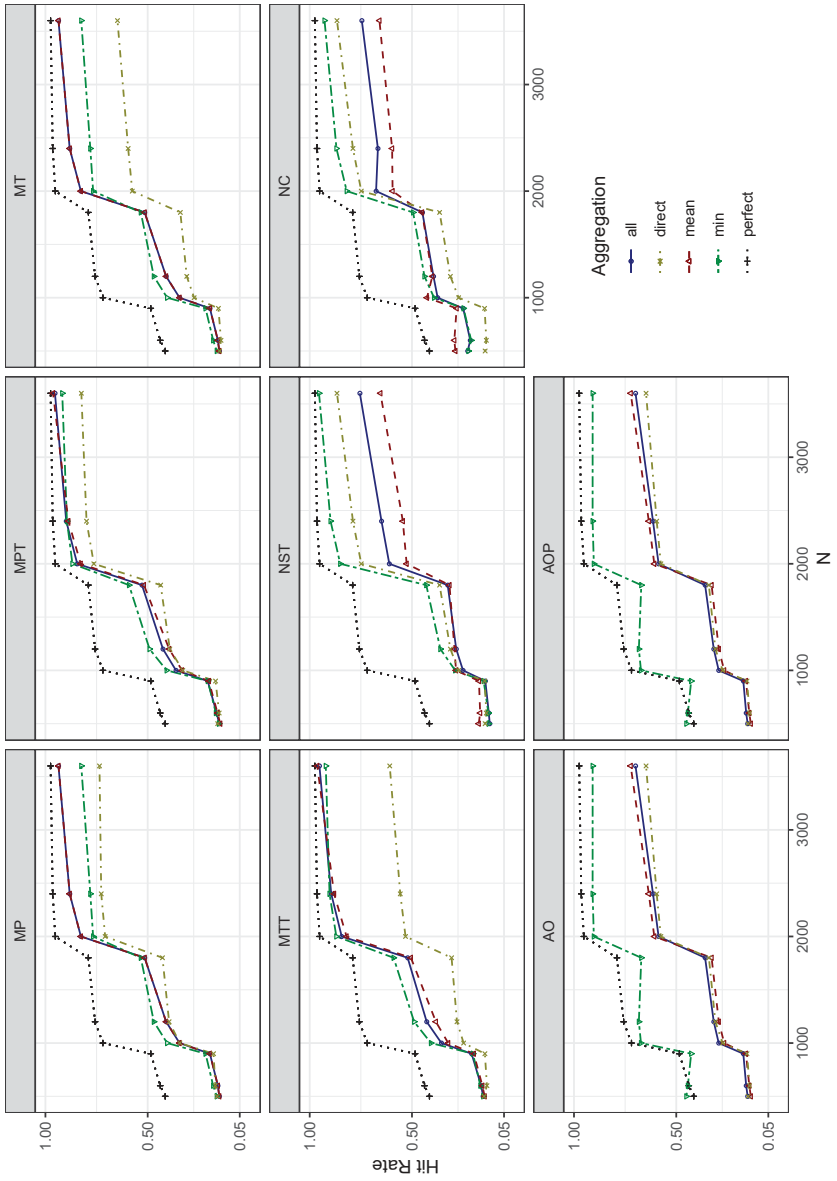

Figure A.3: Hit rates in the six group sceanrio with all DIF items favoring the reference group

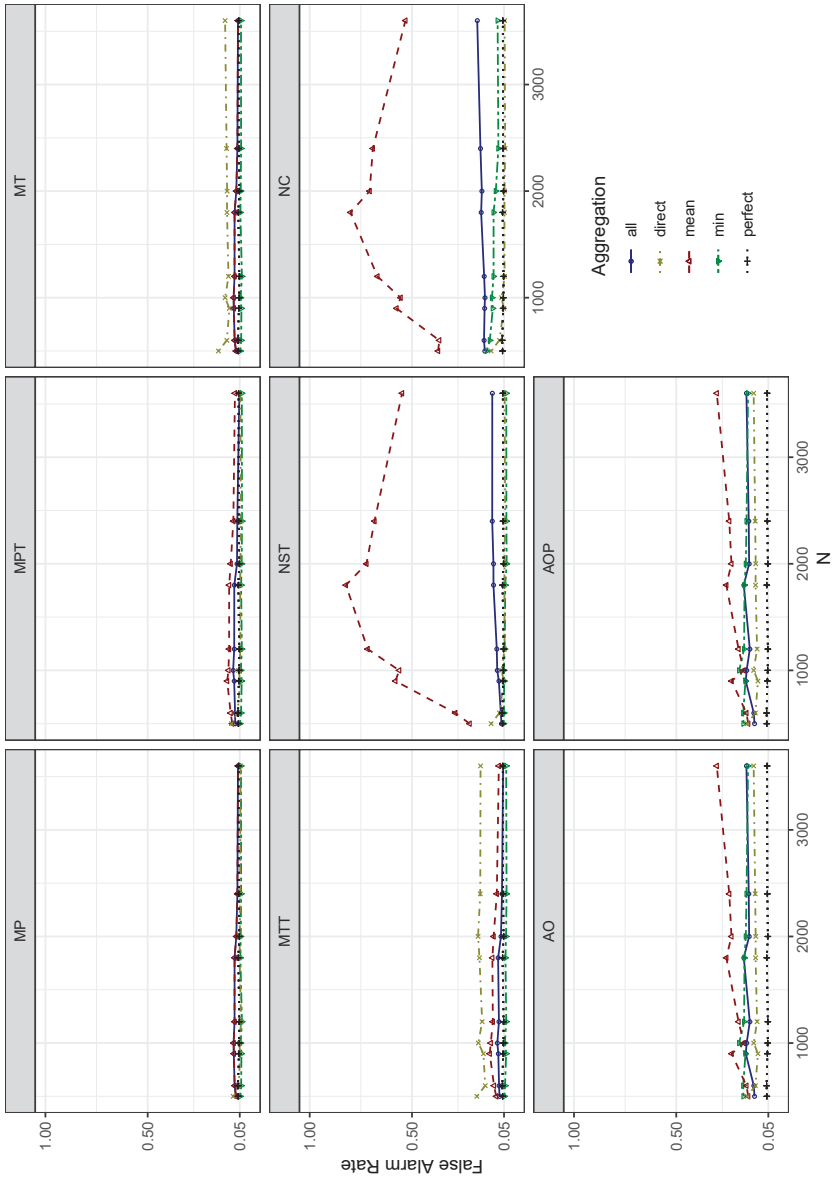

Figure A.4: False alarm rates in the six group sceanrio with half of the DIF items favoring the reference group

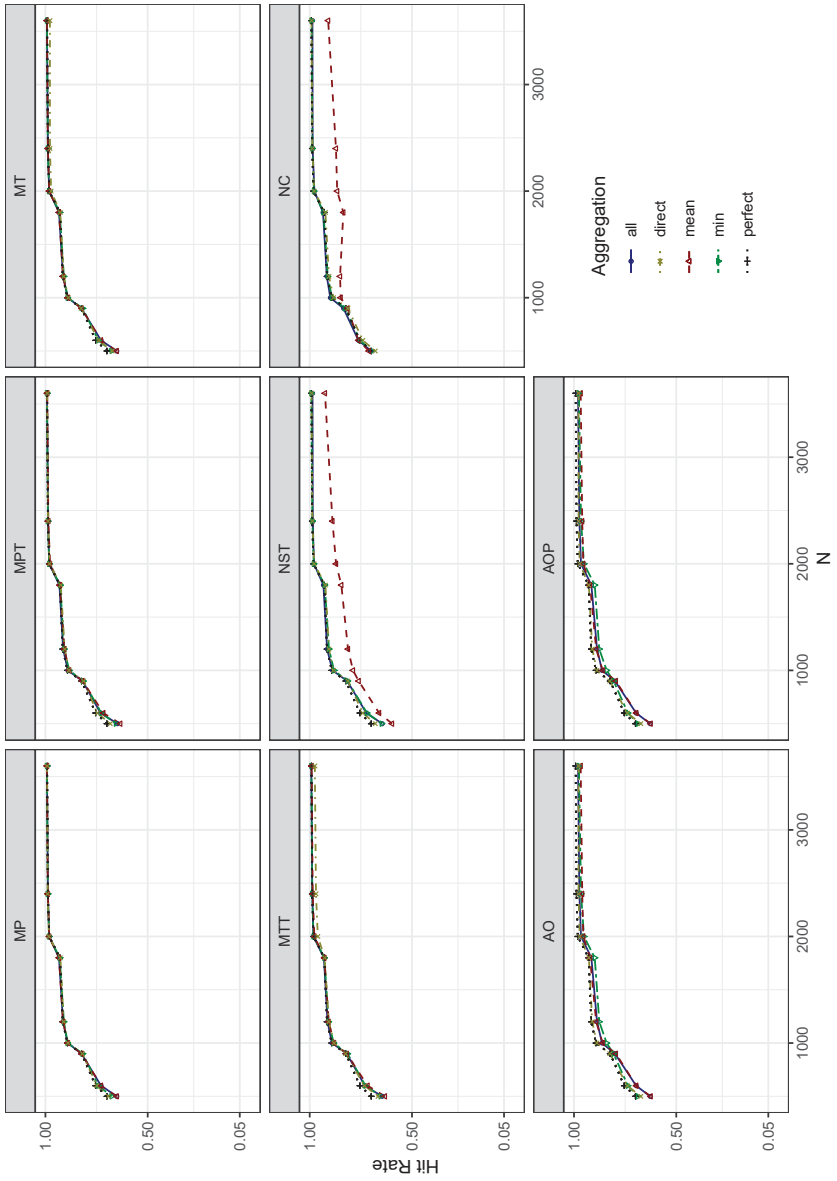

Figure A.5: Hit rates in the six group sceanrio with half of the DIF items favoring the reference group

A.2 Explanation for the limited number of comparisons possible with the generalized Lord's χ^2 test

Kim et al. (1995) define which comparisons are made in the generalized Lord's χ^2 test through a contrast matrix C and the vector of item parameters $\vec{\xi}$ containing the item difficulties α_{ij}, with i denoting the item and j denoting the group. In a three group scenario with only one item, $\vec{\xi}$ would look like this:

$$\vec{\xi} = \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{13} \end{pmatrix}$$

The contrast matrix C could be specified like this:

$$C = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

For the contrast matrix C also other matrices could be considered, as long as it has a full row rank. Which comparisons are made can then be seen by the multiplication of $\vec{\xi}$ and C:

$$\begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{13} \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} \alpha_{11} - \alpha_{12} \\ \alpha_{11} - \alpha_{13} \end{pmatrix}$$

Note that in this case α_{11} is compared to α_{12} and α_{13}. By choosing a different contrast matrix C α_{11} and α_{12} could be compared and simultaneously α_{12} and α_{13}, for example. But if it is desired to compare α_{11} to α_{12} and α_{13} and to simultaneously compare α_{12} and α_{13} a contrast matrix like this is needed:

$$C = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

This matrix, however, no longer has a full row rank and can therefore not be used for the generalized Lord's χ^2 test.

A.3 Items of the SPISA dataset

Table A.1: Items from the SPISA dataset on the topic of politics and whether they were flagged as DIF items (D) by one of the aggregation rules or chosen as an anchor item (A).

Item	Question	Answer	Setting			
			direct	all	min	mean
1	Who determines the rules of action in German politics according to the constitution?	The Bundeskanzler (federal chancellor)	D	D	D	D
2	What is the function of the second vote in the elections to the German Bundestag (federal parliament)?	It determines the allocation of seats in the Bundestag	0	0	0	0
3	How many people were killed by the RAF (Red Army Faction)?	33	0	0	0	0
4	Where is Hessen (i.e., the German federal country Hesse) located?	Indicate location on a map	0	0	0	A
5	What is the capital of Rheinland-Pfalz (i.e., the German federal country Rhineland-Palatinate)?	Mainz	D	D	D	D
6	Who is this?	Horst Seehofer	D	D	D	D
7	Which EU institution is elected in 2009 by the citizens of EU member countries?	European Parliament	0	0	D	0
8	How many votes does China have in the UNO general assembly?	1	D	D	D	0
9	Where is Somalia located?	Indicate location on a map	D	D	D	D

Table A.2: Items from the SPISA dataset on the topic of history and whether they were flagged as DIF items (D) by one of the aggregation rules or chosen as an anchor item (A).

Item	Question	Answer	Aggregation			
			direct	all	min	mean
10	The Roman naval supremacy was established through...	the abolition of Carthage	0	A	A	A
11	In which century did the Thirty Years' War take place?	The 17th century	0	0	A	0
12	Which form of government is associated with the French King Louis XIV?	Absolutism	A	A	A	A
13	What island did Napoleon die on in exile?	St. Helena	0	0	0	0
14	How many percent of the votes did the NSDAP receive in the 1928 elections of the German Reichstag?	About 3 percent	0	0	0	D
15	How many Jews were killed by the Nazis during the Holocaust?	About 6 Million	0	0	0	0
16	Who is this?	Johannes Rau	0	0	0	0
17	Which of the following countries is not a member of the EU?	Croatia[1]	0	0	D	0
18	How did Mao Zedong expand his power in China?	The Long March	0	0	0	0

[1] At the time the quiz was conducted, Croatia was not a member of the EU.

Table A.3: Items from the SPISA dataset on the topic of economics and whether they were flagged as DIF items (D) by one of the aggregation rules or chosen as an anchor item (A).

Item	Question	Answer	Aggregation			
			direct	all	min	mean
19	Who is this?	Picture of Dieter Zetsche, CEO of Mercedes-Benz	D	D	D	D
20	What is the current full Hartz IV standard rate (part of the social welfare) for adults?	351 Euro	0	0	0	0
21	What was the average per capita gross national product in Germany in 2007?	About 29,400 Euro	0	0	0	0
22	What is a CEO?	A Chief Executive Officer	D	D	D	D
23	What is the meaning of the hexagonal 'organic' logo?	Synthetic pesticides are prohibited	D	D	D	D
24	Which company does this logo represent?	Deutsche Bank	D	D	D	D
25	Which German company took over the British automobile manufacturers Rolls-Royce?	BMW	D	D	D	D
26	Which internet company took over the media group Time Warner?	AOL	D	D	D	D
27	What is the historic meaning of manufacturies?	Manufacturies were the precursors of industrial mass production	0	0	0	0

Table A.4: Items from the SPISA dataset on the topic of culture and whether they were flagged as DIF items (1) by one of the aggregation rules or chosen as an anchor item (A).

Item	Question	Answer	Aggregation			
			direct	all	min	mean
28	Which painter created this painting?	Andy Warhol	D	D	D	D
29	What do these four buildings have in common?	All four were designed by the same architects	0	0	0	0
30	Roman numbers: What is the meaning of CLVI?	156	D	D	D	D
31	What was the German movie with the most viewers since 1990?	Der Schuh des Manitu	0	0	0	D
32	In which TV series was the US president portrayed by an African American actor for a long time?	24	0	0	0	0
33	What is the name of the bestselling novel by Daniel Kehlmann?	Die Vermessung der Welt (Measuring The World)	0	0	0	0
34	Which city is the setting for the novel 'Buddenbrooks'?	Luebeck	D	D	D	D
35	In which city is this building located?	Paris	A	A	A	A
36	Which one of the following operas is not by Mozart?	Aida	A	0	0	0

Table A.5: Items from the SPISA dataset on the topic of natural sciences and whether they were flagged as DIF items (D) by one of the aggregation rules or chosen as an anchor item (A).

Item	Question	Answer	Aggregation			
			direct	all	min	mean
37	Why does an ice floe not sink in the water?	Due to the lower density of ice	0	0	0	0
38	What is ultrasound not used for?	Radio	A	A	0	0
39	Which sensory cells in the human eye make color vision possible?	Cones	D	D	D	D
40	What is also termed Trisomy 21?	Down syndrome	0	0	0	0
41	Which element is the most common in the Earth's atmosphere?	Nitrogen	0	0	0	0
42	Which kind of tree does this leaf belong to?	Maple	D	D	D	D
43	Which kind of bird is this?	Blackbird	0	0	0	D
44	Where is the stomach located?	Indicate location on a map of the body	D	D	D	D
45	What is the sum of interior angles in a triangle?	180 degrees	D	D	D	D

Appendix B

Supplementary material: Chapter 3

B.1 Item parameters

Table B.1: Item parameters from Wang, Shih, and Sun (2012)

Item	Parameter	Item	Parameter	Item	Parameter	Item	Parameter
1	-2.522	11	0.295	21	-2.198	31	0.116
2	-1.902	12	0.778	22	-1.621	32	0.273
3	-1.351	13	1.514	23	-0.761	33	0.840
4	-1.092	14	1.744	24	-1.179	34	0.745
5	-0.234	15	1.951	25	-0.610	35	1.485
6	-0.317	16	-1.152	26	-0.291	36	-1.208
7	0.037	17	-0.526	27	0.067	37	0.189
8	0.268	18	1.104	28	0.706	38	0.345
9	-0.571	19	0.961	29	-2.713	39	0.962
10	0.317	20	1.314	30	0.213	40	1.592

B.2 DIF structures

Table B.2: DIF structure "norm"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.30	0.30

Table B.3: DIF structure "var"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	-0.30	0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30
0.00	0.30	-0.30	0.00	0.30	-0.30

Table B.4: DIF structure "inc"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50
0.00	0.10	0.20	0.30	0.40	0.50

Table B.5: DIF structure "bal"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	-0.30	-0.30
0.00	0.30	0.30	0.30	0.30	-0.30
0.00	0.30	-0.30	-0.30	-0.30	0.30
0.00	0.30	-0.30	-0.30	0.30	0.30
0.00	0.30	-0.30	-0.30	-0.30	-0.30
0.00	0.30	0.30	-0.30	0.30	0.30
0.00	0.30	0.30	-0.30	-0.30	-0.30
0.00	0.30	0.30	-0.30	0.30	-0.30
0.00	-0.30	-0.30	0.30	-0.30	0.30
0.00	-0.30	-0.30	0.30	0.30	0.30
0.00	-0.30	-0.30	0.30	-0.30	-0.30
0.00	-0.30	0.30	0.30	0.30	0.30
0.00	-0.30	0.30	0.30	-0.30	-0.30
0.00	-0.30	0.30	0.30	0.30	-0.30
0.00	-0.30	-0.30	-0.30	-0.30	0.30
0.00	-0.30	-0.30	-0.30	0.30	0.30
0.00	-0.30	-0.30	-0.30	-0.30	-0.30

Table B.6: DIF structure "chaos"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)
0	0.3	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)	unif(-0.6,0.6)

Table B.7: DIF structure "groups"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60

Table B.8: DIF structure "equal"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.30	0.30	0.30	0.60
0.00	0.30	0.30	0.30	0.30	0.60
0.00	0.30	0.30	0.30	0.30	0.60
0.00	0.30	0.30	0.30	0.30	0.60
0.00	0.30	0.30	0.30	0.60	0.60
0.00	0.30	0.30	0.30	0.60	0.60
0.00	0.30	0.30	0.30	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.60
0.00	0.30	0.30	0.60	0.60	0.00
0.00	0.30	0.30	0.60	0.60	0.00
0.00	0.30	0.30	0.60	0.00	0.00
0.00	0.30	0.60	0.60	0.00	0.00
0.00	0.30	0.60	0.00	0.00	0.00
0.00	0.30	0.60	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00

Table B.9: DIF structure "incdec"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.70
0.00	0.30	0.40	0.50	0.60	0.00
0.00	0.30	0.40	0.50	0.60	0.00
0.00	0.30	0.40	0.50	0.60	0.00
0.00	0.30	0.40	0.50	0.00	0.00
0.00	0.30	0.40	0.50	0.00	0.00
0.00	0.30	0.40	0.50	0.00	0.00
0.00	0.30	0.40	0.00	0.00	0.00
0.00	0.30	0.40	0.00	0.00	0.00
0.00	0.30	0.40	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00
0.00	0.30	0.00	0.00	0.00	0.00

Table B.10: DIF structure "alter"

Reference Group	Focal Group 1	Focal Group 2	Focal Group 3	Focal Group 4	Focal Group 5
0.00	0.30	0.30	0.30	0.30	0.30
0.00	0.30	0.30	0.30	0.00	0.00
0.00	0.30	0.30	0.30	0.30	0.00
0.00	0.30	0.00	0.00	0.00	0.30
0.00	0.30	0.00	0.00	0.30	0.30
0.00	0.30	0.00	0.00	0.00	0.00
0.00	0.30	0.30	0.00	0.30	0.30
0.00	0.30	0.30	0.00	0.00	0.00
0.00	0.30	0.30	0.00	0.30	0.00
0.00	0.00	0.00	0.30	0.00	0.30
0.00	0.00	0.00	0.30	0.30	0.30
0.00	0.00	0.30	0.30	0.00	0.00
0.00	0.00	0.30	0.30	0.30	0.00
0.00	0.00	0.00	0.00	0.00	0.30
0.00	0.00	0.00	0.00	0.30	0.30
0.00	0.00	0.00	0.00	0.00	0.00

B.3 Hit Rates in Further Simulations

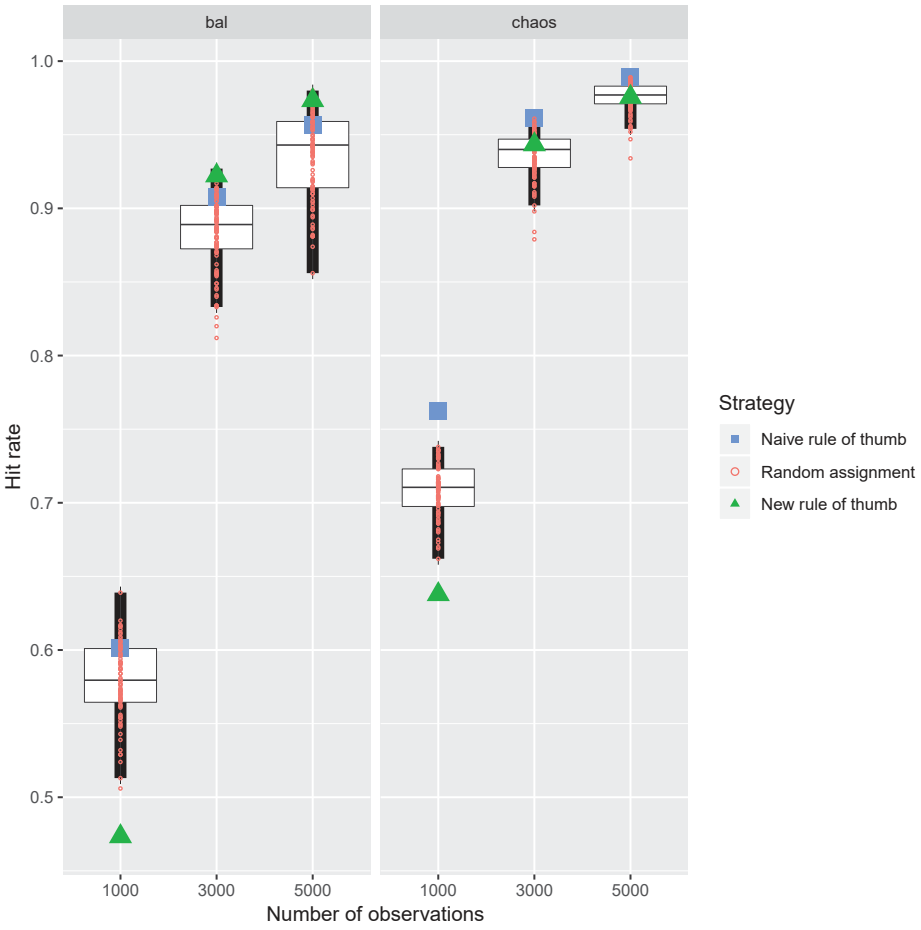

Figure B.1: Hit rates in the six group scenario under the "chaos" and "bal" DIF structure

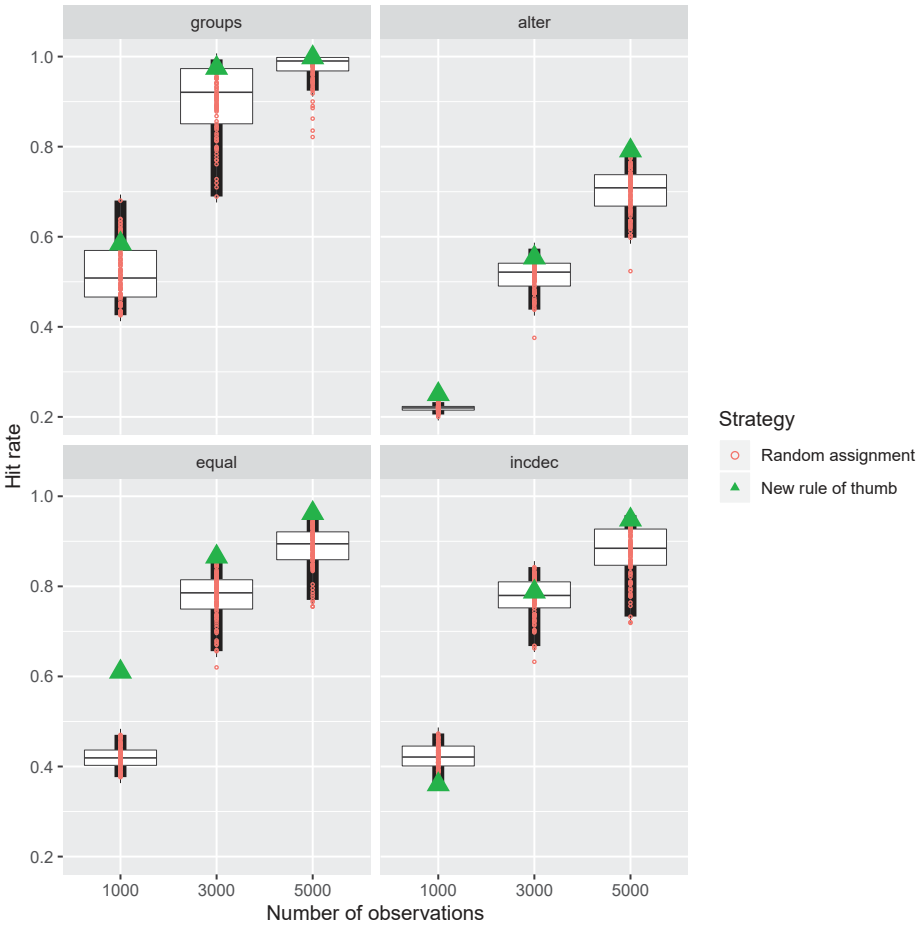

Figure B.2: Hit rates in the six group scenario under the "equal", "incdec", "alter" and "groups" DIF structure

B.4 R Code

```
optgroup <- function(dif.str, N)
{
dimensions <- dim(dif.str)
ngroups <- dimensions[2]
nitems <- dimensions[1]

compmat <- matrix(-99, ncol = ngroups, nrow = ngroups)

for (i in 1:ngroups){
for (j in 1:ngroups){
compmat[i,j] <- sum(dif.str[,i] == dif.str[,j])
}
}

checkmat <- diag(1,ngroups)
gone <- vector()

for(i in 1:(ngroups-1)){
for(j in (i+1):ngroups){
check <- isTRUE(all.equal(compmat[i,], compmat[j,]))
if (check)
{checkmat[i,j] <- checkmat[j,i] <- 1
gone <- rbind(gone,j)}
}
}
unigone <- unique(gone)
l <- 1:ngroups
notgone <- l[-unigone]
compmatnew <- compmat
if(length(gone != 0)) {
compmatnew <- compmat[-unigone, -unigone]
checkmat <- checkmat[notgone,unigone]}

compinv <- solve(compmatnew)
oben <- rowSums(compinv)
unten <- sum(oben)
erg <- round(N*oben/unten, 2)
if(any(erg < 0))stop("At least one group was calculated to be negative")
if (any(erg <= 20) && all(erg >0)) warning("For optimal ratio a group size
```

```
of less then 20 was calculated for at least one group. Check whether you
should increase the total N")
if(any(round(erg) == 0))warning("For optimal ratio a group size of 0 was
calculated for at least one group. Check whether you would like to add a
minimum number to every group")

if(length(erg) == ngroups)return(list(erg, checkmat, dif.str, compmat))
if(is.vector(checkmat)) ndiv <- checkmat
else ndiv <- rowSums(checkmat)

preresult <- erg/(ndiv+1)
result <- rep(-99, ngroups)
result[notgone] <- preresult
missing <- t(preresult)%*%checkmat
result[unigone] <- missing

return(list(result, checkmat, dif.str, compmat))
}
```

Appendix C

Supplementary material: Chapter 4

Table C.1: Item parameters from Wang, Shih, and Sun (2012)

Item	Parameter	Item	Parameter
1	-2.522	11	0.295
2	-1.902	12	0.778
3	-1.351	13	1.514
4	-1.092	14	1.744
5	-0.234	15	1.951
6	-0.317	16	-1.152
7	0.037	17	-0.526
8	0.268	18	1.104
9	-0.571	19	0.961
10	0.317	20	1.314