# Model-based Recursive Partitioning Meets Item Response Theory

## New Statistical Methods for the Detection of Differential Item Functioning and Appropriate Anchor Selection

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Julia Kopf
am 14.08.2013
in München

| | |
|---|---|
| Datum der Einreichung: | 14.08.2013 |
| Erstgutachterin: | Prof. Dr. Carolin Strobl |
| Zweitgutachter: | Prof. Dr. Thomas Augustin |
| Drittgutachter: | Prof. Dr. Achim Zeileis |
| Datum der Disputation: | 28.10.2013 |

*für Agnes*

# Summary

The aim of this thesis is to develop new statistical methods for the evaluation of assumptions that are crucial for reliably assessing group-differences in complex studies in the field of psychological and educational testing.

The framework of *item response theory* (IRT) includes a variety of psychometric models for scaling latent traits such as the widely-used Rasch model. The Rasch model ensures objective measures and fair comparisons between groups of subjects. However, this important property holds only if the underlying assumptions are met. One essential assumption is the invariance property that states constant item parameters for all subgroups. Its violation is extensively discussed in the literature and termed *differential item functioning* (DIF). This thesis focuses on the methodology of DIF detection. Existing methods for DIF detection are briefly discussed and new statistical methods are introduced.

One key aspect are the recently suggested Rasch trees that employ the model-based recursive partitioning algorithm to detect groups that display different item parameters in the Rasch model. In this thesis, it is shown that Rasch trees are also suited to detect non-uniform DIF, if the sample size is sufficiently large. For item-wise DIF detection methods, that rely on the comparison of item parameters, an anchor is necessary to place the estimated item parameters onto a common scale. A conceptual framework for the anchor process is suggested. Existing item-wise DIF tests are improved by new anchor methods and first ideas to generalize the methods for multiple-group comparisons are presented. The methods introduced in this thesis allow to classify items with and without DIF more accurately and, thus, to improve the evaluation of the invariance assumption in the Rasch model. This thesis, thereby, provides a contribution to the construction of objective and fair tests in psychology and educational testing.

# Zusammenfassung

Ziel dieser Arbeit ist die Entwicklung neuer statistischer Methoden zur Überprüfung von Annahmen, die maßgeblich für die Analyse von Gruppenunterschieden in komplexen Studien aus dem Bereich der Psychologie und der empirischen Bildungsforschung sind.

Die *Item Response Theory* (IRT) umfasst eine Vielzahl psychometrischer Modelle zur Skalierung latenter Personeneigenschaften, wie das weit verbreitete Rasch Modell. Das Rasch Modell gewährleistet objektive Messungen und erlaubt dadurch auch faire Vergleiche zwischen Personengruppen. Allerdings gilt diese wichtige Eigenschaft des Rasch Modells nur, solange die zugrundeliegenden Modellannahmen erfüllt sind. Eine zentrale Annahme des Modells ist die der Invarianz, die konstante Aufgabenparameter über Personengruppen hinweg fordert. Ihre Verletzung ist Gegenstand zahlreicher wissenschaftlicher Abhandlungen und wird als *Differential Item Functioning* (DIF) bezeichnet. Die Methodologie zur Prüfung von DIF steht im Vordergrund dieser Arbeit. Bereits vorgeschlagene Verfahren zur DIF Prüfung werden kurz diskutiert und neue statistische Verfahren zur Analyse von DIF werden entwickelt.

Einen Schwerpunkt bilden die jüngst vorgeschlagenen Rasch-Bäume, die das Modell-basierte rekursive Partitionieren einsetzen, um Gruppen mit unterschiedlichen Aufgabenparametern im Rasch Modell aufzudecken. In dieser Arbeit wird gezeigt, dass sich Rasch-Bäume auch eignen, um non-uniform DIF aufzudecken, wenn der Stichprobenumfang ausreichend groß ist. Item-weise DIF Methoden, die auf dem Vergleich geschätzter Aufgabenparameter basieren, verlangen eine Verankerung, um eine gemeinsame Skala der Aufgabenparameter zu konstruieren. Ein konzeptioneller Rahmen für Verankerungsverfahren wird vorgeschlagen. Item-weise DIF Tests werden durch neue Verankerungsmethoden weiter entwickelt und erste Ideen zur Verallgemeinerung dieser Verfahren für den Mehrgruppenvergleich werden vorgeschlagen. Die in dieser Arbeit neu vorgeschlagenen Verfahren erleichtern die Klassifikation von Aufgaben in solche mit und ohne DIF und ermöglichen so eine Verbesserung der Überprüfung der Invarianzannahme im Rasch Modell. Diese Arbeit leistet damit einen Beitrag zur Konstruktion objektiver und fairer Tests in der Psychologie und der empirischen Bildungsforschung.

# Danksagung

Zunächst möchte ich mich bei meinen Betreuern, Carolin Strobl und Thomas Augustin, für ihre fachliche Unterstützung, für ihre Motivation, für die vielen Anregungen und Perspektiven und alles darüber hinaus bedanken. Carolin Strobl danke ich besonders für alles was ich in der Zeit meiner Promotion von ihr lernen durfte, für ihren kritischen Blick und für ihre nicht zu erschöpfende Geduld. Thomas Augustin danke ich insbesondere dafür, dass er mir während meines Studiums der Soziologie das Fach Statistik näher gebracht hat und somit meinen Lebensweg in sehr entscheidender Weise beeinflusst hat.

Weiter möchte ich mich bei Achim Zeileis, der als Entwickler des Modell-basierten rekursiven Partitionierens den Grundstein dieser Arbeit gelegt hat, für die angenehme, hilfreiche Zusammenarbeit sowie für wichtige Vorschläge und Ideen bei gemeinsamen Manuskripten bedanken.

Die vorliegende Arbeit entstand im Rahmen des Projekts *Heterogenität in IRT-Modellen*, das im Förderprogramm *Empirische Bildungsforschung* durch das Bundesministerium für Bildung und Forschung finanziert wurde (Förderkennzeichen 01JG1060). Im Promotionsbegleitprogramm am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) konnten aktuelle Forschungsergebnisse mit ausgewiesenen Experten diskutiert werden, welchen ich für wertvolle Hinweise danken möchte.

Philipp Linzmeier möchte ich für Vertrauen, Liebe und Kraft, die diese Arbeit ermöglicht haben, danken – weiter auch für seinen Humor und seine Nachsicht.

Meinen Eltern, Renate und Michael Kopf, möchte ich danken, dass sie mir alle Bildungsentscheidungen frei überlassen haben und diese bedingungslos und nach Kräften unterstützt haben. Ich danke für die Aufgeschlossenheit, den Rückhalt, den Mut und die Zuversicht. Ich danke meiner Schwester Susanne Kopf für die schöne Zeit zusammen, für ihr Verständnis, für ihr offenes Ohr und für die Rückendeckung. Auch ihrem Freund Christoph Jansen wie meiner Familie danke ich fürs Mitfiebern und meiner Großmutter Agnes Kopf für das Wissen wie stolz sie wäre.

Für schönen Ausgleich zur wissenschaftlichen Arbeit sorgten meine Freunde, insbesondere Christine Feiner, Katharina Scheint, Philipp Albert, Christian Gärtner und Nils Moschüring. Ich danke ihnen dafür und auch für Zuspruch und Ermutigung.

Bei meinen Münchner und Züricher Kollegen möchte ich mich ebenfalls für die gute Zeit bedanken, besonders bei Stephanie Thiemichen, Gero Walter, Andrea Wiencierz, Nora Fenske, Jona Cederbaum, Johanna Brandt, bei meinen Bürokollegen Armin Monecke und Basil Abou El-Komboz und bei meiner Innsbrucker Kollegin Hannah Frick.

Auch bei den Soziologen Christine Feiner, Matthias Pflaum, Franziska Völkel, Anja Strobl, Janine Schmidt, Alexandra Klee, Robert Ivanov und Florian Haas möchte ich mich für die lustigen Stunden bedanken.

München, den 14. August 2013 Julia Kopf

# Contents

# 1 Scope of this work

> „*The importance of the property of invariance of item and ability parameters cannot be overstated. This property is the cornerstone of item response theory*" according to Hambleton *et al.* 1991, p. 25.

The detection of differential item functioning (DIF) – i.e. the violation of the invariant item parameter assumption – is of utmost importance for the valid measurement of latent traits such as abilities in the field of educational testing, personality traits in behavioral research or attitudes in political and social sciences. If DIF is present, fair comparisons between groups of test-takers for example between participating countries in the PISA study and meaningful individual test results for example for test-takers in academic assessments are no longer guaranteed.

Recently, a new statistical method for the detection of uniform DIF in the Rasch model termed *Rasch trees* was suggested (Strobl *et al.*, 2013). This approach brings together the model-based recursive partitioning algorithm (Zeileis *et al.*, 2008) and item response theory (IRT). This work addresses several challenges in the analysis of DIF, some of which are directly related to the newly proposed method and some that are concerned with the analysis of DIF in general.

This work is supported by the German Federal Ministry of Education and Research (BMBF) within the project "*Heterogeneity in IRT-Models*" (grant ID 01JG1060).

In Chapter 2, the model-based recursive partitioning algorithm (Zeileis *et al.*, 2008) is introduced with regard to its potential for the social sciences. Methods developed from machine learning, such as classification and regression trees or random forests, are exploratory data mining techniques. They search for patterns over a set of available covariates and are popular in many fields such as epidemiology, genetics or economics due to their high performance (Strobl, 2008). However, research in the social sciences is based on the development and evaluation of theories to explain the object of research, such as human behavior, and might often not benefit from the exploratory nature of the algorithmic methods. Yet, the variant of model-based recursive partitioning can bring together a statistical modeling part and a variable selection part that indicates whether the parameters of the underlying model are unstable in groups defined by observable covariates. Thus, model-based recursive partitioning allows to evaluate whether statistical models, that represent the substantial theories, are appropriate for the entire sample. It can be used as a broad test of assumptions of pre-defined statistical models, for example to assess violations of Ockham's razor.

When model-based recursive partitioning meets item response theory, it allows to investigate the assumption of parameter invariance that is crucial for psychometric models in the field of psychological and educational testing. The recently suggested Rasch trees (Strobl *et al.*, 2013) are based on the model-based recursive partitioning algorithm for the widely-used Rasch model.

It has already been shown to be a powerful tool for the detection of uniform DIF since it allows to detect observable groups that display different item parameters. The performance in word problems in a math test, for example, may depend on an additional trait dimension such as language skills and, thus, be harder for non-native speakers even if they have the same math ability. Rasch trees were developed as a global model test to answer the question if uniform DIF is present and – if so – which groups display DIF in the test. Chapter 3 briefly introduces the Rasch model, the problem of DIF, its relation to test fairness and several methods that were previously suggested for DIF detection with an emphasis on the Rasch trees.

In Chapter 4, it is evaluated whether Rasch trees are also suited for the detection of non-uniform DIF. Non-uniform DIF – as opposed to uniform DIF – is present when the disadvantages (measured in terms of item parameter differences) vary over the ability range. In the math test example, it may happen that non-native speakers with a lower ability are more strongly affected by DIF compared to non-native speakers with a higher ability level. In this case, non-uniform DIF is present. For a comparison with the global Rasch tree procedure, two extensions of the logistic regression (that was suggested for the item-wise detection of uniform and non-uniform DIF by Swaminathan and Rogers, 1990) that allow to test for DIF in all items are suggested. It is shown that Rasch trees are also suitable for the detection of non-uniform DIF when large sample sizes are present and that they outperform the extensions of the logistic regression when DIF is simulated in the difficulty parameters or when the logistic regression is misspecified. Furthermore, Rasch trees automatically search for the affected groups over all available covariates and maintain their straightforward interpretation.

Up to Chapter 4 this work focused on global tests for DIF. While the global tests can be considered as a first general assessment of measurement invariance, conclusions on the item level cannot yet be drawn. However, for practical research in test or questionnaire development, it is important to know which items display DIF and between which groups so that items can be modified or excluded from the test (Westers and Kelderman, 1992) and research on the underlying causes of DIF (Jodoin and Gierl, 2001) can be carried out. Since DIF generally reflects situations where the latent ability is no longer unidimensional (Mellenbergh, 1982), the knowledge about which items and groups display DIF may help to generate hypothesis about the additional dimensions involved.

To answer the question which items display DIF, item-wise DIF tests are regarded. Throughout this thesis, the classical Wald test based on the item parameters of the Rasch model (which are estimated using the conditional maximum likelihood) is used to assess DIF. For the DIF analysis, the origin of the scales of the item parameters has to be determined. Therefore, one linear restriction is imposed in each group and the items in the restriction are termed *anchor items* or the *anchor*. There are two problems associated with the choice of the anchor that is determined by an anchor method. First, the results of the DIF-analysis strongly depend on the choice of the anchor. Second, little information is at hand how the anchor is selected appropriately (Lopez Rivas *et al.*, 2009). The Wald test is shown to be an appropriate test for DIF if it is combined with a suitable anchor method.

Chapter 5 provides an introduction of the anchor problem together with a new conceptual framework for the classification of previously suggested anchor methods. In this framework, it is distinguished between *anchor classes* that determine characteristics of the anchor methods, such as a predefined number of anchor items, and *anchor selection strategies* that locate the anchor items. The entire procedure, consisting of an anchor class and – if necessary – an anchor selection strategy, is then termed *anchor method*.

In this thesis, several new suggestions for the anchor process are proposed, since an appropriate anchor method is necessary for the correct classification of DIF- and DIF-free items. In case the anchor method does not work appropriately, inflated false alarm rates (i.e. DIF-free items erroneously display DIF) occur, that jeopardize the statistical inference at a predefined significance level and, thereby, the correct classification of DIF- and DIF-free items. The inflated false alarm rate is induced by an artificial difference between the item parameters of the groups that also affects descriptive measures or effect size measures. In Chapter 6, a new anchor class named the *iterative forward anchor class* is developed for the DIF analysis of two pre-specified groups and evaluated in an extensive simulation study. While previously suggested methods often start with a criterion that is severely biased, this problem is avoided by starting with a single anchor item and successively including items in the anchor. The iterative forward anchor class generates a longer anchor and, thereby, allows for a higher hit rate (i.e. more DIF-items are correctly detected).

In Chapter 7, existing anchor selection strategies are compared with three new suggestions: the *mean p-value*, the *mean test statistic threshold* and the *mean p-value threshold* selection. The anchor selections are combined with the newly proposed iterative forward anchor class and the previously suggested constant anchor class. In an extensive simulation study, it is shown that the anchor selection strategies developed in this thesis outperform previously suggested anchor selection strategies in the majority of the simulated settings and allow for a lower false alarm rate (i.e. less DIF-free items display DIF) and a higher hit rate (i.e. more DIF-items are detected).

So far, the development of the anchor class and anchor selection strategies was limited to the comparison of two groups. In Chapter 8, two alternative extensions of the anchor methods for multiple group comparisons are suggested. Even though several authors proposed to conduct post-hoc tests that answer the question which items display DIF between which groups (Kim *et al.*, 1995; Penfield, 2001), the literature on the anchor problem is – to my knowledge – limited to the two-group case. The extension to multiple group comparisons is also important for the previously suggested Rasch trees that automatically detect – potentially more than two – groups of subjects that display DIF. To draw conclusions which items display DIF and between which groups, multiple comparison procedures need to be constructed (independently of whether they are used as post-hoc significance tests or as descriptive measures). Here, it is recommended to generalize the anchor methods for these comparisons by selecting a common set of anchor items. Therefore, two aggregation rules, the *mean* and the *minimax* rule are suggested.

Throughout this thesis, several decisions about assumptions or methods to be investigated have been taken. Chapter 9 is concerned with alternative ideas. First, it is evaluated whether previously suggested quasi-variances (Firth, 2003) can be used for DIF detection. Second, an alternative to the decision that the first anchor item of a test was defined DIF-free is discussed and evaluated.

In Chapter 10, the contents of this thesis are briefly summarized. Furthermore, the main scientific findings are discussed followed by the limitations of this work. Finally, future research questions are addressed.

In summary, the suggestions presented in this thesis by means of new methods for DIF detection (Chapter 4), a new anchor class (Chapter 6) and new anchor selection strategies (Chapter 7) allow for lower false alarm rates (i.e. less items erroneously display significant DIF) and higher hit rates (i.e. true DIF-items are more often detected) compared to previous suggestions. In Chapter 8 extensions for multiple group comparisons were addressed to form the basis for post-hoc measures for the Rasch trees to answer the question which items display DIF between which groups.

By employing the new methods, expenses for test or questionnaire construction can be reduced since fewer items are falsely eliminated and less new items have to be developed. Furthermore, items that are affected by DIF are more often correctly detected. The detection of DIF-items is necessary for fair comparisons between groups. If items of a test are affected by DIF, groups of equally able test-takers display different probabilities of solving the items that may result in different test results. Thus, groups of test-takers – such as male test-takers, female test-takers or native speakers – may have an unfair advantage in the test. On the one hand, groups can no longer be compared in a fair way and, on the other hand, individual test results are no longer valid. This is an important argument, since in contemporary society important personal and political decisions are based on results from IRT models (examples can be found in Chapter 3). Presumably, standardized test results will become even more widely used in educational testing and other settings.

### Contributing Manuscripts

Parts of this thesis are already published as a technical report, as a book chapter or a journal article. The rest are based on yet 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 content and the contributions from all authors.

- Kopf J., Augustin T. and Strobl C. (2013): *The Potential of Model-based Recursive Partitioning in the Social Sciences – Revisiting Ockham's Razor*. In: McArdle J.J. and Ritschard G. (Ed.): *Contemporary Issues in Exploratory Data Mining*, Routeledge, New York, to appear 2013.

  In this manuscript, we review the model-based recursive partitioning technique and the predecessor method classification and regression trees. To highlight the potential of model-based recursive partitioning for the social sciences, we point out the relation of the algorithmic method to the principle of parsimony and Ockham's Razor.
  Julia Kopf identified the relation to Ockham's Razor as one key argument for the use of model-based recursive partitioning in the social sciences, performed the analyses and drafted the manuscript. Carolin Strobl and Thomas Augustin contributed to the conception and presentation of the article.

  Chapter 2 is based on contents of this manuscript.

- Strobl C., Kopf J. and Zeileis A. (2013): *Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model*. Psychometrika, accepted.

  This manuscript proposes a new method for DIF detection termed *Rasch trees* based on model-based recursive partitioning of the Rasch model. With this approach, it is possible to detect groups of subjects exhibiting DIF, which are not pre-specified, but result from combinations of observed covariates.
  Carolin Strobl developed the idea to use the model-based recursive partitioning algorithm for DIF detection in the Rasch model, conducted the final simulation studies and drafted the manuscript. Achim Zeileis created the statistical software, contributed to the methodological aspects and to the manuscript. Julia Kopf reviewed the literature, derived measures necessary for the statistical test and conducted preliminary simulation studies during her master thesis.

  Chapter 3, that forms the statistical background for the following methodological suggestions, is partly based on contents of this manuscript.

- Kopf J. and Strobl C. (2013): *Detecting Non-uniform DIF with Rasch Trees*.

  In this manuscript, we address the question of whether the newly proposed Rasch trees are also suited for the detection of *non-uniform DIF* in the Rasch model. Furthermore, we suggest two extensions of the logistic regression approach, that has previously been

proposed for the detection of uniform and non-uniform DIF, to allow for a direct comparison.
Julia Kopf suggested the extensions of the logistic regression for the comparison, planned and performed the simulation studies and drafted the manuscript. Carolin Strobl identified the problem, developed the idea to detect non-uniform DIF with the Rasch trees and contributed to the presentation.

Chapter 4 is based on contents of this manuscript.

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

  This manuscript proposes a conceptual framework for categorizing anchor methods: The *anchor class* to describe characteristics of the anchor methods and the *anchor selection strategy* to guide how the anchor items are determined. Furthermore, we propose a new anchor class termed the *iterative forward anchor class* and compare it to several previously suggested anchor classes in an extensive simulation study.
  Julia Kopf developed the conceptual framework, suggested the iterative forward anchor method, provided software for the anchor methods, conducted the simulation study and drafted the manuscript. Carolin Strobl contributed to the manuscript, especially regarding the interpretability of the results, and made methodological suggestions for consideration. Achim Zeileis provided one part of the software for the DIF tests and contributed to the formal notation in the manuscript.

  Chapters 5 and 6 and are based on contents of this manuscript.

- Kopf J., Zeileis A. and Strobl C. (2013): *Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches*. Technical Report 150, Department of Statistics, LMU Munich.

  In this manuscript, we review existing anchor selection strategies that do not require any knowledge prior to DIF analysis – as this is typically not available in practical applications – offer a formal notation for these strategies and propose three new anchor selection strategies. We evaluate the appropriateness of the anchor selection strategies by conducting an extensive simulation study.
  Julia Kopf developed the new threshold anchor selection strategies, implemented the anchor selection strategies in statistical software, conducted the simulation study and drafted the manuscript. Carolin Strobl contributed to the presentation in the manuscript. Achim Zeileis provided parts of the software for the DIF tests, suggested the usage of p-values for anchor selection and contributed to the presentation of the manuscript.

  Chapter 7 is based on contents of this manuscript.

- Kopf J., and Strobl C. (2013): *Outlook on Anchor Selection Strategies for Multiple Group Comparisons in DIF Analysis*.

Ideas how anchor selection strategies can be generalized to paired multiple group comparisons are presented in this manuscript draft. We argue to select a common set of anchor items instead of a different anchor set for each paired comparison. Two aggregation rules are introduced to generalize the anchor selection strategies for multiple group comparisons.
Julia Kopf developed the generalization to multiple groups and drafted the manuscript. Carolin Strobl contributed to the manuscript.

Chapter 8 is based on contents of this manuscript draft and gives a first impression of current work in progress.

- Kopf J., and Strobl C. (2013): *On Quasi-variances for DIF Detection.*

  Some alternative approaches to the strategies used in this thesis are presented in this manuscript draft. First, quasi-variances are evaluated regarding their appropriateness in DIF analysis. Second, the question whether the test decision for the first anchor item based on quasi-variances is better suited compared to the decision to declare the first anchor item as DIF-free is addressed.
  Achim Zeileis developed the idea to use quasi-variances for DIF detection. Julia Kopf reviewed the methodological aspects, carried out the simulation study and drafted the manuscript. Carolin Strobl contributed to the manuscript.

  Chapter 9 is based on contents of this manuscript draft and also reflects ongoing work.

# 2 The potential of model-based recursive partitioning in the social sciences – Revisiting Ockham's Razor

***Summary:*** *A variety of new statistical methods from the field of machine learning have the potential to offer new impulses for research in the social, educational and behavioral sciences. In this chapter we focus on one of these methods: model-based recursive partitioning. This algorithmic approach is reviewed and illustrated by means of instructive examples and an application to the Mincer equation, that is commonly used to describe the association between education, job experience and income in econometric and sociological research. For readers unfamiliar with algorithmic methods, the explanation starts with the introduction of the predecessor method classification and regression trees. As opposed to classification and regression trees that search for groups of observations that differ in the values of a response variable, model-based recursive partitioning searches for groups differing in their estimated parameters of a postulated statistical model. With respect to the application and interpretation of model-based recursive partitioning, we highlight the principle of parsimony and Ockham's Razor. To facilitate the applicability in the social sciences, we close with a section on recursive partitioning software available in the free R system for statistical computing.*

***Keywords:*** *model-based recursive partitioning; structural change; parsimony; classification and regression trees (CART); algorithmic methods; data analysis*

## 2.1 Introduction

The aim of this chapter is to demonstrate the potential of model-based recursive partitioning (Zeileis *et al.* 2008; related approaches have previously been suggested by Loh 2002; Li *et al.* 2000; Chaudhuri *et al.* 1995, Wang and Witten 1997), a statistical method adopted from the field of machine learning, for applications in the social sciences. In particular, we will point out that this algorithmic method provides a powerful tool to evaluate whether relevant covariates have been omitted in a statistical model and, therefore, whether a theoretically postulated model is in conflict with Ockham's Razor.

As a prototypical example the method is employed for evaluating the appropriateness of the so called Mincer equation (Mincer, 1974), which explains different income levels through rates of return from schooling and work experience by means of a linear model. The analysis relies on data of the German Socio-Economic Panel Study (SOEP) from 2008, provided by DIW Berlin (German Institute for Economic Research).

Model-based recursive partitioning can be considered as a powerful synthesis between nonparametric partitioning methods and parametric regression models. In contrast to standard multiple regression approaches, model-based recursive partitioning is based on the successive segmentation of the sample used: the data are split further as long as different groups of observations still display substantially different values of the estimated parameters of the statistical model of interest.

Hence, the objective of model-based recursive partitioning is related to the objective of latent class or mixture models where different regression parameter estimates are permitted between subgroups of the data set (see e.g. Vermunt, 2010; Leisch, 2004, for general introductions to mixture or latent class models, and Ünlü, 2011, for a specific application to knowledge structures). In latent class regression models, these groups are unobserved, whereas in model-based recursive partitioning the groups are determined from combinations of observed covariates.

For example, in our investigation of the Mincer equation we will see that the intercept and the estimated coefficient for further education vary across groups of men and women working full-time in east or west Germany. Thus additional sociological and economic theories, such as discrimination in labor markets (e.g. Aigner and Cain, 1977; Phelps, 1972), need to be considered for explaining these differences.

The method of model-based recursive partitioning forms an advancement of classification and regression trees, which are widely used in life sciences (cf. e.g. Hannöver *et al.*, 2002; Kitsantas *et al.*, 2007; Romualdi *et al.*, 2003; Zhang *et al.*, 2001) and have recently been applied to social and behavioral sciences (e.g. Berk, 2006). Classification and regression trees will be summarized briefly in the following section, beginning with an informal description of the resulting tree-structure.

After some technical details of classification and regression trees are reviewed, the advanced method of model-based recursive partitioning is addressed in Section 2.3, firstly by pointing out the main differences and similarities to classification and regression trees. The review of model-based recursive partitioning will be continued by interpreting an instructive example and recapitulating the statistical background. To facilitate the use of this powerful algorithmic method in the social sciences, this chapter highlights the interpretation with regard to the principle of parsimony in the context of model construction (Section 2.4). Moreover, the application to the Mincer equation in Section 2.5 demonstrates the potential of model-based recursive partitioning in empirical research. For further research, software available in the R system for statistical computing is indicated in Section 2.6.

In summary, in this chapter we show how model-based recursive partitioning allows to decide whether a postulated model fails to describe the whole sample in a suitable way, because the method may detect varying parameter estimates in different subgroups of the sample. Model-based recursive partitioning therefore offers a synthesis of the theory-based and the data-driven approach. In particular, it can be used for detecting violations of Ockham's Razor. If subgroups with different parameter estimates are found, the postulated model is too simple and not appropriate for the entire sample.

## 2.2 Classification and regression trees

Classification and regression trees (cf. e.g. Breiman *et al.*, 1984) are based on a purely data-driven paradigm. Without referring to a concrete statistical model, they search recursively for groups of observations with similar values of the response variable by building a tree structure. If the response is categorical, one refers to classification trees; if the response is continuous, one refers to regression trees. The basic principles of this approach will be explained by means of an exemplary application in the following section.

### 2.2.1 Basic principles of classification and regression trees

As a first example, we consider the respondents of the SOEP study 2008 (see Wagner *et al.*, 2007, for details about SOEP). In this data set, groups of subjects vary with respect to whether they participate in full-time labor or not (the latter including all categories like part-time or marginally employment, civil or military service, vocational training and unemployment labeled here as ‘other’).

These groups can be described by means of covariates, such as age and gender. The covariates (here: age and gender), together with the response variable (full time or other), are handed over to the algorithm. The resulting tree-structure is displayed in Figure 1.

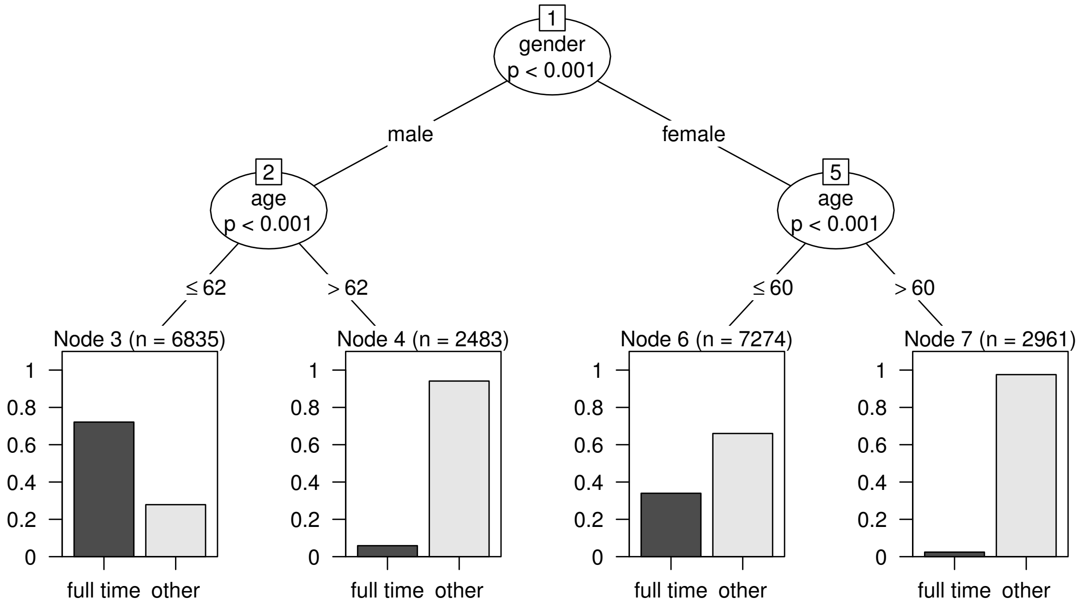

**Figure 1 –** *Classification tree: Assessing different frequencies of full-time jobs in Germany (SOEP 2008). The resulting tree-structure shows varying participation rates in full-time labor in three splits according to the covariates gender and age.*

From the entire sample of about 19,553 respondents living in private households, the covariate with the highest association (for technical details see Section 2.2.2) to the response is chosen for the first split. It is the participant’s gender, and, thus, 9,318 male respondents (represented in the left branch) are separated from the rest of the sample (10,235 female respondents, represented in the right branch). In the next step the male group is further diversified: it is split into two new subgroups, over the age of 62 or not (node 3 and node 4). Figure 1 also shows that the majority of women in the lower age group respond differently (node 6) compared to those in node 7. Here we stop the algorithm for simplicity.

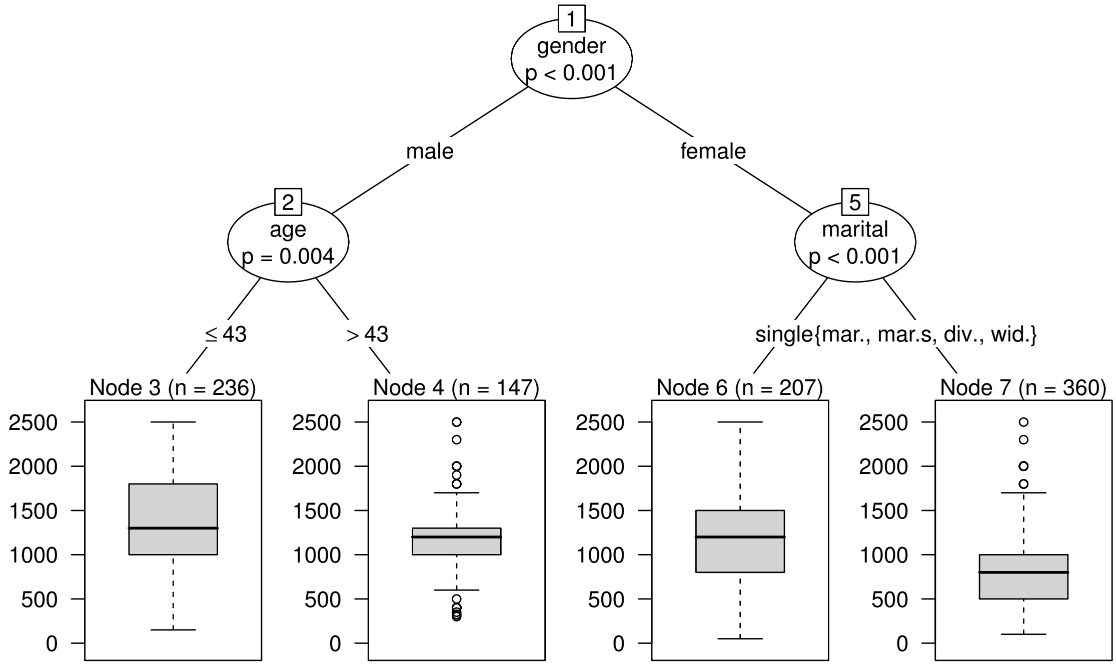

**Figure 2 –** *Regression tree: Assessing different requested incomes of unemployed respondents (SOEP 2008). Three different levels are obtained in groups related to gender, age and marital status.*

The respective cutpoint for these splits depends on the type of the covariate: while gender has only two categories – male and female – and thus offers only one cutpoint, referring to age the algorithm must also find the ‘best’ cutpoint within this variable. This optimal cutpoint turns out to be located at the threshold of 62 years for the male subsample and 60 years for the female subsample (technical details are given below).

The resulting tree-structure is interpreted easily and shows groupwise frequencies for full-time and non-full-time workers in the end nodes: the left node indicates that the majority of men up to 62 years in Germany work full-time, while the majority of women up to 60 do not. Women over 60 years are hardly ever employed full-time. The tree-structure in this example represents an interaction effect between gender and age (see e.g. Strobl *et al.*, 2009, for details on the interpretation of main effects and interactions in classification trees).

In contrast to this classification problem, regression trees focus on continuous response variables. Instead of regarding the frequencies of the categories, groups with different average response values are separated and visualized, e.g. by means of box plots (like in Figure 2). These groups are again detected automatically.

In this example the regression tree searches for different patterns of the (outlier adjusted) requested income at which 950 unemployed respondents would take a job (outliers are defined as participants with a requested income higher than the third quartile plus 1.5 times interquartile range). Additional covariates that are handed over to the algorithm are gender, age, nationality (nation) and marital status (marital). Figure 2 again shows the first split in the variable gender (node 1). The second split in the male subsample is again related to age (node 3 and node 4), while the third split in the female subsample is associated with marital status (node 6 and node 7). The cutpoint in a categorical variable is chosen automatically in an optimal way from all possible combinations of categories. Here the requested mean income is associated with marital status, in particular the request of female singles differs from the other categories (married, married but separated, divorced and widowed). The latter categories have smaller values of the requested income (median $x_{\text{med}} = 800$, mean $\overline{x} = 870$) than female singles ($x_{\text{med}} = 1200$, $\overline{x} = 1166$), who seem to be in the same magnitude as men over 43 years ($x_{\text{med}} = 1200$, $\overline{x} = 1159$). The highest average of requested income occurs within the male subsample up to the age of 43 years (node 3, $x_{\text{med}} = 1300$, $\overline{x} = 1349$). After these three splits, all of the determined groups are homogeneous enough to let the algorithm come to a stop, without further splitting, e.g. according to the nationality of the respondent. This exemplifies another attractive feature of partitioning methods: they implicitly perform a flexible variable selection.

A more detailed description of the technical procedure underlying classification and regression trees is given in the next section.

### 2.2.2 Some technical details

Classification trees search for different patterns in the response variable according to the available covariates. Since the sample is divided in rectangular partitions defined by values of the covariates and since the same covariate can be selected for multiple splits, classification trees can assess even complex interactions, non-linear and non-monotone patterns. The structure of the underlying data-generating process is not specified in advance, but is determined in an entirely data-driven way. These are the key distinctions between classification and regression trees, and classical regression models. The approaches differ, firstly, with respect to the functional form of the relationship that is limited to e.g. linear influence of the covariates in most parametric regression models and, secondly, with respect to the pre-specification of the model equation in parametric models.

Historically, the foundations for classification and regression trees were first developed in the sixties as Automatic Interaction Detection (Morgan and Sonquist, 1963). Later the most popular algorithms for classification and regression trees were developed by Quinlan (1993) and Breiman *et al.* (1984). Here we concentrate on a more recent framework by Hothorn *et al.* (2006b), which is based on the theory of conditional inference developed by Strasser and Weber (1999). The major advantage of this approach is that it avoids two fundamental problems of earlier algorithms for classification and regression trees: variable selection bias and overfitting (cf. e.g. Strobl *et al.*, 2009).

The algorithm of Hothorn *et al.* (2006b) for binary recursive partitioning can be described in three steps: firstly, beginning with the whole sample, the global null hypothesis that there is no relationship between any of the covariates and the response variable is evaluated. If no violation of the null hypothesis is detected, the procedure stops. If, however, a significant association is discovered, the variable with the largest association is chosen for the split. Secondly, the best cutpoint in this variable is determined and used to split the sample into two groups according to values of the selected covariate. Then the algorithm recursively repeats the first two steps in the subsamples until there is no further violation of the null hypothesis, or a minimum number of observations per node is reached.

In the following, we briefly summarize which covariates can be analyzed using classification and regression trees, how variables are selected for splitting and how the cutpoint is chosen.

*The response variable in the end nodes*

As outlined in the previous section, classification trees search for groups of similar response values with respect to a categorical dependent variable, whereas regression trees focus on continuous variables. Hothorn *et al.* (2006b) stress that their conditional inference framework can be applied beyond that to situations of ordinal, censored survival times and multivariate response variables.

Within the resulting tree-structure, all respondents with the same covariate values – represented graphically in one end node – obtain the same prediction for the response, i.e. the same class membership for categorical responses or the same value for continuous response variables.

*Selection of splitting variables*

The next question is how the variables for the potential splits are chosen and how the related cutpoints can be obtained. As outlined above, Hothorn *et al.* (2006b) provide a statistical framework for tests applicable to various data situations. In the binary recursive partitioning algorithm, each iteration is related to a current data set (beginning with the whole sample), where the variable with the highest association is selected by means of permutation tests as described in the following. The usage of permutation tests allows for evaluating the global null hypothesis $H_0$ that none of the covariates has an influence on the dependent variable. If $H_0$ holds (in other words, if the independence between any of the covariates $Z_l$ ($l = 1, \ldots, L$) and the dependent variable $Y$ cannot be rejected), the algorithm stops. Therefore the statistical test acts both for variable selection and as a stopping criterion.

Otherwise the strength of the association between the covariates and the response variable is measured in terms of the p-value that corresponds to the test of the null hypothesis that the specific covariate is not associated with the response. Thus, the variable with the smallest p-value is selected for the next split. The advantage of this approach is that the p-value criterion guarantees an unbiased variable selection regardless of the scales of measurement of the covariates (cf. e.g. Hothorn *et al.*, 2006b; Strobl *et al.*, 2007, 2009).

Permutation tests are constructed by evaluating the test statistic for the given data under $H_0$. Monte-Carlo or asymptotic approximations of the exact null-distribution are employed for the computation of the p-values (see Hothorn *et al.*, 2006a,b; Strasser and Weber, 1999, for more details).

*Selection of the cutpoints*

After the variable for the split has been selected, we need a cutpoint within the range of the variable to find the subgroups that show the strongest difference in the response variable. In the procedure described here, the selection of the cutpoint is also based on the permutation test statistic: the idea is to compute the two-sample test statistic for all potential splits within the covariate. In the case of continuous variables all potential cutpoints between any two successive observations are investigated (except for a certain percentage of the smallest and largest observations to avoid too small nodes). In the case of ordinal variables the ordering of the categories is accounted for. The resulting split is located where the binary separation of two data sets leads to the highest test statistic. This reflects the largest discrepancy in the response variable with respect to the two groups.

In the case of missing data, the algorithm proceeds as follows: observations that have missing values in the currently evaluated covariate are ignored in the split decision, whereas the same observations are included in all other steps of the algorithm. The class membership of these observations can be approximated by means of so called surrogate variables (Hothorn *et al.*, 2006b; Hastie *et al.*, 2008).

## 2.3 From classification and regression trees to model-based recursive partitioning

Model-based recursive partitioning was developed as an advancement of classification and regression trees. Both methods originate from the field of machine learning, which is influenced by both statistics and computer sciences.

The algorithmic rationale behind classification and regression trees is described by Berk (2006, p. 263) in the following way:

> ”*With algorithmic methods, there is no statistical model in the usual sense; no effort has been made to represent how the data were generated. And no apologies are offered for the absence of a model. There is a practical data analysis problem to solve that is attacked directly with procedures designed specifically for that purpose.*”

In that sense, classification and regression trees are purely data-driven and exploratory – and thus mark the entire opposite of the theory-based approach of model specification that is prevalent in the empirical social sciences.

The advanced model-based recursive partitioning method, however, brings together the advantages of both approaches: at first, a parametric model is formulated to represent a theory-driven research hypothesis. Then this parametric model is handed over to the model-based recursive partitioning algorithm that checks whether other relevant covariates have been omitted which would alter the parameters of the model of interest. Note that, as opposed to latent class regression the groups yielding differing parameter estimates are explained by covariates and not by a latent class approach.

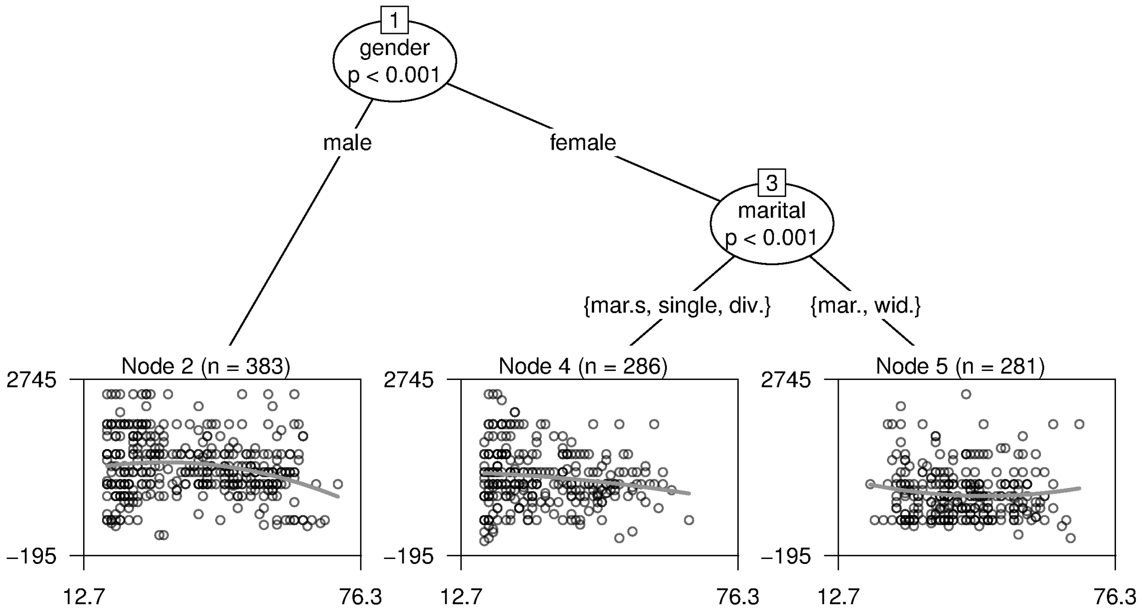

**Figure 3** – *MOB: Assessing different relationships between age and requested income of unemployed respondents in Germany (SOEP 2008). The line pictures the estimated relationship in the current subsample and indicates the varying parameters according to groups related to age and marital status.*

Technically, the tree-structure obtained from classification and regression trees remains the same for model-based recursive partitioning. However, instead of splitting for different patterns of the response variable, now we search for different patterns of the association between the response variable and other covariates, that has been pre-specified in the parametric model. Therefore the end nodes in the model-based tree represent statistical models, such as linear models, and no longer mere values of the response variable. The execution of a split in the model-based tree then indicates a parameter instability in the original model, i.e. the postulated model is too simple to explain the data.

### 2.3.1 Basic principles of model-based recursive partitioning

As an instructive example for a partitioned model, Figure 3 shows the tree-structure for a sample of unemployed respondents from the SOEP study. The model of interest here is the relationship between the requested income, at which respondents would take on a new job, and the age. The functional form of this relationship is fixed to a quadratic polynomial as often found intuitive for models relating age and income:

$$\text{requested income} = \tau_0 + \tau_1 \cdot \text{age} + \tau_2 \cdot \text{age}^2 + \varepsilon.$$

Additional covariates passed over to the algorithm are marital status, gender and nationality.

Beginning with the whole outlier adjusted sample of 950 unemployed respondents the model with the linear and quadratic term is fitted, where the estimated coefficients $\hat{\tau}_0, \hat{\tau}_1$, and $\hat{\tau}_2$ indicate parameter instability. The highest instability is related to gender and thus a split in this variable is performed. While in the sample of the male respondents (node 2) no more instabilities are detected, the female subset is again divided into two subgroups with differing parameter estimates. The end node in the middle shows the result for married but separated, single or divorced women (node 4). The rightmost end node contains the linear model for married and widowed women (node 5). Interestingly, even the direction of the relationship changes from a parabola on the left, where men of higher age tend to request less income, to a slight u-shape on the right, where married or widowed women request more income in higher age. The attractive feature of implicit variable selection is also maintained in model-based recursive partitioning: the nationality does not occur in any split decision in this example.

| | **estimated coefficients** | | |
|---|---|---|---|
| **node number** | $\hat{\tau}_0$ | $\hat{\tau}_1$ | $\hat{\tau}_2$ |
| 2 | 1014.5837 | 22.5446 | -0.3618 |
| 4 | 1212.6621 | -0.0236 | -0.0871 |
| 5 | 1390.9708 | -25.3983 | 0.2737 |

**Table 1** – *Estimated coefficients of the regression models in the end nodes estimated from the data sets that correspond to the end nodes.*

In Table 1 the parameter estimates for the different groups – represented in the end nodes of the tree-structure – are displayed. The varying signs of the coefficients confirm what is illustrated in Figure 3: the inverse u-shape holds only for part of the sample and is reversed for other parts. Thus, the example illustrates that model-based recursive partitioning is indeed able to detect different functional forms which might be masked when a single model is fit to the data.

The example also shows that, as opposed to classification and regression trees, the end nodes in model-based recursive partitioning do not contain values of a response variable, but represent a statistical model for each specific subpopulation. Between these groups the estimated parameters of the common underlying model vary significantly, but the postulated basic functional form (here polynomial) stated by the researcher is fixed. Within the subgroups no significant parameter instability is present.

Hence, the interpretation of a tree without any split is quite simple: there are no significant parameter instabilities found in any of the covariates handed over to the algorithm. If, however, a tree-structure is displayed, it reveals that the postulated model is not appropriate for describing the entire sample. The variation of the parameters highlights structural differences in the obtained subgroups, which can be easily interpreted by examining the estimates or the graphical output.

In the next section, important steps of the model-based partitioning algorithm are outlined. Then we take a closer look at the interpretation in social or behavioral sciences in Section 2.4

### 2.3.2 Some technical details

The model-based recursive partitioning algorithm maintains the fundamental steps of the partitioning method reviewed in Section 2.2, but coherently extends them in the light of the model-based paradigm. According to this paradigm, the recursive process now estimates the basic statistical model beginning with all available observations. The result of this step is the estimated parameter vector from the optimization of the objective function, typically the (log-)likelihood. In almost the same manner as classification trees, the recursive process starts: instead of testing the association, now the parameter instability is assessed using so called generalized M-fluctuation tests (Zeileis, 2005; Zeileis and Hornik, 2007). If the data indicate parameter instability, the split of the parent node in two daughter nodes is executed. Relying on the data points in the new subgroups only, the algorithm again searches for parameter instability until no further significant instability is found, or another stopping criterion is fulfilled. The resulting tree structure can be visualized as illustrated in the examples presented below, so that the different groups can be compared. Note, however that statistical tests conducted after model selection – such as significance tests for group differences after recursive partitioning – may be affected by the effects described by Leeb and Pötscher (2005) and Berk *et al.* (2010), and should thus be based on new data.

This brief overview about the similarities and differences in the algorithms leaves some questions that have yet to be explained: Which models can be partitioned recursively? How can we assess parameter instability and where are the optimal cutpoints in the covariates in model-based recursive partitioning? These questions are addressed in the next subsections, which are structured in the same way as Section 2.2.2.

*The statistical model in the end nodes*

The foundation of a general statistical framework for model-based recursive partitioning by Zeileis *et al.* (2008) allows using a variety of underlying statistical models, such as linear and logistic regression models. The wide range of applications emerges from the inclusion of several widely used test statistics in a unified approach (Zeileis, 2005) called generalized M-fluctuation tests.

Technically, the generalized M-fluctuation test used for the split decisions relies on the objective function $\Psi(.)$ of the parameter estimation, like least-squares and maximum-likelihood-estimation:

$$\hat{\vartheta} = \arg\max_{\vartheta} \sum_{i=1}^{n} \Psi(y_i, \vartheta), \tag{1}$$

where $y_i$ $(i = 1, \ldots, n)$ symbolizes the vector of all values of the dependent and independent variables in the postulated model for subject $i$, and $\vartheta$ represents the (potentially vector-valued) parameter. For reasons of simplicity, here we use the full sample notation and do not distinguish whether the underlying observations are the entire sample or a specific subgroup arising from the recursive application of the procedure.

The estimation process is based on the individual contributions of each subject $i$ to the score function

$$\psi(y_i, \vartheta) = \frac{\partial \Psi(y_i, \vartheta)}{\partial \vartheta}, \tag{2}$$

as outlined below.

In addition to the model specification, the algorithm requires categorical or numeric covariates – denoted as $Z_l$ ($l = 1, \ldots, L$) – for potential splits in the model-based tree.

*Selection of splitting variables*

After the first step of the algorithm – fitting the underlying model for the whole sample and obtaining a preliminary estimate $\hat{\vartheta}$ – a test of parameter instability is performed. It is based on the statistical framework developed by Zeileis and Hornik (2007) to detect structural changes by fluctuation tests. In econometrics, these tests for structural changes are widely used to detect e.g. a drop in the expected value of a time series for a stock exchange due to an economic crisis.

To detect a systematic change in the parameter over the range of a covariate $Z_l$, the observations are ordered according to their values of $Z_l$. Under the null hypothesis of parameter stability, no systematic structural change is present. The null hypothesis is rejected if one or more parameters of the postulated model change significantly over the ordering induced by the covariate $Z_l$.

The construction of general test statistics relies on the partial derivatives of the objective function, e.g. of the log-likelihood. The contributions of each individual observation $i$ to the derivative of the objective function (i.e., to the score function) evaluated at the current parameter estimate, $\psi\left(y_i, \hat{\vartheta}\right)$, are ordered with respect to the potential splitting variable $Z_l$. The individual contributions $\psi\left(y_i, \hat{\vartheta}\right)$ are depicted as vertical dashed lines for an instructive example in Figure 4 (left).

Under the null hypothesis, the individual contributions $\psi\left(y_i, \hat{\vartheta}\right)$ should fluctuate randomly around the mean zero, whereas in Figure 4 (left) a clear structural change can be detected. To grasp this structural change statistically, we turn from the individual contributions to their cumulative sums in Figure 4 (right). Zeileis and Hornik (2007) proved the convergence of the cumulative sum process (also termed decorrelated empirical fluctuation process)

$$W_l(t) = \hat{J}^{-\frac{1}{2}} n^{-\frac{1}{2}} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \psi\left(y_i, \hat{\vartheta}\right) \tag{3}$$

against a Brownian bridge. The first part of the formula, $\hat{J}^{-\frac{1}{2}}$, denotes an estimator of the covariance $Cov\left(\psi(Y, \hat{\vartheta})\right)$. The summation over all $\lfloor n \cdot t \rfloor$ refers to the first $n \cdot t$ (with $t \in [0; 1]$) observations according to the order with respect to covariate $Z_l$ (for example the first 50%, where the $\lfloor . \rfloor$ indicates that the integer part of $n \cdot t$, i.e. the lower whole number, is used).

The instructive example in Figure 4 can be interpreted as the variation of income before and after the simulated threshold of 40 years. The path of the cumulative sum process increases until the age of 40 and decreases after that threshold, with a sharp peak at the change-point

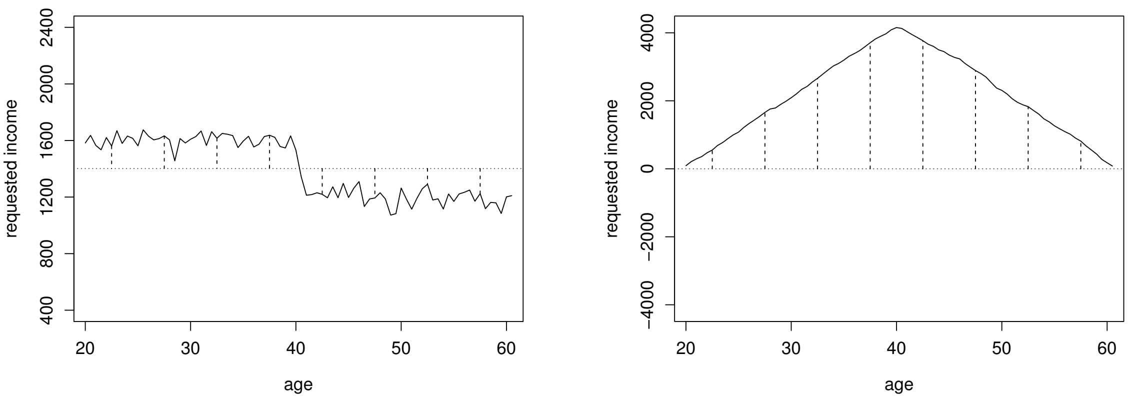

**Figure 4 –** *Structural change in the mean over age (artificial data). The left plot displays the mean income over all age groups (dotted line) and the individual deviations (dashed lines), the right the cumulated deviations over the variable age.*

40. The strength of this peak is used as a statistical measure for the strength of the parameter instability.

The asymptotic properties of the cumulative sum process allow for the construction of test statistics that are used for detecting the structural change (see Chapter 3.4). The test statistic for numeric variables is directly build from the empirical fluctuation process $W_l(t)$, while the test statistic for categorical variables takes into account that the categories and the observations within the category are not ordered. The result of Zeileis *et al.* (2008) also permits the computation of p-values and thus the statistical decision whether the parameters differ significantly from parameter stability. If parameter instability is detected, the algorithm selects the variable with the smallest p-value. Splitting continues until there is no further instability in any current node.

*Selection of the cutpoint*

In case of a splitting decision the cutpoint can be sought by a criterion that also includes the maximization of the objective function in the two potential subsamples. In the case of ordered or numeric covariates, these subsamples can easily be defined as $L(\xi) = \{i \,|\, z_{il} \leq \xi\}$ and $R(\xi) = \{i \,|\, z_{il} > \xi\}$ for a candidate cutpoint $\xi$ and the component $z_{il}$ of $z_l$.

The optimal cutpoint $\xi^\star$ is determined by maximizing

$$\sum_{i \in L(\xi)} \Psi\left(y_i, \hat{\vartheta}^{(L)}\right) + \sum_{i \in R(\xi)} \Psi\left(y_i, \hat{\vartheta}^{(R)}\right) \tag{4}$$

over all candidate cutpoints $\xi$. $\hat{\vartheta}^{(L)}$ and $\hat{\vartheta}^{(R)}$ are the estimated parameters in the subsets. In case of unordered categorical covariates all potential binary partitions need to be evaluated and the partition with the highest criterion is chosen for the split (Zeileis *et al.*, 2008).

Both parts of the binary split generate new parent nodes. Unless there is no further parameter instability found or another stopping criterion is satisfied (such as a minimum sample size in the current node) the algorithm continues searching for instability and splitting the current (sub-)data set in daughter nodes.

## 2.4 Potential in the social sciences

The application of model-based recursive partitioning offers new impulses for research in the social, educational and behavioral sciences. For the interpretation of model-based recursive partitioning, we would like to point out the connection to the principle of parsimony: following the fundamental research paradigm that theories developed in the social sciences should yield falsifiable hypotheses, the latter are translated into statistical models. The aim of model construction is thus to simplify the complex reality.

The decision on the complexity of the formulated model can be guided by "*a working rule known as Occam's Razor whereby the simplest possible descriptions are to be used until they are proved to be inadequate*" (Richardson, 1958, p. 1247). This rule implies the objective of parsimonious model formulation: a model should be no more complex than necessary, but it also needs to be complex enough to describe the empirical data.

In the regression context usually the usage of sparse and simple models with few variables explaining the response are propagated (e.g. Gujarati, 2003) – as long as no relevant explanatory variables are omitted. The strength of model-based recursive partitioning in this context lies in the power to let the data decide this question. Indeed, it offers the possibility to detect whether the suggested model is inadequate because relevant covariates are missing and it explicitly selects these relevant covariates. If the algorithm executes at least one split, we obtain the statistical decision that the parameters are instable and the data are too heterogeneous to be explained by the postulated model. In this case, the presumed functional form does not describe the entire sample in an appropriate way and thus subgroups have to be constructed.

Moreover, the tree-structured results provide information which subgroups differ in their association patterns. This information can either be integrated into a revision of the substantial theory and the formulation of a new parametric model, or it should be pointed out in the interpretation that the postulated model applies only to a limited scope of subjects.

Consequently, model-based recursive partitioning can identify different shapes of a parametric model stated by the researcher in different subgroups of the sample. Model-based recursive partitioning offers a synthesis of the theory-based and the data-driven approach that can be used for evaluating violations of the 'working rule' Ockham's Razor: if the method detects no instability of the model parameters, the model is not rejected based on the additional covariates provided to the algorithm. If, however, the method does detect instability, the postulated model is too simple.

## 2.5 Empirical example

To illustrate the potential of model-based recursive partitioning further, we turn to another example, based on an extension of the so called Mincer equation. In the seminal econometric work of Mincer (1974) the logarithmic income is described as a function of the variables years of schooling (time_edu) and full-time experience (included in linear and squared terms, full_ex, full_ex$^2$).

The Mincer equation owes its popularity to the straightforward interpretation of the coefficients as approximated rates of return from education (cf. Björklund and Kjellström, 2002, for a critical discussion). We focus on the following extension of the Mincer equation that also includes a dummy variable for further education on the job (further_edu):

$$\ln(\text{income}) = \tau_0 + \tau_1 \text{ time\_edu} + \tau_2 \text{ full\_ex} + \tau_3 \text{ full\_ex}^2 + \tau_4 \text{ further\_edu} + \varepsilon,$$

with $\varepsilon \sim N(0, \sigma^2 I)$. Here we restrict the observations from the SOEP study to over 6000 respondents in full-time employments who are not in vocational training and earn more than 500 Euros monthly.

The examination of the Mincer equation, which is driven by the principle of parsimony, via model-based recursive partitioning is illustrated in Figure 5. The model formulation involves the effects (which are displayed as symbols in the end nodes) of years of schooling, further education and work experience in full-time jobs (linear and squared term) on the logarithmic gross income of fully employed respondents in Germany. Again, further potentially influencing variables are passed over to the algorithm, namely the location of the employer in east or west Germany, gender and the size of the company. The results show significantly different parameter estimates related to each of the additional covariates. These estimated coefficients of the Mincer equation are approximated rates of return e.g. from schooling. A closer look at the estimated parameters for the detected subgroups (Table 2) shows quite similar effects on the logarithmic income for some covariates from the original Mincer equation, such as the percentaged change for the time of education on earnings ($\hat{\tau}_1$). However, the estimated coefficients for further education ($\hat{\tau}_4$) and the intercept ($\hat{\tau}_0$) differ more strongly between the groups. In particular, the effect of further education ($\hat{\tau}_4$) is higher for employers in east as opposed to west Germany.

Our results are in accordance with current empirical social and economic research on heterogeneous effects of further education for men in Germany by Kuckulenz and Zwick (2005). One reason for the violation of a joint model for all respondents may lie in the strong assumption of the Mincer equation that there is no relevant change in the economy under research. In the SOEP study, this assumption is clearly violated by the reunification of the eastern and western parts of Germany. As a consequence, we find a split according to the location of the employer in east and west Germany in Figure 5.

Our findings imply that more elaborate theories explaining the different income levels in these subgroups, such as discrimination theories (e.g. Aigner and Cain, 1977; Phelps, 1972), and a more specific investigation of the differential effects of further education may be necessary to

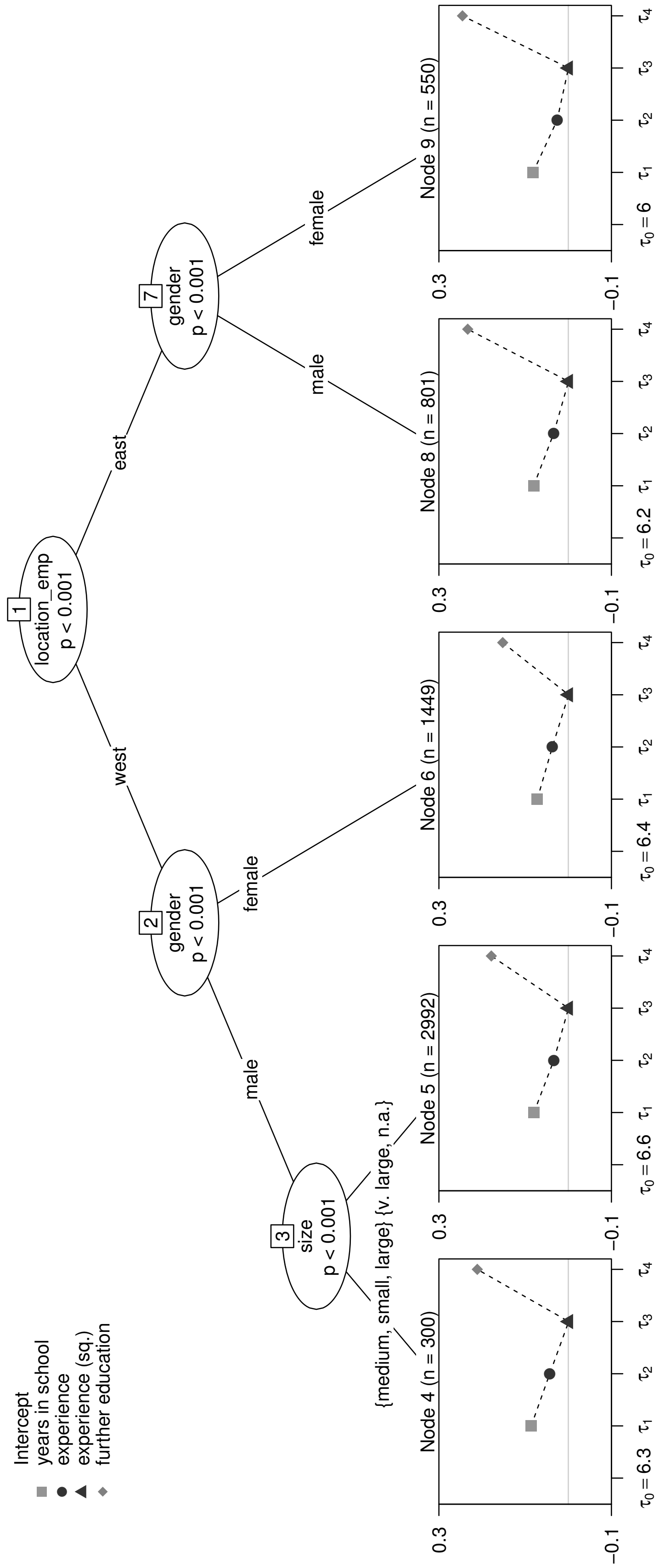

**Figure 5 –** *Model-based recursive partitioning of the extended Mincer equation (SOEP 2008). The symbols in the end nodes illustrate the estimated coefficients in the subgroups related to the location of the employer, gender and size of the company.*

| | **estimated coefficients** | | | | |
|---|---|---|---|---|---|
| **node number** | $\hat{\tau}_0$ | $\hat{\tau}_1$ | $\hat{\tau}_2$ | $\hat{\tau}_3$ | $\hat{\tau}_4$ |
| 4 | 6.2743 | 0.0860 | 0.0430 | -0.0009 | 0.2110 |
| 5 | 6.5620 | 0.0796 | 0.0335 | -0.0005 | 0.1785 |
| 6 | 6.4486 | 0.0718 | 0.0369 | -0.0007 | 0.1520 |
| 8 | 6.1543 | 0.0801 | 0.0340 | -0.0006 | 0.2332 |
| 9 | 6.0173 | 0.0817 | 0.0258 | -0.0004 | 0.2454 |

**Table 2** – *Estimated coefficients of the regression models in the end nodes in Figure 5 estimated from the data sets that correspond to the end nodes.*

explain the observed group differences.

## 2.6 Software

The data analysis presented here uses the R system for statistical computing (R Development Core Team, 2011), which is freely available under terms of the GNU General Public Licence (GPL) from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. Methods for classification, regression and model-based trees are provided in the package `party`.

- *Conditional inference trees*

  The conditional inference framework is implemented in the function `ctree()` (Hothorn *et al.*, 2006b). It allows to compute both classification and regression trees.

- *Model-based recursive partitioning*

  The model-based recursive partitioning algorithm is available via the function `mob()` (Zeileis *et al.*, 2008). At the moment the algorithm can be applied to various types of generalized linear models, survival models or linear models. Moreover, the authors allow the users to build their own model classes and pass them on to the existing `mob()` function. A vignette explaining the use of the software for linear regression and logistic regression trees including the R-code is also available (Zeileis *et al.*, 2010).

Ongoing research expands the unbiased recursive partitioning approach presented here to psychometric models such as the Bradley-Terry model for detecting different preference structures (Strobl *et al.*, 2011) as well as the Rasch model (Strobl *et al.* 2010, Strobl *et al.* 2013), that will be addressed in the next chapter, and factor analytic and structural equation models (Merkle and Zeileis, 2012) for the assessment of measurement invariance.

## 2.7 Concluding remarks

Algorithmic procedures, such as classification and regression trees, have become popular and widely used tools in many scientific fields. Our aim here was to highlight that the recent development of model-based recursive partitioning allows to combine the power of these algorithmic methods with that of theory-based parametric models by means of enhancing the purely data-driven approach towards a segmentation procedure for postulated models. Our presentation has highlighted the relation between this approach and the principle of parsimonious model construction. The tree-structured results allow straightforward interpretations of potential parameter instabilities that have been detected via empirical fluctuation tests. The detection of parameter instability leads to the interpretation that the statistical model under investigation cannot describe the whole sample appropriately, because relevant covariates have been omitted. Thus, model-based recursive partitioning can be used as a diagnostic check for inadequately simple descriptions of the relationship between response and explanatory variables.

The application in social science research is eased by the freely accessible and well documented packages provided in the statistical software R.

# 3 The issue of differential item functioning in the Rasch model

***Summary:*** *The analysis of differential item functioning (DIF) in item response theory (IRT) research investigates the violation of the invariant measurement property among subgroups of examinees, such as male and female test-takers. Invariant item parameters are necessary to assess ability differences between groups in an objective, fair way. Questions addressed in this chapter are: What does DIF mean? What are potential consequences if DIF is present? How can DIF be detected?*

*The current state of research cannot be covered completely in the following sections. The introductory section raises historical and political aspects of DIF and is intended to motivate the analysis of DIF. Therefore, instructive examples as well as the main idea of commonly used statistical methods for DIF detection are included. The reader is referred to the monograph* Differential Item Functioning *by Osterlind and Everson (2009), the like-titled collection by Holland and Wainer (1993) for more detailed information on DIF and to the broad collection* Rasch Models - Foundations, Recent Developments, and Applications *by Fischer and Molenaar (1995).*

***Keywords:*** *Differential Item Functioning, item bias, test fairness, statistical methods for DIF detection*

## 3.1 The Rasch model

In the social and behavioral sciences, researchers are often confronted with the fact that the variable of interest is not directly observable. In sociology or in political science, attitudes towards certain political or social systems or subjects are not observable similar to personality traits in psychology or abilities and skills in educational research. Following the idea that the response behavior of test-takers contains information on the *latent variables*, models from item response theory (IRT) are intended to measure variables that are not directly observable. The Rasch model (Rasch, 1960) is a widely used IRT model that displays unique statistical properties (Wang, 2004). When the assumptions of the Rasch model for measurement hold, measures of both, the item difficulty and the person ability, are generated on an interval scale level with a common measurement unit (Fischer, 1995). Furthermore, the assumptions of the model can be tested and, thus, the measurement process can be evaluated (Molenaar, 1995b).

In the Rasch model, the variable $U_{ij}$ contains the dichotomous information whether the item or the statement is solved or agreed to ($U_{ij} = 1$) or not ($U_{ij} = 0$) with the respective observed value ($u_{ij} \in \{0, 1\}$). Here, $\theta_i$ denotes the *latent trait*, e.g. the attitude, the ability or the personality trait of person $i$ ($i = 1, ..., n$) and $\beta_j$ denotes the *difficulty* of the item (also termed *item location*) or the extremeness of the statement $j$ ($j = 1, ..., k$). For convenience, the latent trait will be called ability and the answer will be termed as solved an item ($u_{ij} = 1$) or not ($u_{ij} = 0$) in the following. However, note that the Rasch model can be used to measure latent traits that are not limited to variables measuring abilities or skills and it may provide useful insight to evaluate the appropriateness of the items or statements for measuring the latent trait in an objective, fair way.

In the Rasch model, the response of subject $i$ to item $j$ is modeled by

$$P(U_{ij} = u_{ij}|\theta_i,\beta_j) = \frac{e^{u_{ij}\cdot(\theta_i-\beta_j)}}{1+e^{\theta_i-\beta_j}} \tag{5}$$

and, thus, depends on two types of parameters, the item difficulty parameters $\beta_j$ and the person ability parameters $\theta_i$. The relationship between the latent trait and the probability of solving the item can be displayed in the so called *item characteristic curves* (ICCs) or *trace lines* (Thissen *et al.*, 1988) or *item response functions* (IRF) (Kim *et al.*, 1994) as illustrated in Figure 6. Here, four ICCs are displayed. In the Rasch model, the ICCs do not cross and all items are assumed to have to same discriminatory power. The item difficulty is defined as the value of the latent trait where the probability of solving the item equals .5.

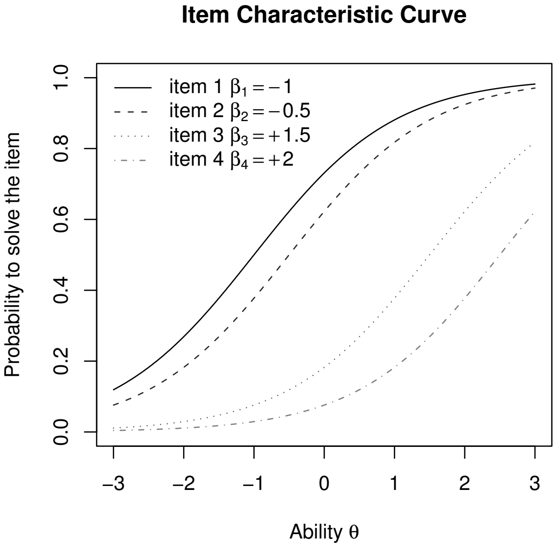

**Figure 6** – *Different item characteristic curves that follow the Rasch model.*

The ICCs in Figure 6 reflect the stringent assumptions of the Rasch model. Other IRT models for dichotomous responses such as the two-parameter logistic (2pl) model

$$P(U_{ij} = 1|\theta_i, \alpha_j,\beta_j) = \frac{e^{\alpha_j\cdot(\theta_i-\beta_j)}}{1+e^{\alpha_j\cdot(\theta_i-\beta_j)}} \tag{6}$$

or three-parameter logistic (3pl) model allow for more flexible forms of the ICCs by e.g. allowing the items to have different discriminatory powers. In empirical research, thus, the construction of a scale conform with the Rasch model is more difficult. The appropriateness of the Rasch model is controversially discussed. While opponents claim that the Rasch model was inappropriate for educational testing in the past as well as will be in the future (McLean and

Ragsdale, 1983), proponents argue that the benefits of the Rasch model (regarding the estimation process, the model evaluation and the interpretation) in comparison with the more flexible IRT models outweigh the higher effort in practical data analysis and, thus, the aim in practical research should be to find items in accordance with the Rasch model (e.g. Molenaar, 1995b).

The foundation of the Rasch model can be carried out from different perspectives. While often the model formula forms the starting point to derive estimation equations, Fischer (1995) shows that the assumptions of strictly monotone increasing ICCs of an unidimensional IRT model with the upper limit of 1 and the lower limit of 0 (that excludes the possibility of guessing), of local stochastic independence (the items are not allowed to be built on other items and the persons are not allowed to copy answers) and of the sufficience of the person raw-scores $r_i = \sum_{j=1}^{k} u_{ij}$ for the latent trait yield a resulting IRT model that is equivalent to the Rasch model where, then, the ICCs include the substractive form $\theta_i - \beta_j$. The author further shows that the substractive form implies that $\theta$ and $\beta$ are measured on metric scales (with the same measurement unit) and that $\theta$ and $\beta$ are unique except for linear transformations $\alpha\theta + b_1$ and $\alpha\beta + b_2$ with $\alpha > 0$ that yield the formula of the family of Rasch models

$$P(U_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\alpha\cdot(\theta_i - \beta_j)+b}}{1 + e^{\alpha\cdot(\theta_i - \beta_j)+b}} \tag{7}$$

with a common discrimination parameter $\alpha$ and $b := b_1 - b_2$ representing the fact that shifts on the $\theta$ scale are balanced by the corresponding shifts on the $\beta$ scale (Fischer, 1995). Further derivations for the Rasch model that rely on assumptions related to concept of sufficiency, such as a derivation from the stochastic consistent ordering or on the specific objectivity and likelihood principle, are reviewed in Fischer (1995).[1]

In the following, the estimation will be discussed only for complete cases. Missing values in the Rasch model may occur due to the test design when, for example, different test booklets contain different questions for the respondents. These missing values result per design since, of course, respondents will not answer questions that have not been posed (for the resulting consequences in the estimation process see Molenaar, 1995a). Missing values may also occur when respondents refuse answering certain questions (for item non-response, see Schnell *et al.*, 2008). In situations of ability testing, it is likely that values are not missing at random. Since the aim is to measure ability through the correct responses to items, one possible way of handling missing values is to define these answers as not solved and, consequently, wrong. In situations of attitude measurement, the refusal might be caused by social desirability and especially delicate questions might be affected. Thus, the response mechanism might as well be related to item extremeness and to the latent attitude, and, consequently, not be missing at random (Molenaar, 1995a). Furthermore, in attitude measurement related problems such as acquiescence where respondents tend to answer "yes" or "I agree" or socially expected answers might occur, that can result in untruthfully answered items (Diekmann, 2007). In this thesis, ability measurement is focused and it is assumed for the simulation studies that no missing values are present. In

[1] The author further states that the interpretation of resulting measurement as difference scales for $\theta$ and $\beta$ – where the scales can be shifted only by an additive constant – only holds when the ICCs are assumed to have a common discrimination parameter $\alpha$ of the value one.

situations where missing values occur by design, they can be included in the estimation process (Mair and Hatzinger, 2007).

The estimation of the item parameters can be carried out using several different estimation techniques. The joint or unconditional maximum likelihood estimation relies on the product of (5) over all persons and items by applying the assumption of local stochastic independence (see e.g. Molenaar, 1995a). When the derivation of the log-likelihood is set to zero, it leads to (under complete data at most $2k - 2$) estimation equations that have to be solved iteratively. Unfortunately, the estimates are inconsistent for $n \rightarrow \infty$ and $k$ fixed and lack properties for hypothesis testing (Molenaar, 1995a). The conditional maximum likelihood (CML) estimation is preferable since it leads to consistent item parameter estimates for $n \rightarrow \infty$ and $k$ fixed (see Molenaar, 1995a, and the references therein for the assumptions).

In comparison with the alternative marginal maximum likelihood (MML) approach, that relies on either a parametric distributional assumption or non-parametric estimation of the distribution of the latent trait, the CML approach is "*closer to the concept of person-free item assessment*" (Molenaar, 1995a, p. 47). Here, the item assessment is given priority compared to the assessment of ability distributions of specific groups of subjects, where the MML approach is said to be more suitable (Molenaar, 1995a) and, thus, the CML estimation is described here in detail (for the above mentioned estimation techniques or other approaches that yield no asymptotic standard errors, see e.g. Molenaar, 1995a).

It is assumed here that all items and persons with zero or perfect sums, that would result in infinite parameter estimates (Molenaar, 1995a), are removed before the estimation is carried out. The existence and uniqueness of the CML estimates (i.e. of the normalized estimates defined by a common discrimination of one and $\sum_{j=1}^{k} \beta_j = 0$) requires a so called well-conditioned response matrix as discussed in Fischer (1981).[2] The person raw-scores $r_i = \sum_{j=1}^{k} u_{ij}$ that form sufficient statistics for the person parameters under the Rasch model, are then used to estimate the item parameters by means of iterative procedures from the conditional likelihood

$$L_c(\beta | r_1, \ldots, r_n) = \prod_{i=1}^{n} L_c(\beta | r_i) = \prod_{i=1}^{n} \frac{e^{-\sum_{j=1}^{k} u_{ij} \cdot \beta_j}}{\gamma_{r_i}(\beta)}, \tag{8}$$

where $\gamma_{r_i}$ denote the elementary symmetric function of order $r_i$ (cf., e.g., Fischer, 1974, 1981; Molenaar, 1995a). One linear restriction has to be imposed to identify the item parameters. Generally, the sum-zero-restriction $\sum_{j=1}^{k} \beta_j = 0$ or the set-the-first-zero restriction $\beta_1 = 0$ are used. More generally, Eggen and Verhelst (2006) state the equation

$$d_0 + \sum_{i=1}^{k} d_i \beta_i = 0\,, \tag{9}$$

allowing arbitrarily chosen constants $d_i$ with $\sum_{i=1}^{k} d_i \neq 0$.

[2] This well-conditioned response matrix is almost surely present when the sample size converges to infinity (Fischer, 1995). In practical research situations, computer programs such as the function `RM()` provided by the R add-on package `eRm` by Mair *et al.* (2012) states an error message if the matrix is not well-conditioned.

The authors specify two commonly used normalization equations under the assumption $d_0 = 0$ (w.l.o.g.): Firstly, one item parameter may be set zero per definition ($\beta_i = 0$, for one $i \in 1, 2, \ldots, k$). Equation 9 is fulfilled since all $d_{j \neq i} \overset{!}{=} 0$ despite $d_i = 1$, $\sum_{i=1}^{k} d_i \neq 0$ and $\sum_{i=1}^{k} d_i\beta_i = 0$. The second common practice to restrict the item parameters is by defining a index set $\mathcal{I}$ containing indices from $1, 2, \ldots, k$. Then the linear restriction is defined by $\sum_{i \in \mathcal{I}} \beta_i = 0$ or on the mean of the item parameters.

## 3.2 Differential item functioning

At the beginning of the discourse about different test properties in the 1960s (Angoff, 1993), the problem was referred to as *item bias* (cf., e.g., Lord, 1980). Since the term *bias* has been applied in an ambiguous way, it was replaced by the expression *differential item performance*, *unexpected differential item performance* or by the – contemporary commonly used – term *differential item functioning* or the abbreviation *DIF* (Angoff, 1993; Thissen *et al.*, 1988).

Related concepts to DIF exist. *Measurement invariance* (also termed *measurement equivalence*) represents the absence of DIF and is an important requirement of IRT models. There might be a tendency of favoring the term *measurement equivalence* instead of DIF at the moment (Paul De Boeck 2011, personal communication).

However, in this thesis, the term DIF is preferred since it better reflects the fact that items function in a different way for certain groups. The term *impact* is used for differences in the performance of both groups for one item when no matching is carried out (Holland and Thayer, 1988). A conceptualization of test bias and differential test functioning can be found in Shealy and Stout (1993a, p. 159). The distinction of *measurement equivalence* and *relational equivalence* is explained in Drasgow (1987, p. 19).

### 3.2.1 Historical development and test fairness

In order to illustrate the "*testing controversy*" in the United States, Drasgow (1987) summarizes two court proceedings. In 1984, the Educational Testing Service (ETS) was sued since black participants failed more often in insurance licensing tests which were designed by the ETS. As a consequence of the lawsuit, the ETS agreed outside court to modify the test items in a way such that the percentages of correct answers differ no more than 15% between black and white test-takers (known as the "golden rule", see Gómez-Benito *et al.* 2010). Similarly, a lawsuit from the year 1985 against the Alabama State Board of Education resulted in the obligation to use only test items that differ no more than 5% between black and white test-takers regarding the percentages of correct answers (Drasgow, 1987).

Drasgow (1987) states that the specifications of the test items cannot be justified scientifically but politically. The reason for this is that the analysis of certain characteristics of test items by the current state of scientific knowledge is always conditioned on the variable of interest (e.g. the ability variable). If two groups of participants have different mean levels on the latent trait, different percentages of correct answers are expected and even intended. Otherwise, the information about the probabilities of solving the item does not allow to draw conclusions about the latent trait. On the other hand, if two groups of participants have the same latent ability, the probabilities of solving the item and, hence, the observed proportions of correct answers should not differ substantially. Thus, the analysis of DIF requires to match the examinees on their ability variable, as will be discussed in more detail in Section 3.3.

This was not always the case. Early attempts to investigate DIF also used the percentages of correct answers in the groups. These proportions were depicted against each other. In the case of equally functioning items, the points should lie on a line that begins at a chance level (c,c) and ends in (1,1), where items are on a difficulty level such that each respondent may solve it, or on a curve if one group performs better. Transformations have been used to restructure the functional form (Lord, 1980). Nevertheless, the percentage of correct answers per item with respect to a certain group membership is not an appropriate measure of DIF, since it does not allow to distinguish properties of the item and of the groups themselves and "[t]*hus the* $\pi_i$ [the proportions of correct answers to item i in the population], *however transformed, are not really suitable for studying item bias*" (Lord, 1980, p. 217). In the following, only DIF methods that condition on ability are discussed.

Even though first approaches to DIF analysis have already been made half a century ago, the ongoing research on DIF methods continues. Finch and French (2007, p. 565) explain this fact "*due to the increased reliance on standardized achievement testing for assessing educational progress*". A well-known example supporting this argument is the OECD Programme for International Student Assessment (PISA) initiated in 1997. The aim is to assess the preparation of 15 year old students for future challenges (OECD, 2009). The number of participating countries[3] increased from 43 countries in 2000 to 64 countries in 2012. Even if the PISA study is predominant in the media, many other large-scale assessments are conducted, too. Standardized large-scale testings in Germany under participation of the German Institute for International Educational Research[4] (DIPF, Deutsches Institut für Internationale Pädagogische Forschung) for example include the National Educational Panel Study (NEPS), the study "*Persönlichkeits- und Lernentwicklung von Grundschulkindern*" (PERLE, a study about personality and learning development of primary school children), the Programme for the International Assessment of Adult Competencies (PIAAC) and the study "*Deutsch Englisch Schülerleistungen International*" (DESI, a study of student achievements in German and English). Thus, important political decisions depend on standardized testing results. Predominantly in the United States, also personal occupational histories are influenced by standardized achievement tests such as academic assessments. In Germany, academic assessments are less common, but e.g. test results from the TOEFL test of the ETS are sometimes required for the admission to master programs. In Spain, examples where test results are used include the participation in intervention programs and the assessment in human resources as well as in universities (Gómez-Benito *et al.*, 2010). Thus, test items need to be subject to an extensive analysis before they are used in real testing situations. The aspect of test fairness will be addressed in the next paragraph.

[3] http://www.oecd.org/pisa/participatingcountrieseconomies, access on 14.01.2013

[4] http://www.dipf.de/de/themen/large-scale-assessment, access on 14.01.2013

The problem of DIF is closely related to the concept of test fairness. When a test is intended to measure a certain ability such as maths skills, the items should only measure the intended variable. A common perception of justice follows the equity principle wherein rewards are distributed according to the input. Following this principle, it is considered fair, if a respondent obtains a higher test score if he or she has a higher ability compared to another test-taker. Similarly, if groups of subjects have different mean abilities, it is considered fair, if these groups display different mean test scores. In contrast to this, it is not considered fair, if equally able test-takers from different groups obtain different test results or if the differences in the test scores occur for equally able groups.

Therefore, the analysis of DIF focuses on the question whether equally able test-takers display different probabilities of solving an item. Tests that are constructed to measure maths skills, for example, may include word problems. It may happen that word problems are harder to solve for participants with a different mother tongue. In this case, the item is no longer unidimensional and native speakers obtain an unfair advantage in the test. If we would compare equally able groups of native and non-native speakers, a higher probability of solving the item would be expected for the former group and, correspondingly, a larger proportion of correct answers. If all other items of the math test function similarly for the groups (i.e. are invariant), a higher test score would be expected for the native speakers and a higher math ability would be reported even if the groups are in fact equally able. This situation reflects an unfair advantage in the math test.

This example illustrates the relationship between DIF and differential test functioning (DTF). The latter is defined by a lack of invariance on the scale or test level (cf. Stark *et al.* 2006, and the references therein, or Shealy and Stout 1993a). If items from a test display DIF, the test may function differently for the groups. This is the case if the DIF effects across all items do not cancel out on the test level. Hence, DIF and DTF are present. Otherwise, if the DIF effects cancel out on the test level, DIF is present but DTF is not (Stark *et al.*, 2006). Even if the possibility of cancellation effects exists, DIF analysis has to be carried out to not only legitimate the test items but also to examine the underlying causes of DIF (Jodoin and Gierl, 2001).

The examples from Section 3.2.1 are now addressed with respect to test fairness. Going back to the lawsuits, the difference in the proportions of test failures between black and white respondents per se cannot be considered as unfair, since no information on the latent variable is taken into account yet.

Only an analysis that conditions on the latent variable is able to distinguish between true ability differences and DIF: When a black and a white test-taker have the same latent ability, but still perform differently on some of the test items, the test is considered unfair. One reason for the underperformance of black test-takers given the same ability level could be the so called stereotype threat (for more details see, e.g., Wicherts *et al.*, 2005).

Returning to the PISA study, it is not intended to report individual test results, but focuses on comparisons between the participating countries – with major political impact – and between male and female students. Therefore, measurement invariance is important in the PISA study to allow for test fairness. Before items are administered in the final test set, characteristics of the items are studied in pretests and DIF tests are conducted regarding effects of the gender and the country variable (Joachim Funke, Chairman of the International PISA Problem Solving Expert Group, 2011, personal communication).

Test fairness is related to test validity, that is often addressed as an argument for the importance of testing the assumptions of IRT models. Gómez-Benito *et al.* (2010, p. 76) refer to the statement of a test as valid as the claim that "*the score obtained has a specific meaning, assuming that this meaning is the same in the different groups for which the test has been validated*". The presence of DIF-items "*on a test is a threat to validity*" (Cohen *et al.*, 1996, p. 15). Accordingly, Magis and De Boeck (2011) highlight that identifying and excluding DIF-items is necessary for validity. Alternatively, modifications of the DIF-items may be considered (Kelderman and MacReady, 1990).

Ackerman (1992, p. 69) explains this relationship between DIF items, unwanted multidimensionality and validity more precisely: "*If a test lacks construct validity, it contains items that are measuring skills other than those purported to be measured, and, hence, the potential for item bias also exists. This bias may be realized if groups of interest differ in their underlying distribution of these extraneous skills. Simply put, items invalid in the construct sense are a necessary but not sufficient cause of item bias.*" This is an important argument since the lack of validity implies that wrong conclusions might be drawn. The variable of interest is not measured appropriately and the measurement includes distracting information from other nuisance dimensions.

In summary, a given educational or psychological test consisting of many items with substantial DIF may be unfair for certain subgroups and lacks validity, and it is important to identify these items, so that they can be improved or deleted from the test (Westers and Kelderman, 1992). With a rising reliance on standardized achievement test results, DIF analysis gets even more important and sophisticated methods to detect DIF are required.

### 3.2.2 Uniform and non-uniform differential item functioning

DIF exists if items show different measurement characteristics for at least two groups of respondents, given the latent trait (Woods, 2009) or if matched test-takers from one group are favored over those from another (Shealy and Stout, 1993a). An item is unbiased if the ICCs (for the definition of the ICCs, see again Section 3.1) are identical for all groups. Then, respondents with the same ability value have exactly the same probability of solving each item, regardless of any group membership (Lord, 1980). In contrast to this, DIF occurs, if the probability to solve an item varies for at least two groups given the same latent ability.

Large parts of this thesis focus on the comparison of two groups, as is usually done in research on DIF. These groups are termed the *reference* and the *focal* group. In the substantial sciences, the term focal group refers to the group the researcher is interested in, whereas the reference group is the standard group (Holland and Thayer, 1988), often a majority group with which the focal group – often a disadvantaged minority – is to be compared (Gómez-Benito *et al.*, 2010).

The assignment of reference and focal groups in this thesis is arbitrary and, thus, unprejudiced and without the expectation of one group performing better.

The variable defining the groups is called *external variable* or *background variable* in the literature (Glas, 1998). It should be noted that (as with all observational data) the background variable used to define the groups cannot simply be interpreted as the causal source of the observed DIF, because the background variable may only serve as a proxy for the unobserved (and potentially unobservable) true cause. For example, if DIF is detected between men and women, gender should not be considered as the actual cause of the DIF, but as an indicator of a variety of educational and social influences (Strobl *et al.*, 2013).

*Example: An item that functions differently*

Now, an example of a DIF-item is considered. Dorans (1989) compared male and female test-takers and presented a highly differentially functioning item from the originally termed *Scholastic Aptitude Test* (SAT) study. The SAT from the College Board is used for college admissions in the USA. In the example, the task is to build verbal analogies:

✎ DECOY:DUCK = □ net:butterfly
□ web:spider
□ lure:fish
□ lasso:rope
□ detour:shortcut

In this case, female respondents display a lower probability of solving the item. This does not necessarily imply DIF. It could also result from a lower ability. Therefore, the author distinguishes between impact and DIF (termed *unexpected differential item performance*). Under the control of ability – by stratifying the sample in ability groups according to the observed SAT score – male respondent have a 15% to 20% percent higher chance of solving the item in regions of 250 – 500 points of the SAT score. The explanation at hand for this effect is additional knowledge of terminology in hunting and fishing.

So far, the definition of DIF only stated that the ICCs (or IRFs) differ between the groups (Kim *et al.*, 1994). However, the ICCs can differ in various forms: A DIF-item may favor one group over the entire latent space or the ICCs of reference and focal group may cross (Lord, 1980). A common classification of the type of DIF, which will be presented in the next paragraphs, distinguishes between uniform and non-uniform DIF.

In this thesis, DIF is defined as *uniform* DIF if two conditions are met: Firstly, one group has a higher probability of solving an item (given the latent trait) over the entire latent continuum and, secondly, the group differences in the logits, i.e. the logarithmic odds of correctly answering the item, remain constant over the ability levels (Mellenbergh, 1982). This definition is consistent with the definition used by Finch and French (2007).

The term uniform DIF is not consistently used. Jodoin and Gierl (2001) and French and Maller (2007) use only the first condition to declare uniform DIF. This situation is referred to as *uni-directional* DIF in this thesis and will be addressed in the following paragraphs. To use the second condition is motivated from a modeling perspective: Mellenbergh (1982, p. 114) describes uniform DIF as "*a constant difference between group-item performance over ability levels*". Thus, if the group differences on the logit scale remain constant, uniform DIF can be grasped statistically in differing item difficulty parameters in the Rasch model that are already conditioning on ability levels.

Furthermore, Mellenbergh (1982) explains unidimensional DIF with regard to the multidimensionality aspect, in case of e.g. two dimensions, uniform DIF occurs when the additional trait difference is constant over the prime dimension.

*Example: An item that displays uniform DIF*

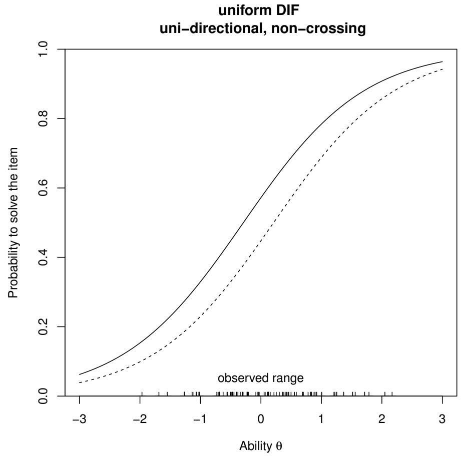

**Figure 7** – *An item displaying uniform DIF (artificial data).*

In this example, uniform DIF is represented by displaying the ICCs, which connect the ability to the probability of solving the item. In Figure 7, the ICC for the reference group (solid line) and the ICC for the focal group (dashed line) do not cross each other and the differences in the logits remain constant. The reference group has an unfair advantage, since the probability of solving the item is higher, even if both groups have the same ability.

Mellenbergh (1982), who is said to be the first researcher to conceive the classification of uniform and non-uniform DIF (Li and Stout, 1996), approached the situation of non-uniform DIF by considering three-way contingency tables. They consist of the group membership indicator, the ability level (approximated through a test score) and the interaction of both variables. Mellenbergh (1982) defines items that require the interaction term to fit the data adequately as non-uniformly biased. Thus, *non-uniform* DIF is present when the group differences in the logit are no longer constant but differ over the range of the ability levels. Thus, from a statistical perspective, non-uniform DIF corresponds to an interaction effect between the group membership and the ability variable on the probability of solving the item (Mellenbergh, 1982).

According to Li and Stout (1996), situations of non-uniform DIF may be further characterized according to whether the ICCs cross in the range of the observed ability levels (i.e. only those values that were actually in the sample), what is referred to as *crossing* DIF, or not, what is termed *non-crossing* DIF. Furthermore, situations of uniform or non-uniform DIF where one group has an advantage over the observed ability levels are referred to as *uni-directional* DIF as opposed to *non-directional* DIF as illustrated in Figure 8 (Shealy and Stout, 1993a; Li and Stout, 1996; Finch and French, 2007).

*Example: Items that display uniform and non-uniform DIF*

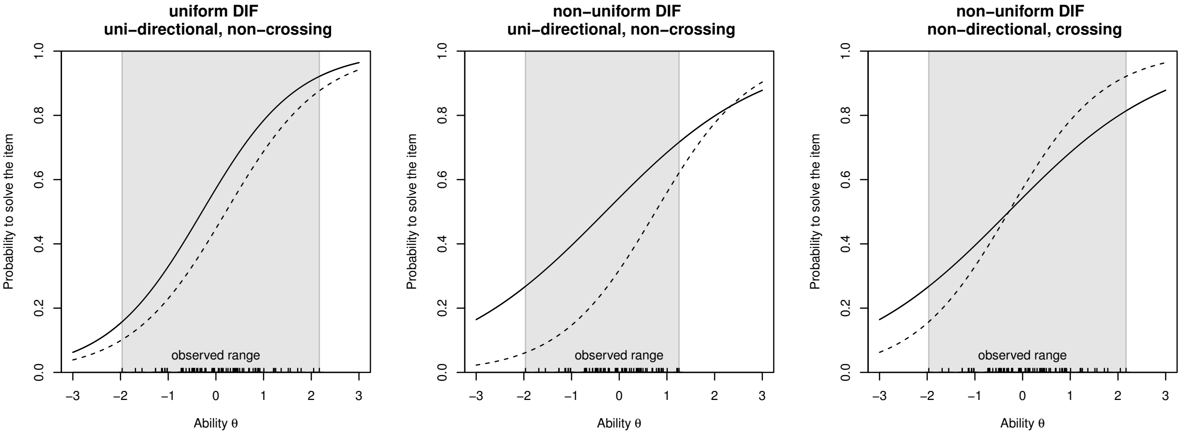

**Figure 8 –** *Items displaying different types of DIF (artificial data).*

Figure 8 (left) again displays the situation of uniform DIF. The ICCs do not cross each other at any point of the x-axes and, thus, uniform DIF is non-crossing and uni-directional. Figure 8 (middle) represents a situation of non-uniform DIF. The ICCs cross each other but not within the range of the observed ability. Hence, in the region where data was observed, the reference group (solid line) has an unfair advantage by yielding a higher probability of solving the item given the ability variable compared to the focal group (dashed line). Thus, the situation is uni-directional and non-crossing. In the right panel of Figure 8, the ICCs do cross within the observed ability range, and hence, crossing DIF is present. Crossing DIF corresponds to the situation where DIF no longer favors one group but the advantage depends on the ability level; In lower regions of the ability variable, the reference group has an advantage (i.e. a higher probability of solving the item given the ability variable), whereas in higher regions of the ability variable, the focal group has the advantage. This situation is referred to as non-directional.

Generally, situations with multiple crossing points at various ability levels are possible. However, as cancellation effects are conceivable, the main research focuses in one-point-crossing at ability levels with a high density (Li and Stout, 1996). No difference between crossing and non-uniform DIF is made by Finch and French (2007). The author defines non-uniform DIF when items discriminate differently for groups across the ability continuum. The approach of modeling non-uniform DIF by means of varying discrimination parameters is common in IRT-based DIF analysis.

As DIF generally corresponds to situations where the latent trait is no longer unidimensional, non-uniform DIF occurs in situations where the nuisance dimension depends on the target ability or, put differently, the differences on the nuisance dimension now vary over values of the prime dimension (Mellenbergh, 1982).

Mellenbergh (1982) for example found a considerable proportion of non-uniformly biased items in a word analogies test and highlights the importance of both types, uniform and non-uniform DIF, to compare groups appropriately.

## 3.3 Methods to detect DIF or DIF groups

A variety of methods has been proposed for the analysis of DIF in the Rasch model. Here, the review is limited to some commonly used DIF detection methods and does not cover all proposed methods (for more detailed overviews see Osterlind and Everson 2009, Holland and Wainer 1993 and Fischer and Molenaar 1995). Firstly, global goodness-of-fit tests for the Rasch model, that also allow to test for DIF, are briefly discussed. Secondly, methods to detect uniform DIF in individual items in two or more pre-defined groups are regarded. Thirdly, methods to detect item-wise non-uniform DIF in two or more pre-defined groups are discussed. Lastly, Rasch trees, that combine model-based recursive partitioning and IRT, are introduced in Section 3.4. Rasch trees proceed in a different way: they do not only allow to detect DIF but also determine the groups that display DIF.

In the following, only methods that condition on ability and, thus, focus on DIF instead of impact are regarded. Flier *et al.* (1984) discuss unconditional and conditional methods and list arguments in favor of the latter: unconditional methods focus on impact and, thus, depend on the true ability distributions as well as on other items, since bias is defined relatively. The dependence for conditional methods is reduced to the matching criterion for the estimation. The problem of defining a common metric for the item parameters of the groups that are to be compared when conditional item-wise DIF methods are employed is a main aspect of this thesis (see Chapter 5 through Chapter 8).

### 3.3.1 Global model tests

Global model tests are employed to answer the question of whether DIF is present anywhere in the entire set of test items and, thus, do not allow to specify directly which items display DIF.

One global goodness-of-fit tests for the Rasch model is the widely used likelihood ratio (LR) test (Anderson, 1973; Gustafsson, 1980). The LR test was first suggested by Anderson (1973) as a test for different slope parameters by means of dividing the subjects into groups according to their ability raw scores. However, it has long been noted (e.g. by Gustafsson, 1980) that it can also be used as a test for DIF when the subjects are divided into groups according to covariates such as gender and social background. The test then compares the item parameter estimates between two or more pre-specified groups of subjects, such as males and females, as reference and focal group. The parameters need to be estimated both for the full sample and for all subsamples. The full sample likelihood for the full sample item parameter estimates is then compared to the product of all subsample likelihoods for the subsample item parameter estimates by calculating minus two times the logarithmic likelihood ratio which follows a $\chi^2$-distribution with the degrees of freedom representing the number of additional parameters for the unrestricted model.

Similarly, the Wald test could be used for this situation. The item parameters need to be estimated for all subsamples only. The subsample item parameter estimates are then directly compared, so that the Wald test does not require the computation of the full sample item parameter estimates. Also the Lagrange multiplier (LM) or score test employs only the item parameter estimates from the full sample, and evaluates group differences by means of the individual score contributions. Asymptotically all three tests are equivalent and hence, in practice, the choice of test is often guided by computational considerations: The LR test is often found to perform slightly better in finite samples; however, it also poses the highest computational burden (see also Merkle and Zeileis, 2012, who discuss similarities and differences of the three types of tests in more detail in a structural change setting). All three tests rely on the estimation of an IRT model and are referred to as IRT-based tests in the literature.

Another way to approach a global goodness-of-fit test is the latent class (or mixture) approach of Rost (1990). It tests for item parameter differences between all possible groups of subjects regardless – and even in the absence – of person covariates (see also Kelderman and MacReady, 1990; Mislevy and Verhelst, 1990). In this sense, the latent class approach can be considered as a very stringent model test (even though it has a lower statistical power than tests for given groups when informative covariates are available, cf. Smit *et al.*, 2000). However, the latent class approach provides no straightforward interpretation of the resulting groups. Therefore, often latent class approaches are used only as a first step in the analysis, where the second step is to attempt to describe the latent classes by person covariates for interpretability (see, e.g., Cohen and Bolt, 2005; Hancock and Samuelsen, 2007; Maij-de Meij *et al.*, 2008, and the references therein).

### 3.3.2 Methods to detect item-wise uniform DIF

Item-wise DIF tests are intended to answer the question of whether an individual item displays DIF. A variety of methods to test for DIF is available. Some of these methods have been adopted from statistics or biostatistics to analyze DIF in the Rasch model, while other methods have directly been developed for the purpose of investigating DIF.

One IRT-based method adopted from statistics is the use of the item-wise Wald test to assess DIF in a single item. Frederick M. Lord (1912-2000), who was called the *father of modern testing*, first suggested to use the Wald test for the assessment of different item parameters (Lord, 1980). The Wald test is, thus, often referred to as Lord's test statistic (Thissen *et al.*, 1988) and was generalized for multiple groups by Kim *et al.* (1995). The idea behind this approach is that DIF is present if the ICCs are not identical across groups (Mellenbergh, 1982). When the Rasch model is considered, this is translated to the null hypothesis for uniform DIF, that states the equality of the item difficulty parameters in the Rasch models that are estimated separately for both groups. The same hypothesis can be tested with the item-wise likelihood ratio test proposed for DIF analysis by Thissen *et al.* (1988) and available in the software program IRTLRDIF (Thissen, 2001). A LM DIF statistic for MML-estimation is discussed by Glas (1998).

Another example for the adoption of a statistical test is the Mantel-Haenszel (MH) test, originally developed in biostatistics (Mantel and Haenszel, 1959). The MH test is a non-IRT test statistic, since it is not necessary to compute an IRT model for the DIF test. Instead of using e.g. CML-estimated item parameters, the matching criterion used is the total test score, i.e. the sum of correctly answered items. Contingency tables are constructed for the potential DIF-item including only respondents with a certain test score. All tables resulting from the number of possible raw test scores are regarded and the expected frequencies are compared to the observed frequencies using the so called common odds ratio. The resulting asymptotic distribution is $\chi^2$ (cf. Holland and Thayer, 1988, also for further statistical properties). The MH procedure is not the only test that has been proposed for the investigation of matched tables (see, e.g. Mellenbergh, 1982; Scheuneman and Bleistein, 1989), but it is widespread in DIF research. An introduction to related DIF methods as well as the MH estimate for the common odds ratio are presented by Holland and Thayer (1988).

Swaminathan and Rogers (1990) proposed the usage of the logistic regression to detect uniform or non-uniform DIF in one item. To accomplish this, the linear predictor is specified including an intercept, the ability variable $x_\theta$, the group membership variable $x_{\text{group}}$, and their interaction $x_{\text{group},\theta}$. The sum score is commonly used as the continuous ability measure $x_\theta$ (Finch and French, 2007) and thus functions as the conditioning variable. The model formula specifies the probability of solving one specific item $j$ as $P(u_j = 1|x) = \frac{\exp(\eta_j)}{1+\exp(\eta_j)}$ through the linear predictor $\eta_j = \tau_{j0} + \tau_{j1}x_\theta + \tau_{j2}x_{\text{group}} + \tau_{j3}x_{\text{group},\theta}$. The test is originally intended to assess uniform and non-uniform DIF simultaneously and to specify the type of DIF post-hoc. If only the group membership is statistically significant, but the interaction term is not, uniform DIF is present. An iterative extension was suggested by Flier *et al.* (1984). Magis *et al.* (2011) generalize the logistic regression approach to multiple group comparisons. In Chapter 4, extensions to test for DIF in multiple items simultaneously will be presented.

The logistic regression approach is similar to the currently discussed broader mixed models framework for IRT models (Boeck *et al.*, 2011). In this approach, the linear predictor usually includes a random effect for the ability variable and fixed effects for the item difficulties. Potential DIF effects are grasped by means of person-by-item interactions. Therefore, similar to logistic regression, groups to be investigated have to be pre-specified and complex DIF-patterns such as focal groups that consist of person-covariate interactions might be missed.

Yet other methods to test for DIF are the multiple indicators, multiple cause (MIMIC) approach (Finch, 2005), the Simultaneous Item Bias (SIBTEST) procedure (Shealy and Stout, 1993a), confirmatory factor analysis (CFA) (see, e.g. Meade and Lautenschlager, 2004), structural equation modeling (SEM) (for an literature overview, see French and Maller, 2007), random item mixture (RIM) (Frederickx *et al.*, 2010), a robust outlier approach (Magis and De Boeck, 2011) or a penalty approach (Tutz and Schauberger, 2012). However, this thesis focuses on the itemwise Wald test that is presented in more detail – together with a simulation study addressing its performance under ideal conditions – in Chapter 5.

### 3.3.3 Methods to detect non-uniform DIF

In DIF research, non-uniform DIF might often be considered subordinate, since it is less often analyzed (Finch and French, 2007) and methods are primarily developed to detect uniform DIF (Li and Stout, 1996). Nevertheless, non-uniform DIF has also been found in empirical testing situations (e.g. a considerable proportion of non-uniformly biased items in a word analogies test in Mellenbergh, 1982).

Until the year 1996, Li and Stout (1996) observed less research that focused on non-uniform DIF as many standard procedures, such as the MH test, are specifically designed for the purpose to detect uniform DIF, even though empirical applications imply the presence of non-uniform DIF. The subordinate role of non-uniform DIF analysis persists, as Finch and French (2007) mention less insight in non-uniform DIF methods in the year 2007.

However, some methods to detect uniform DIF have been extended for the analysis of non-uniform DIF. Lord (1980) recommends to simultaneously test for uniform and non-uniform DIF by also including the estimated discrimination parameters in the Wald test. To simultaneously test for equal discrimination and difficulty parameters, the Rasch model does not suffice and the comparison is based on the estimated item parameters of the 2pl model. The approach of modeling non-uniform DIF by means of varying discrimination parameters is common in IRT-based DIF analysis, but – from a more general point of view – not the only possibility.

The logistic regression by Swaminathan and Rogers (1990) was already proposed to test for uniform and non-uniform DIF simultaneously. The probability of solving a specific item is modeled through the sum score, the group membership and their interaction. The post-hoc interpretation is suggested in the following way; If the interaction term is statistically significant, regardless of the other effects, non-uniform DIF is present. A generalization for multiple groups can be found in Magis *et al.* (2011).

To assess non-uniform DIF, Li and Stout (1996) suggested the crossing SIBTEST procedure that relies on a different test statistic. Finch and French (2007) systematically compared the Crossing SIBTEST procedure, the logistic regression, the likelihood ratio test (IRTLRDIF) suggested by Thissen *et al.* (1988), and the confirmatory factor analysis (CFA, see, e.g., Meade and Lautenschlager 2004) and found that the Crossing SIBTEST method and the logistic regression clearly outperformed the IRTLRDIF and the CFA method.

However, this thesis focuses on those methods that have previously been shown to perform well (e.g. by Finch and French, 2007). Furthermore, Rasch trees will also be extended to allow for non-uniform DIF detection in Chapter 4. Therefore, Rasch trees are introduced in more detail in the next section.

## 3.4 Rasch trees for uniform DIF detection

Rasch trees for uniform DIF detection have been previously proposed by Strobl *et al.* (2013). Rasch trees result when the model-based recursive partitioning algorithm is applied to the Rasch model as will be pointed out in detail below. With this approach it is possible to detect groups of subjects exhibiting DIF, that are not pre-specified nor latent classes, but result from combinations of observed covariates. Note that, all previously mentioned DIF detection methods do not share this characteristic, because the majority of DIF detection methods assumes that the groups which are tested for DIF are pre-defined.

Rasch trees are intended for detecting groups of subjects with DIF and are based on the technique of model-based recursive partitioning, that employs statistical tests for structural change adopted from econometrics. Model-based recursive partitioning is a semi-parametric approach. The aim is to detect differences in the parameters of a statistical model between groups of subjects defined by (combinations of) covariates. Such parameters could be, e.g., intercept and slope parameters in a linear regression model (as was illustrated in Chapter 2) or, as in our case, the item difficulty parameters of a Rasch model, that may vary between groups of subjects, and thus directly correspond to uniform DIF. If the procedure determines varying item parameters, it means that the null hypothesis of one joint Rasch model for the entire sample (i.e. measurement invariance) must be rejected. In this sense, the proposed method is a global test for DIF as well as an overall model test for the Rasch model. In addition to this, the graphical visualization allows to identify which groups are affected by DIF as will be illustrated in an instructive example in Chapter 4.

Technically, the following consecutive steps are used to infer the structure of a Rasch tree from the data:

1. Estimate the item parameters jointly for all subjects in the current sample, starting with the full sample.
2. Assess the stability of the item parameters with respect to each available covariate.
3. If there is significant instability, split the sample along the covariate with the strongest instability and in the cutpoint leading to the highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the resulting subsamples until there are no more significant instabilities (or the subsample becomes too small).

These four steps are now explained in more detail.

*1. Estimating the item parameters*

The common CML approach is used for estimating the item parameters. Let $\theta_i$, $i = 1, \ldots, n$, denote the person parameters, $\beta_j$, $j = 1, \ldots, k$, denote the item parameters and $u_{ij}$ denote the response of subject $i$ to item $j$. Since under the Rasch model

$$P(U_{ij} = u_{ij}|\theta_i,\beta_j) = \frac{e^{u_{ij}\cdot(\theta_i-\beta_j)}}{1 + e^{\theta_i-\beta_j}}$$

the person raw-scores $r_i = \sum_{j=1}^{k} u_{ij}$ form sufficient statistics for the person parameters, the item parameters can be estimated by means of iterative procedures from the conditional likelihood

$$L_c(\beta|r_1, \ldots, r_n) = \prod_{i=1}^{n} L_c(\beta|r_i) = \prod_{i=1}^{n} \frac{e^{-\sum_{j=1}^{k} u_{ij} \cdot \beta_j}}{\gamma_{r_i}(\boldsymbol{\beta})}, \tag{10}$$

where $\gamma_{r_i}$ is the symmetric function of order $r_i$ (cf., e.g., Molenaar, 1995a). To fix the origin of the scale, some constraint has to be applied, e.g., setting the first item parameter or the sum of all item parameters to zero as discussed earlier, leaving $k - 1$ free parameters.

*2. Testing for parameter instability*

In order to test whether the item parameters vary between groups of subjects defined by covariates, we use the approach of structural change tests from econometrics. These tests are usually employed for detecting, e.g., a drop in stock returns over time, whereas here we employ the same methodology for detecting parameter changes over person covariates. The rationale of the employed structural change tests is the following: The item parameters are first estimated jointly for the entire sample. Then the individual deviations from this joint model are ordered with respect to a covariate, such as age (as was illustrated in Figure 4 in Chapter 2). If there is systematic DIF with respect to groups formed by the covariate, the ordering will exhibit a systematic change in the individual deviations. If, on the other hand, no DIF is present, the values will merely fluctuate randomly.

For statistically testing structural change in the model parameters, we suggest the usage of generalized M-fluctuation tests (Zeileis and Hornik, 2007) that form the basis of the model-based recursive partitioning framework of Zeileis *et al.* (2008). The idea of this class of tests is to compute the subject-wise score contributions and derive test statistics with known distributions from them.

The individual score function $\psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})$, for $i = 1, \ldots, n$ observations, i.e., the derivative of the individual contributions to the log-likelihood $\Psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})$ with respect to the parameter vector, is a general measure of deviation for likelihood-based models. For the Rasch model these individual contributions can easily be computed from the conditional likelihood as outlined below.

For the construction of the test statistic, the individual contributions to the score function are cumulated according to the order induced by a variable such as age, or any other covariate. The systematic change from positive to negative in the individual contributions to the score function is then captured as a clearly noticeable peak in the cumulative sum process (see again Figure 4 in Chapter 2).

The cumulative sum process is defined as

$$W_l(t) \quad = \quad \widehat{\boldsymbol{V}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \psi(\boldsymbol{u}_{(i|l)}, \hat{\boldsymbol{\beta}}) \qquad (0 \leq t \leq 1), \tag{11}$$

where the index $(i|l)$ denotes the $i$-th ordered observation with respect to the $l$-th covariate, $\lfloor \cdot \rfloor$ denotes the integer part, $\widehat{\boldsymbol{V}} = \sum_{i=1}^{n} \psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}) \psi(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}})^\top$ is the outer-product-of-gradients estimate of the covariance matrix, and $t$ is a fraction of the sample size. Under the null hypothesis of parameter stability, the cumulative sum process $W_l(\cdot)$ can be shown to converge to an $(k-1)$-dimensional Brownian bridge (Zeileis and Hornik, 2007), which can be used as the basis for statistical inference.

The cumulative aggregation runs over the order induced by the $l$-th covariate: The $i = 1, \dots, n$ individual deviations are ordered with respect to the covariate and aggregated up to the $\lfloor n \cdot t \rfloor$-th element in each step. When $W_l(t)$ is considered as a function of the fraction $t$ of the sample size, under the null hypothesis of parameter stability the cumulative sum process follows the path of a random process with constant zero mean (whereas under the alternative hypothesis of parameter instability the path deviates from this random fluctuation).

The advantage of this approach is that the model does not have to be reestimated for all splits in all covariates, because the individual deviations remain the same and only their ordering (and the corresponding path of $W_l(t)$) needs to be adjusted for evaluating the different covariates.

To capture systematic deviations in $W_l(\cdot)$, different test statistics can be used depending on whether the $l$-th covariate is a numeric or a categorical variable. If it is numeric, Zeileis *et al.* (2008) point out that a natural test statistic is

$$S_l \quad = \quad \max_{i=\underline{i},\dots,\overline{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n}\right)^{-1} \left\| W_l\left(\frac{i}{n}\right) \right\|_2^2 . \tag{12}$$

This can be interpreted as the maximum LM statistic (also known as score statistic) for a single shift alternative over all conceivable cutpoints in $[\underline{i}, \overline{\imath}]$. The limiting distribution is the supremum of a tied-down Bessel process, from which p-values can be computed (for details see Zeileis *et al.*, 2008; Merkle and Zeileis, 2012).

If, on the other hand, the $l$-th covariate is categorical (with values $z_{il}$ taking categories $q = 1, \dots, Q$), it is more natural to use the following test statistic

$$S_l \quad = \quad \sum_{q=1}^{Q} n \left( \sum_{i=1}^{n} I(z_{il} = q) \right)^{-1} \left\| \Delta_q W_l\left(\frac{i}{n}\right) \right\|_2^2 , \tag{13}$$

where $\Delta_q$ is the increment within the $q$-th category. This test statistic is invariant to reordering of the $Q$ categories and the subjects within each category. The test statistic captures the instability over the $Q$ subsamples. Its limiting distribution is $\chi^2$ with $(Q-1) \cdot (k-1)$ degrees of freedom, from which p-values can be computed. This test is employed for both nominal and ordinal categorical variables. A potential ordering of the categories is accounted for in the next step, when the cutpoint is selected.

For the Rasch model, the objective function used for parameter estimation is the conditional log-likelihood. The individual contributions to the conditional log-likelihood can be easily computed as $\log L_c(\boldsymbol{\beta}|r_i)$ (cf. Equation 10), yielding

$$\Psi(\boldsymbol{u}_i, \boldsymbol{\beta}) = -\sum_{j=1}^{k} u_{ij} \cdot \beta_j - \log\left(\gamma_{r_i}(\boldsymbol{\beta})\right). \tag{14}$$

For the computation of the structural change tests, the individual contributions to the score function are derived from Equation 14. The contribution of the $i$-th subject for the $j$-th item parameter is:

$$\psi(\boldsymbol{u}_i, \boldsymbol{\beta})_j = \frac{\partial \Psi(\boldsymbol{u}_i, \boldsymbol{\beta})}{\partial \beta_j} = -u_{ij} - \frac{1}{\gamma_{r_i}(\boldsymbol{\beta})} \cdot \frac{\partial \gamma_{r_i}(\boldsymbol{\beta})}{\partial \beta_j}. \qquad (15)$$

The derivatives of the symmetric functions $\gamma_{r_i}(\boldsymbol{\beta})$ are again symmetric functions with certain terms omitted (cf., e.g., Glas and Verhelst, 1995). In our implementation of the Rasch trees, the sum algorithm of Liou (1994) is used (by default) for computing these derivatives.

When the individual contributions to the score function of the Rasch model from Equation 15 are ordered with respect to covariate $l$ and inserted in Equation 11, parameter instabilities in the item parameters can be statistically tested using the model-based recursive partitioning approach outlined above.

The results of this procedure are also easy to interpret: The parameter instability test statistics $S_l$ with associated (Bonferroni adjusted) p-values are provided for each candidate variable. The p-values are derived from the respective limiting distributions. Splitting continues until all p-values exceeded the significance level (commonly 5%), indicating that there is no more significant parameter instability, or until the number of observations in a subsample falls below a given threshold.

*3. Selecting the cutpoints*

After a covariate has been selected for splitting, the optimal cutpoint is determined by maximizing the partitioned log-likelihood (i.e., the sum of the log-likelihoods for two separate models: one for the observations to the left and up to the cutpoint, and one for the observations to the right of the cutpoint) over all candidate cutpoints within the range of this variable.

For a split within a binary variable, the selection of the cutpoint is trivial – e.g. gender only allows for a single split between the subgroups of females and males. In splits within numeric variables or categorical variables that are no longer binary, all possible cutpoints are considered and the optimal cutpoint is chosen according to the highest value of the partitioned log-likelihood.

Formally, for a numeric splitting variable $l$ with values $z_{il}$ we can define the subsamples $L(\xi) = \{i \,|\, z_{il} \leq \xi\}$ and $R(\xi) = \{i \,|\, z_{il} > \xi\}$ on the left and right, respectively, of some cutpoint $\xi$. For both subsamples, the parameters $\hat{\boldsymbol{\beta}}^{(L)}$ and $\hat{\boldsymbol{\beta}}^{(R)}$ can be estimated separately as described above. To determine the optimal cutpoint $\xi$, the partitioned log-likelihood

$$\sum_{i \in L(\xi)} \Psi\left(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}^{(L)}\right) + \sum_{i \in R(\xi)} \Psi\left(\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}^{(R)}\right)$$

is maximized over all candidate cutpoints $\xi$ (typically requiring a certain minimal subsample size).

While this approach can be applied to numeric and ordered covariates, for unordered categorical covariates the $Q$ categories can be split into any two groups. From all these candidate binary partitions, again the one that maximizes the partitioned log-likelihood is chosen.

Note that choosing the optimal cutpoint by maximizing the partitioned (log-)likelihood corresponds directly to using the maximum LR statistic of the joint vs. the partitioned model. Thus, for selecting the optimal cutpoint the computationally more expensive LR test is implicitly used in the Rasch tree algorithm. However, it is not employed in the first step for *testing whether* there is significant DIF in a covariate, but only for the second step of *estimating where* the strongest DIF occurs by obtaining the maximum likelihood estimator for the cutpoint. Unlike the tests in the previous sections, this computationally costly LR test is not applied to all potential splitting variables but only to those selected for splitting in the first place.

From a statistical point of view, this two-step approach – where the variable selection is made independently from the cutpoint selection – has two important advantages: Not only does it considerably reduce the computational burden (Strobl *et al.*, 2013), but at the same time it also prevents an artifact termed variable selection bias (cf., e.g., Dobra and Gehrke, 2001; Shih, 2004; Hothorn *et al.*, 2006b; Strobl *et al.*, 2007), that was inherent in earlier recursive partitioning algorithms.

Variable selection bias occurs when first the best cutpoint is determined in each variable and then the best splitting variable is selected by means of evaluating some splitting criterion or test statistic, that was computed exactly for the cutpoint producing the highest value of this criterion or statistic. In this case, variables offering more cutpoints (such as numeric variables or variables with many categories) have a higher chance of being selected only due to multiple testing, which does not reflect the actual quality of the splitting variable. Therefore, with respect to selecting the variable with the strongest parameter instability in the Rasch model it would be statistically incorrect to select the best splitting variable by means of the standard LM or LR test (based on the standard $\chi^2$ distribution) when the test statistic is computed in the best cutpoint offered by that variable. The reason is that – due to the optimal selection of the cutpoint – the asymptotic distribution of this optimally selected statistic is no longer $\chi^2$ (Andrews, 1993). Therefore, the correct distribution has to be derived for any optimally selected statistic (cf., e.g., Miller and Siegmund, 1982; Koziol, 1991; Hothorn and Lausen, 2003; Boulesteix, 2006) – as in our case for the optimally selected LM statistic from Equation 12, that is employed in the variable selection step of the Rasch tree method. This approach guarantees that the selection of the best splitting variable is not affected by the number of cutpoints offered by each candidate variable, but can make the test decision a little conservative in small to moderate samples.

*4. Stopping criteria*

For creating a Rasch model, the four basic steps outlined above – (1) estimating the item parameters of a joint model, (2) testing for parameter instability, (3) selecting the splitting variable and cutpoint and splitting the sample accordingly – are repeated recursively until a stopping criterion (4) is reached.

Two kinds of stopping criteria are currently implemented: Splitting continues only as long as significant parameter instability is detected. If there is no (more) significant instability with respect to any of the covariates, the splitting stops. Thus, the significance level – usually set to 5% – serves as the most important stopping criterion.

In addition to that, as a second stopping criterion a minimum sample size per node can be specified. This minimal node-size should be chosen such as to provide a sufficient basis for parameter estimation in each subsample, and should thus be increased when the number of item parameters to be estimated is large.

Finally, one should keep in mind that when a large number of covariates is available in a data set, and all those covariates are to be tested for DIF, multiple testing becomes an issue – as with any statistical test for DIF. To account for the fact that multiple testing might lead to an increased false-positive rate when the number of available covariates is large, a Bonferroni adjustment for the p-value splitting criterion is applied internally (so that all p-values reported for Rasch trees throughout this thesis have already been Bonferroni-corrected).

Another issue related to stopping criteria in recursive partitioning algorithms is their potential for overfitting: In classical algorithms (such as CART; Breiman *et al.*, 1984) a pruning step (i.e. cutting back branches at the bottom of the tree that do not add to the prediction accuracy in cross-validation) is necessary to make sure that any splits detected for the learning data do not only reflect random variation but also generalize to other samples from the same data generating process. As opposed to these classical algorithms, the model-based recursive partitioning approach employed here is already based on statistical inference tests (rather than merely descriptive statistics) and uses their p-values (together with several precautions against multiple testing) for stopping before overfitting occurs (cf. also Hothorn *et al.*, 2006b). Therefore, pruning is not necessary in this approach.

Moreover, it is important to note that our model-based recursive partitioning algorithm is not affected by an inflation of chance due to its recursive nature. Indeed, several statistical tests are successively conducted in a Rasch tree – but each test is conducted only if the previous test yielded a significant result. In this sense, the recursive approach forms a closed testing procedure, which does not lead to an inflation of chance as is well known from the literature on multiple comparisons (Marcus *et al.*, 1976; Hochberg and Tamhane, 1987). For the Rasch trees this means that the postulated significance level holds for the entire tree, not only for each individual split. This ensures that DIF is not erroneously detected as an artifact of the recursive nature of the algorithm.

Rasch trees have been previously proposed for the detection of uniform DIF (Strobl *et al.*, 2013). In the next chapter of this thesis, the question of whether Rasch trees are also suited to detect non-uniform DIF is addressed. Rasch trees will be illustrated by means of two instructive examples including uniform and non-uniform DIF in the next chapter.

# 4 Detecting non-uniform DIF with Rasch trees

***Summary:*** *A variety of statistical methods have been suggested for the analysis of differential item functioning (DIF) in the Rasch model. Nevertheless, the detection of non-uniform DIF has attracted little attention in the development and evaluation of statistical methods for DIF detection and in empirical studies on DIF. In this chapter, we address the question of whether the newly suggested Rasch trees are suited for the detection of non-uniform DIF in the Rasch model. In an extensive simulation study, we evaluate the appropriateness of Rasch trees as a global test for the detection of uniform and non-uniform DIF. To allow for a direct comparison, we suggest two extensions of the logistic regression approach that has previously been suggested for item-wise detection of uniform and non-uniform DIF. The results show that Rasch trees yield well-controlled type one error rates and are well suited for all different types of DIF when the sample size is large. The extensions of the logistic regression were superior when the DIF groups are known and the DIF-items were simulated with varying discrimination parameters. Still, Rasch trees are able to automatically detect observable groups that display DIF and maintain their straightforward interpretation in non-uniform DIF situations. Thus, they allow for a flexible analysis of DIF even if no reliable prior knowledge of the composition of the DIF groups is available.*

***Keywords:*** *Rasch model, Rasch trees, differential item functioning (DIF), non-uniform DIF, crossing DIF*

## 4.1 Introduction

In item response theory (IRT) research, invariant item parameters are necessary to assess ability differences between groups in an objective, fair way. If the invariance assumption is violated, different ICCs occur in subgroups and DIF is present as discussed in Chapter 3. Most statistical methods developed for the assessment of DIF focus on *uniform* DIF, where one group has a higher probability of solving an item (given the latent trait) over the entire latent continuum and the group differences in the logit, i.e. the logarithmic odds of correctly answering the item, remain constant (Mellenbergh, 1982; Swaminathan and Rogers, 1990). In contrast to this, *non-uniform* DIF is present when the group differences in the logit are no longer constant but differ over the range of the latent continuum. Thus, from a statistical perspective, non-uniform DIF corresponds to an interaction effect between the group membership and the ability variable on the probability of solving the item (Mellenbergh, 1982).

Empirical research in educational and psychological testing confirmed the existence of non-uniform DIF-items (cf. e.g., Linn *et al.*, 1981; Mellenbergh, 1982; Li and Stout, 1996; Hambleton and Rogers, 1989; Ellis, 1989; Maller, 2001; Pae, 2004) and "*[a]lthough nonuniform DIF is said to occur at a lower rate than uniform DIF in practice, accurate detection is no less important*" (Finch and French, 2007, p. 566). Nevertheless, the development and evaluation of statistical methods to detect DIF mainly focus on uniform DIF and DIF analysis is more frequently carried out using methods designed for the detection of uniform DIF (Li and Stout, 1996; Pae, 2004; Finch and French, 2007).

In this chapter, we address the question of whether the recently suggested Rasch trees (Strobl *et al.*, 2013) are able to detect non-uniform DIF. Rasch trees were initially designed to automatically detect groups that display uniform DIF in the Rasch model. These groups are defined by combinations of available covariates. The resulting Rasch trees can be graphically displayed and thus provide a straightforward interpretation. The method employs a global test for model violations where all items are simultaneously investigated (see again Section 3.4), like in the commonly used likelihood ratio test (Anderson, 1973; Gustafsson, 1980).

In the following section, previous studies on the detection of non-uniform DIF are reviewed. After that, the rationale for detecting non-uniform DIF using Rasch trees is explained from a theoretical perspective and by means of an illustrative example. Furthermore, existing methods to detect non-uniform DIF are extended to allow for a meaningful comparison with the Rasch trees. In Section 4.3 the simulation design is presented and the results are discussed in Section 4.4. A concluding summary of the simulation results as well as practical recommendations are given in Section 4.5.

## 4.2 Methods

In the following, we briefly review previous simulation studies that compare methods to detect non-uniform DIF (for a more detailed literature overview including studies on uniform DIF detection see Finch and French, 2007).

In the simulation study of Narayanan and Swaminathan (1996), three different methods were compared regarding their appropriateness for non-uniform DIF detection in settings of different sample sizes, ability distributions, proportions of DIF items, effect sizes, and types of non-uniform DIF. The compared methods were the logistic regression proposed for DIF detection by Swaminathan and Rogers (1990), an extension of the original Simultaneous Item Bias (SIBTEST) procedure (Shealy and Stout, 1993a) named Crossing SIBTEST, that was developed by Li and Stout (1996) and allows for non-uniform DIF detection, and the Mantel-Haenszel (MH) procedure (Holland and Thayer, 1988). The logistic regression and the Crossing SIBTEST procedure reached comparable hit rates for item-wise DIF detection. However, both methods yielded inflated false alarm rates when differences in the ability distributions of reference and focal group were present. Unlike the MH test, which displayed low hit rates in the majority of the simulated settings, the logistic regression and the Crossing SIBTEST procedure were recommended for non-uniform DIF detection together with an adjustment of the alpha level when unequal ability distributions are present.

In their comparison of the Crossing SIBTEST procedure with the logistic regression, Li and Stout (1996) found the logistic regression to reach a higher statistical power of detecting non-uniform DIF in settings where the data generating process relied on the two-parameter logistic (2pl) model. The false alarm rate of the logistic regression was acceptable in this setting as opposed to settings where e.g. the three-parameter logistic (3pl) model was used as the data generating process.

Finch and French (2007) systematically compared four methods for detecting non-uniform DIF in a simulation study: the Crossing SIBTEST procedure, the logistic regression, a likelihood ratio test (IRTLRDIF) suggested by Thissen *et al.* (1988), and confirmatory factor analysis (CFA, see, e.g., Meade and Lautenschlager 2004). In their simulation study, item-wise non-uniform DIF tests were conducted for dichotomous items in conditions of varying sample sizes, ability differences, and proportions of DIF-items. All procedures yielded type one error rates close to the expected level of 0.05. Regarding the power, the Crossing SIBTEST method and the logistic regression clearly outperformed the IRTLRDIF and the CFA method.

Even though the findings agree that the logistic regression and the Crossing SIBTEST method outperformed other methods in the detection of non-uniform DIF, the results are not clear with respect to which of the two methods is superior. While the Crossing SIBTEST procedure showed the highest power in the study of Finch and French (2007), Narayanan and Swaminathan (1996) found the methods to be comparable and Li and Stout (1996) found the logistic regression to be superior. In this study, we will use the logistic regression for comparing it with the Rasch trees. There are several arguments for this choice.

Firstly, the Crossing SIBTEST method requires a so called set of valid items to construct a matching subtest (for details see Shealy and Stout, 1993b). Typically, this set of valid items is a subset of the items. Technically, this subset is defined as a strict subset of the items and cannot consist of all items that should simultaneously be tested for DIF (Shealy and Stout, 1993a, p. 652). Therefore, this test procedure requires more information than one actually has in many practical applications and more than given in our simulation design. This kind of additional information is neither required when the logistic regression nor when the Rasch trees are used for DIF detection.

Secondly, the logistic regression allows to simultaneously test for uniform and non-uniform DIF (Finch and French, 2007) and can, thus, easily be extended to allow for a direct comparison with the Rasch trees that also allow to test for uniform and non-uniform DIF simultaneously without specifying the type of DIF prior to data analysis.

Thirdly, similar to the Rasch trees, logistic regression allows to include more than one group variable to test for DIF. In our simulation design, this property is highlighted by including a noise variable to see whether DIF related to that variable is falsely detected. Moreover, logistic regression can also be used for multiple group comparisons (Magis *et al.*, 2011) that are, however, not considered in this chapter.

As a result of these arguments, Rasch trees will be compared with logistic regression in this chapter to evaluate their performance in detecting non-uniform DIF. One drawback in this decision is that logistic regression is designed for the assessment of DIF in one item only and is thus not directly suited to detect non-uniform DIF in more than one item simultaneously (Li and Stout, 1996). Therefore, the method first has to be extended to allow for a direct comparison with the Rasch trees. Moreover, these extensions can also be of interest for a variety of applications as a global model test particular sensitive to non-uniform DIF. We present two extensions that allow a global assessment of DIF: a multiple comparison procedure based on a correction of the p-values of several item-wise logistic regressions and a model comparison procedure for the global null hypothesis of no DIF in any item.

In the following, the logistic regression and the Rasch tree approach are reviewed and the assessment of non-uniform DIF will be illustrated. Then, the two extensions of logistic regression are introduced in order to allow a global assessment of (uniform and non-uniform) DIF.

### 4.2.1 Logistic regression

Swaminathan and Rogers (1990) proposed the usage of the logistic regression to detect uniform or non-uniform DIF in one item.
The model formula specifies the probability of solving one specific item $j$

$$P(u_j = 1|x) = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)}$$

through the linear predictor $\eta_j$. The authors suggest the following form:

$$\eta_j = \tau_{j0} + \tau_{j1} x_\theta + \tau_{j2} x_{\text{group}} + \tau_{j3} x_{\text{group},\theta}.$$

The linear predictor thus consists of an intercept, the ability variable $x_\theta$, the group membership variable $x_{\text{group}}$, and their interaction $x_{\text{group},\theta}$ (an abbreviation for $x_{\text{group},} \cdot x_\theta$). The sum score is commonly used as the continuous ability measure $x_\theta$ (Finch and French, 2007).
The simultaneous null hypothesis for uniform and non-uniform DIF can be written as a linear hypothesis of the form $H_0 : C\tau = 0$, more precisely,

$$H_0 : C\tau = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_{j0} \\ \tau_{j1} \\ \tau_{j2} \\ \tau_{j3} \end{pmatrix} = \begin{pmatrix} \tau_{j2} \\ \tau_{j3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

or using the notation of Magis *et al.* (2011) where $\mathbf{0}_{n\times m}$ represents a $n$-by-$m$ matrix containing zeros and $\mathbf{I}_n$ the $n$-by-$n$ identity matrix consisting of zeros except for the diagonal one entries

$$H_0 : C\tau = \begin{pmatrix} \mathbf{0}_{2\times 2} & \mathbf{I}_2 \end{pmatrix} \tau = \begin{pmatrix} \tau_{j2} \\ \tau_{j3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The alternative hypothesis $H_1$ corresponds to $C\tau \neq 0$ and indicates DIF. Note that the effect $\tau_{j1}$ for the sum score variable is not included in the hypothesis since it is included in the model only to condition on ability. The authors suggested to use a Wald test for testing the hypothesis for $\tau_{j2}$ and $\tau_{j3}$. Whether uniform DIF or non-uniform DIF is present cannot be inferred from the simultaneous significance test itself. Post-hoc tests represent one alternative, which is used in this chapter, to decide which type of DIF is present. A significant effect of $\tau_{j2}$ alone corresponds to uniform DIF since $\tau_{j2}$ measures the effect of the group membership on the probability of solving the item conditioned on ability. A significant effect $\tau_{j3}$ (independent of the effect $\tau_{j2}$) captures the interaction of the ability and the group membership and thus corresponds to non-uniform DIF. Other alternatives, e.g. to explore the type of DIF graphically or by comparing $R^2$ values (Gelin and Zumbo, 2003; Zumbo, 1999), are not discussed here.

### 4.2.2 Rasch trees

Rasch trees were initially designed for the detection of uniform DIF in the Rasch model. By following the recursive partitioning algorithm (see again Section 3.4 and for technical details cf. Strobl *et al.* 2013), Rasch trees are able to detect groups of subjects – defined by (combinations of) covariates – that display DIF. Rasch trees determine the groups displaying (non-uniform) DIF by searching over all available covariates whether instable item difficulty parameters are found (for a general introduction of recursive partitioning, see Strobl *et al.*, 2009).

As an exemplary illustration for uniform DIF detection, we show a Rasch tree based on simulated data. In the example, a test consisting of 10 items includes one uniform DIF-item (item 3) that differs in its difficulty parameter between the reference group and the equally able focal group by $\Delta_{\text{DIF}} = \beta_{3,\text{foc}} - \beta_{3,\text{ref}} = 0.6$ and has been manually highlighted in the following figures for clarity. In our example, the focal group consists of female test-takers up to the age of 60. All remaining items follow the Rasch model with the parameters described in Section 4.3.

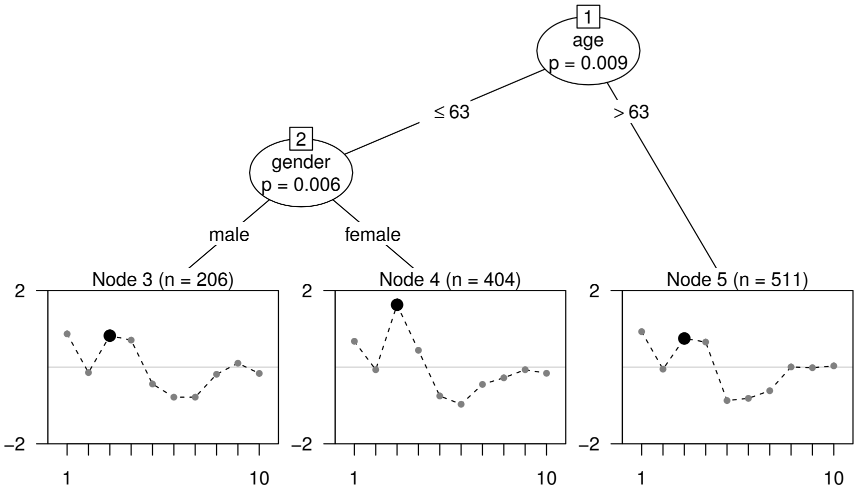

**Figure 9** – *Rasch tree for uniform DIF based on simulated data.*

The fact that the tree splits at all means that DIF is present in the test and one joint Rasch model for the entire sample is rejected. As can be seen from Figure 9, the Rasch tree can identify the correct combination of variables to discriminate between groups of subjects with different item parameters in the case of uniform DIF. Furthermore, a noise variable was handed over to the method which was correctly not selected for splitting. Note that the cutpoint in the continuous variable age was not specified but detected by the algorithm in a data driven way.

For the logistic regression and from the definition of non-uniform DIF, we have seen that non-uniform DIF can be modeled as an interaction between a grouping variable and an ability variable. Since recursive partitioning is particularly suited for detecting interactions (Strobl *et al.*, 2009), the Rasch tree approach can be extended in a straightforward way for detecting non-uniform DIF by handing over an ability variable in addition to the covariates. Hence, Rasch trees are able to detect both types of DIF, uniform and non-uniform DIF. Here, we use again the sum score as a continuous measure of ability, as is often done in practice (Finch and French, 2007).

As an exemplary illustration for non-uniform DIF, we again show a Rasch tree based on simulated data similar to the previous example, but now one non-uniform DIF-item (item 3) differs in its discrimination parameter between the reference group ($\alpha_{\text{ref}} = 1$) and the focal group ($\alpha_{\text{foc}} = 0.4$). Here, the focal group consists of test-takers over the age of 60.

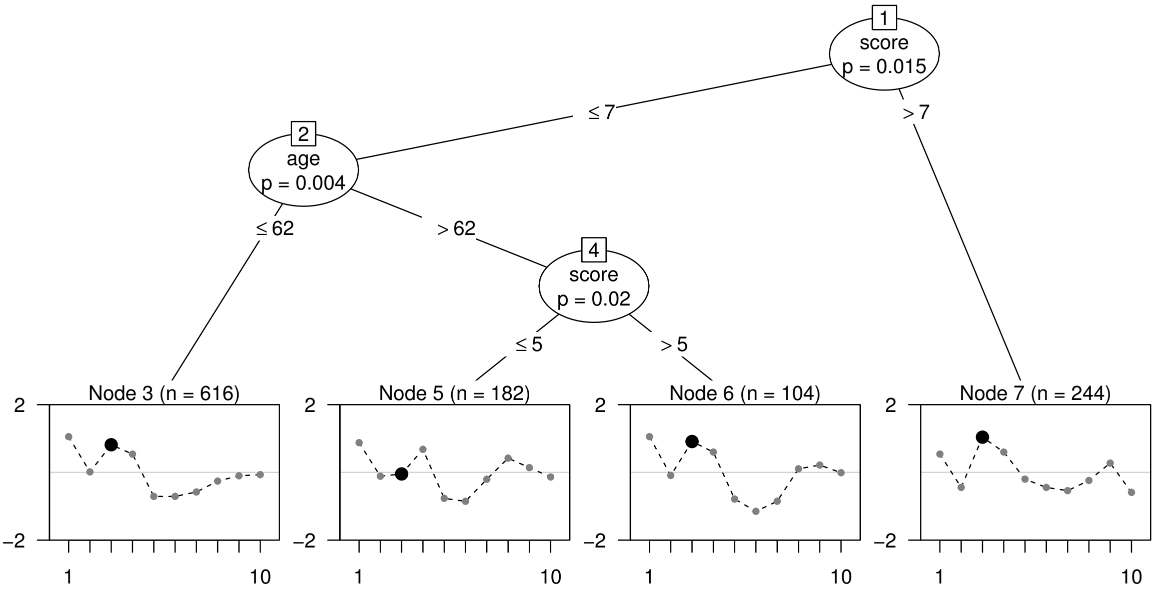

**Figure 10** – *Rasch tree for non-uniform based on simulated data.*

The sum score variable, the age variable and several other potential grouping variables are handed over to the algorithm. The resulting Rasch tree in Figure 10 can, again, be interpreted straightforwardly. The fact that there is more than one node, means that DIF is present in the test.

As can be seen from the graphical representation of the end nodes, the third estimated item difficulty varies between the groups. The fact that both the variable age and the ability measure were used for determining the subsamples displaying different item parameters means that it is an interaction of the two that determines the item parameters – which is exactly the definition of non-uniform DIF. The two splits in the ability measure are a means of the recursive partitioning algorithm to approximate the underlying continuous change in ability by means of several piecewise constant parts (cf. Strobl *et al.*, 2009, for details).

While in generalized linear models the information from different predictor variables is combined linearly, here the range of possible combinations includes all rectangular partitions that can be derived by means of recursive splitting – including multiple splits in the same variable as pointed out by Strobl (2013). In particular, this includes nonlinear and even non-monotone association rules (such as a u-shaped association, where for example very young and very old persons are more likely to show certain measurement properties different than those of middle ages, or even more complex shapes). The advantage of Rasch trees is that the functional form need not be known or specified in advance, but is automatically detected by the algorithm in a data driven way.

However, the rectangular splits can only give a rough approximation of any smooth function. So if, for example, the relationship was truly linear, a tree-based method would not give a very good approximation of this function (it could only approximate it with many rectangular steps), whereas e.g. a linear model would be able to describe the linear relationship perfectly well. On the other hand, if the true relationship was, say, quadratic, one would have to explicitly include the quadratic term in the model – and if this is not done, the linear model will fail to detect the association while a recursive partitioning algorithm can at least approximate it, as illustrated in Figure 11.

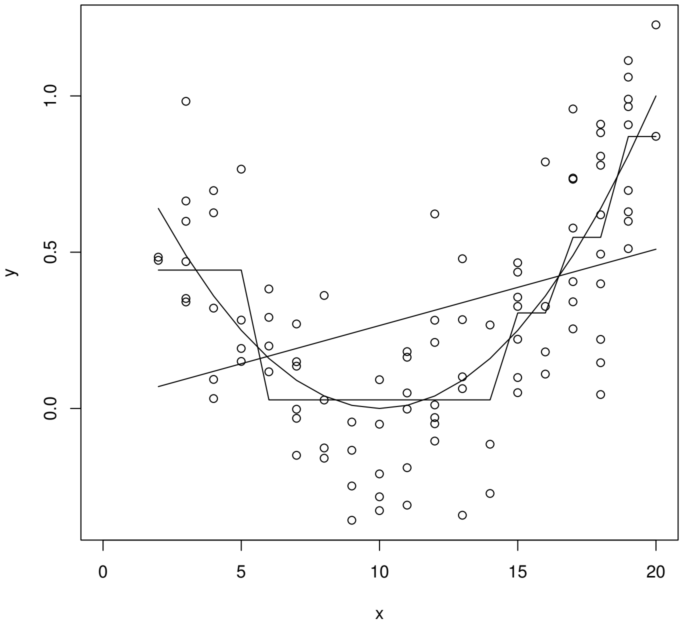

**Figure 11** – *Illustration of the approximation of a smooth quadratic function by a tree-based approach and by regression.*

What follows from the example and what is important to note here, is that – in case of non-uniform DIF – the resulting tree can display various patterns that approximate non-uniform DIF as illustrated in Figure 12.

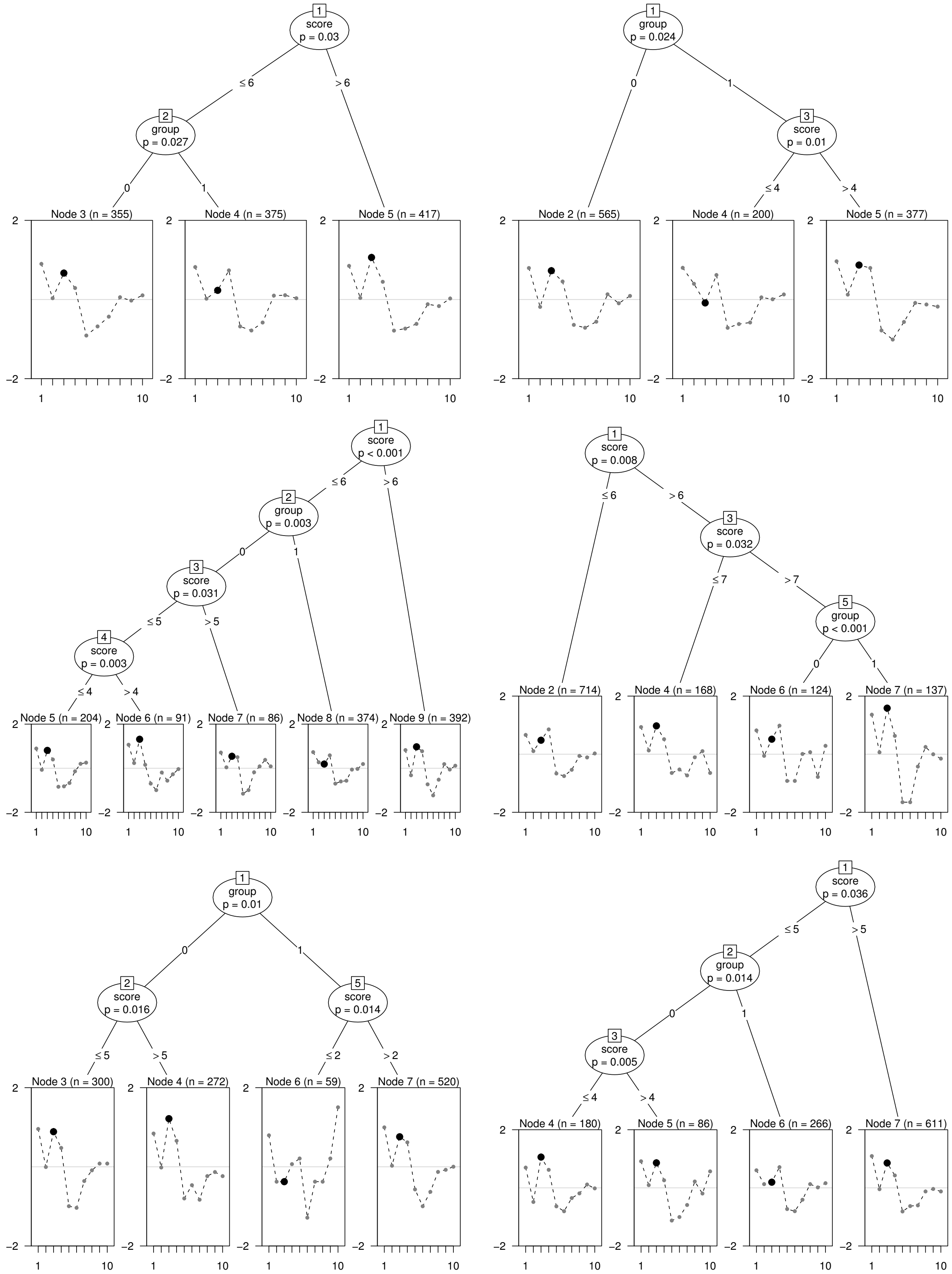

**Figure 12 –** *Illustration of various Rasch trees in the case of non-uniform DIF.*

Even though the pattern obtained in the resulting tree structure may appear complex, the interpretation is straightforward; if both types of variables – one or more group membership variables and the ability variable – are selected for splitting, the test displays non-uniform DIF.

In contrast to existing methods such as the logistic regression approach discussed in Section 4.2.1, Rasch trees detect groups that are not pre-defined, but empirically determined by the method itself. In particular, the interaction effects do not have to be specified in advance as would be the case for the logistic regression. Even though Rasch trees exhaustively search for subgroups displaying different item parameters over all available covariates and possible cutpoints, the type one error rate is well-controlled and the computational effort is reasonable (for details, see again Strobl *et al.*, 2013).

Since Rasch trees rely on tests for the instability of any item difficulty parameters, they may on one hand lack statistical power compared to methods that capture non-uniform DIF in one single parameter (as is done e.g. in the logistic regression). On the other hand, Rasch trees are far more flexible so that a variety of different non-standard patterns of non-uniform DIF can be detected.

The two most crucial differences between Rasch trees and classical approaches like logistic regression are now addressed. First, while other procedures that test for uniform and non-uniform DIF simultaneously, such as the logistic regression, require post-hoc tests that test the individual model coefficients separately to identify the type of DIF after the decision of whether any type of DIF is present (see again Section 4.2.1), Rasch trees provide both decisions simultaneously: If only group membership variables are selected for splitting, it means that only uniform DIF is detected. If both types of variables, the sum score and group membership variables are selected for splitting, it means that non-uniform DIF is detected.

Second, the groups are found automatically from (combinations of) all available variables and not only those main effects and interactions that are explicitly included e.g. in the logistic regression model formula (cf. Section 4.2.1) are considered.

### 4.2.3 Extensions of the logistic regression

The aim in this chapter is to test whether any uniform or non-uniform DIF is present simultaneously for all $k$ items. Since the logistic regression in the form suggested by Swaminathan and Rogers (1990) and presented in Section 4.2.1 is not able to test more than one item simultaneously (Li and Stout, 1996), the method is extended in this section. Since there is not one unique way of constructing a global DIF test from the logistic regression approach, here we describe two possible extensions and include both in our simulation study.

*Item-wise logistic regression (i-logit)*

The first extension of the logistic regression is based on item-wise logistic regressions and thus called *i-logit* method in order to avoid confusion. The linear predictor for each item $j$ is again specified as

$$\eta_j = \tau_{j0} + \tau_{j1} x_\theta + \tau_{j2} x_{\text{group}} + \tau_{j3} x_{\text{group},\theta} .$$

The item-wise logistic regressions are used to test for DIF in each item separately as described in Section 4.2.1 and can, then, be combined by using a correction of the p-values for the multiple testing problem.

For each item, one logistic regression is computed and the p-values $p_j$ of the resulting Wald tests (that are testing two coefficients simultaneously and, thus, follow a $\chi^2$ distribution with 2 degrees of freedom) are corrected for the multiple testing problem by using the Benjamini-Hochberg (BH) correction (Benjamini and Hochberg, 1995) that has previously been applied in DIF testing situations (Thissen *et al.*, 2002; Woods, 2009; Setodji *et al.*, 2011; Kim and Oshima, 2013; Raykov *et al.*, 2013). By focusing on the control of the false discovery rate, this procedure allows for higher hit rates than procedures designed to control the familywise error rate such as the Bonferroni correction (Benjamini and Hochberg, 1995; Williams *et al.*, 1999).

The resulting procedure tests whether DIF is present in at least one item. However, like for the item-wise version of Swaminathan and Rogers (1990), it is still to be decided which type of DIF – uniform or non-uniform – is present. Thus, post-hoc tests are needed to check whether the items display uniform or non-uniform DIF.

Every item that displayed statistically significant DIF is included in the post-hoc tests for the i-logit method. The significance tests of the regression coefficients are grouped by the corresponding covariate (the group variable $x_{\text{group}}$ and the interaction $x_{\text{group},\theta}$). Then, the p-values are again corrected separately for the group and the interaction effects. If only the coefficients of the group variable are significant (and all other post-hoc tests are not significant), uniform DIF is reported. If the interaction is significant (regardless of the main effect for the group membership variable), non-uniform DIF is reported.

*Vector logistic regression (v-logit)*

The i-logit method uses one logistic regression per item and adjusts the p-values for the multiple testing problem. However, the power of the adjusted tests will decrease due to the stricter significance level. Thus, in the case of several items displaying a small DIF effect, this procedure might fail to detect the existing DIF (Shealy and Stout, 1993a).

Therefore, we present another straightforward generalization to test for uniform or non-uniform DIF in one or more items simultaneously. Here we illustrate a global test for all items, but our approach could also be used for investigating a pre-defined subset of items simultaneously.

Analogously to the modeling approach of Swaminathan and Rogers (1990), we adopt the vector logistic regression (abbreviated *v-logit*) from the class of vector generalized linear models that are also termed multivariate generalized linear models (Yee and Wild, 1996; Fahrmeir and Tutz, 2001). For each of the item responses $u_j \in (u_1, u_2, \cdots, u_k)$ the probability of solving item $j$ is again described through a linear predictor:

$$P(u_j = 1|x) = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)}.$$

In the so called unrestricted model the linear predictor

$$\eta_j = \tau_{j0} + \tau_{j1} x_\theta + \tau_{j2} x_{\text{group}} + \tau_{j3} x_{\text{group},\theta}$$

again includes the intercept, the ability measure (here again the sum score), the group indicator and the interaction term. The response variables are assumed to be conditionally independent given the covariates (Yee, 2010b), similar to the local stochastic independence assumption of the Rasch model. A joint model for the response vector including all items is fitted. The estimation can be carried out using Maximum Likelihood (for details on the estimation see Yee and Wild, 1996; Yee and Hastie, 2003).

In order to construct a global DIF test statistic, we additionally estimate the restricted model where the coefficients that correspond to the potential DIF effects (all $\tau_{j2}$'s and $\tau_{j3}$'s) are set to zero

$$\eta_j = \tau_{j0} + \tau_{j1} x_\theta \,.$$

This model corresponds to the null hypothesis of no DIF using again the notation of Magis *et al.* (2011)

$$H_0 : \, C\tau = \begin{pmatrix} \mathbf{0}_{2\times2} & \mathbf{I}_2 & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{I}_2 & \cdots & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ & \vdots & & & & \vdots & \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \cdots & \mathbf{0}_{2\times2} & \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \tau_{10} \\ \tau_{11} \\ \tau_{12} \\ \tau_{13} \\ \vdots \\ \tau_{k0} \\ \tau_{k1} \\ \tau_{k2} \\ \tau_{k3} \end{pmatrix} = \begin{pmatrix} \tau_{12} \\ \tau_{13} \\ \vdots \\ \tau_{k2} \\ \tau_{k3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Both models are nested and, therefore, the likelihood ratio (LR) test can be used to compare the likelihoods of both models. The LR test employs the test statistic

$$-2\,(\text{log-likelihood}_{\text{restricted}} - \text{log-likelihood}_{\text{unrestricted}})$$

that – under the null hypothesis – follows a $\chi^2$-distribution with the degrees of freedom equal to the difference in the number of model parameters (e.g. if all $k$ items are tested for DIF, the degrees of freedom are equal to $2 \cdot k$). If the test is significant, the null hypothesis of no DIF is rejected. Again, this first result provides no information about whether uniform or non-uniform DIF is present and, thus, post-hoc tests must be carried out.

Since – in contrast to the i-logit method – no information about which items display DIF is yet available, all items need to be included in the classification of the type of DIF. Thus, for each variable of interest, a BH-corrected simultaneous Wald test is carried out at the 5% significance level to check whether each variable has an effect. The type of DIF is classified in the same way as for the i-logit method described above.

## 4.3 Simulation study

A simulation study is conducted in the free R system for statistical computing (R Development Core Team, 2011).

- i-logit method: We implemented the i-logit method using the R base system and the add-on R-package aod by Lesnoff and Lancelot (2012).
- v-logit method: We implemented the v-logit method using the add-on R-package VGAM by Yee (2010a,b) that provides the estimation of vector generalized linear models.
- Rasch trees: The implementation was based on the add-on R-package psychotree by Zeileis *et al.* (2011) that provides a freely available implementation of the Rasch trees.

In the simulation study, two questions are addressed. In the first part, the focus is on evaluating the ability of the Rasch trees, the i-logit, and the v-logit method to detect uniform and non-uniform DIF. Therefore, the methods are compared in several conditions including no DIF, uniform DIF and non-uniform DIF. Parts of the simulation design were inspired by the settings used by Narayanan and Swaminathan (1996), Li and Stout (1996), and Finch and French (2007). Each setting is replicated 1000 times to ensure reliable results. The results will show that Rasch trees have a lower power for small sample sizes when non-uniform DIF is simulated using different discrimination parameters.

In the second part of the simulation study, the focus is on illustrating that there are still situations where Rasch trees profit from their ability to automatically detect non-standard groups even for non-uniform DIF.

Furthermore, a noise variable is handed over to all methods to allow for a comparison of the logistic regression methods and the Rasch trees in a situation where distracting information is present. To test whether the logistic regression methods erroneously show a significant effect of the noise variable, the model formula for each item $j$ in both logit methods also includes the additional noise variable $x_{\text{noise}}$ (only the main effect is included to avoid a systematic disadvantage for the i-logit or the v-logit method in the simulation study). Thus, the linear predictor fitted in the simulation study is:

$$\eta_j = \tau_{j0} + \tau_{j1} x_\theta + \tau_{j2} x_{\text{group}} + \tau_{j3} x_{\text{group},\theta} + \tau_{j4} x_{\text{noise}} \,.$$

The Wald tests for the i-logit method now include also the coefficient $\tau_{j4}$ and, thus, follow a $\chi^2$ distribution with 3 degrees of freedom. The resulting p-values are corrected using the BH-correction and the post-hoc tests including the items displaying DIF are again corrected separately for the group, the interaction and the noise effects. If only the coefficients of the group variable are significant (and all other post-hoc tests are not significant), uniform DIF is reported. If the interaction is significant (regardless of the result for the group membership variable but with no significant result for the noise variable), non-uniform DIF is reported. Accordingly, the LR test statistic for the v-logit method also includes the effects for the noise variable and, hence, follows a $\chi^2$ distribution with $3 \cdot k$ degrees of freedom. Post-hoc tests are carried out and are interpreted in the same way as for the i-logit method.

### 4.3.1 Manipulated variables

The manipulated variables in the first part of our simulation study are the ability distribution, the sample size, the DIF type and the DIF proportion. These manipulated variables are fully crossed in the first part of the simulation under the alternative hypothesis where DIF is present (see Section 4.3.2).

- *Ability distribution*

  The abilities of the focal group are generated from a standard normal distribution. In relation to the focal group, the reference group is either equally able, $\theta_{\text{ref}} \sim N(0, 1)$, or more able, $\theta_{\text{ref}} \sim N(0.5, 1)$.

- *Sample size*

  The sample sizes from the reference and the focal group $(n_{\text{ref}}, n_{\text{foc}})$ are varied $(n_{\text{ref}}, n_{\text{foc}}) \in \{(500, 500), (1000, 500), (1000, 1000), (1500, 1000), \ldots, (2500, 2500)\}$. Thus, both equal and different group sizes are considered.

- *DIF type*

  The *DIF type* is varied in four different main conditions that will be explained in more detail in the next section.

- *DIF proportion*

  In addition to the null hypothesis, where no DIF is present, we regard different situations where either one or three out of ten items display DIF. This corresponds to DIF proportions of 10% or 30%, that were also found in empirical studies investigating non-uniform DIF: Maller (2001) examined the Wechsler Intelligence Scale for Children (third edition) and found e.g. 2 out of 21 studied items displaying non-uniform DIF in the Information subtest and 2 out of 20 items in the Vocabulary subtest. Pae (2004) found 3 out of 10 studied items to display non-uniform DIF in a Listening Comprehension subtest and 9 out of 31 studied items in a Reading Comprehension subtest of the English subtest of the 1998 Korean National Entrance Exam for Colleges and Universities.

In the second part of our simulation study, we focus on one exemplary setting with $(n_{\text{ref}}, n_{\text{foc}}) = (1500, 1500)$ where no ability differences are present to save space. The manipulated variables in the second part of the simulation study are the DIF proportion (defined similar to the first part) and the composition of the focal group.

- *Composition of the focal group*

  Strobl *et al.* (2013) have shown that a strength of Rasch trees is their ability to detect non-standard reference and focal groups. To illustrate this, in the second part of the simulation study the composition of the focal group is defined by either a binary group membership variable, an interaction of two binary group membership variables or by values of a numeric covariate. In the latter case, the groups form a u-shaped DIF pattern that reflects a situation where for example the middle aged subjects differ in their measurement properties from young and old subjects (cf. Strobl *et al.*, 2013).

### 4.3.2 Data generating processes

*Part I of the simulation study*

In order to evaluate the ability of the Rasch trees, the i-logit, and the v-logit method to detect uniform and non-uniform DIF, four main conditions are considered that will be described in the following.

The first main condition of *no DIF* reflects the null hypothesis for every method. All item difficulty parameters and all item discrimination parameters are equal in reference and focal group and over the ability range. The difficulty parameters are set to $\beta$ = (0.55, -0.29, 0.55, 0.30, -1.05, -1.07, -0.85, -0.36, -0.34, -0.37) and were selected from the parameter values used in the study of Finch and French (2007) in a way such that the mean test difficulty is almost identical to the mean test difficulty used by Finch and French (2007) and all discrimination parameters are set to 1. In the condition of no DIF, the ICCs are identical in both groups as displayed in Figure 13 (top-left). Thus, the Rasch model holds under the null hypothesis.

The remaining conditions reflect alternative hypotheses where DIF is present. In this case, DIF is simulated in one (namely item 2) or in three (namely item 2, 3, and 4) out of ten items. In the case of the *uniform DIF* condition, either one item parameter (10%) or three item parameters (30%) are more difficult for the focal group over the entire range of the latent variable as shown in Figure 13 (top-right) that represents the data generating process for one uniform DIF-item.

The difference in the item parameters between the focal group and reference group varies according to the proportion of DIF-items. If 10% DIF-items are present, the difference in the item parameters is set to $\Delta_{\text{DIF}} = \beta_{\text{foc}} - \beta_{\text{ref}} = 0.6$, whereas with 30% DIF-items the difference is reduced to $\Delta_{\text{DIF}} = 0.4$ to mimic a situation where a larger number of smaller DIF effects is present.

Non-uniform DIF is implemented in two different ways: the *jump condition* and the *discrimination condition*. The latter condition reflects the typically investigated condition (see, e.g., Swaminathan and Rogers, 1990; Finch and French, 2007) where non-uniform DIF is simulated by means of varying discrimination parameters (see the next paragraph). However, the definition of non-uniform DIF also includes more general shapes, such as the one in Figure 13, bottom-left, where the focal group consists of only of those group members that have a true ability of $-0.5$ or higher.[5] The DIF-item parameters of this group are, again, assigned the same values as in the uniform condition and, thus, the focal group has a disadvantage in this condition. Due to the shape of the ICCs for the DIF-items, this condition is termed jump.[6]

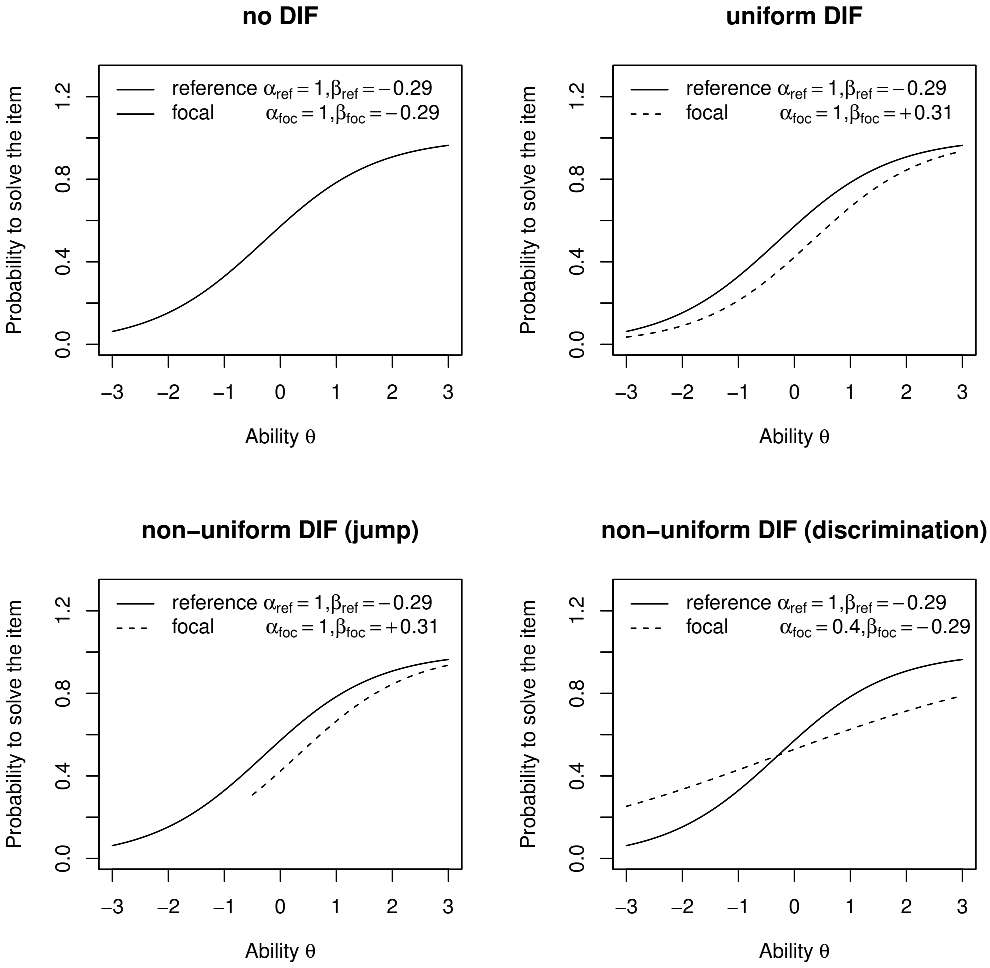

**Figure 13 –** *Four main conditions of no DIF, uniform DIF, non-uniform DIF (interaction), and non-uniform DIF (discrimination).*

In the fourth condition, the discrimination condition, non-uniform DIF is implemented by dif-

[5] The sample sizes always correspond to the members of the focal group. The ability distribution does not correspond to the focal group in the jump condition. In this condition, the abilities are drawn from the ability distribution $\theta \sim N(0,1)$ until the sample size $n_{\text{foc}}$ of the group members with $\theta_{\text{foc}} > -0.5$ is reached. The values lower than $-0.5$ are used for the reference group. The remaining observations that are required for the reference group to reach $n_{\text{ref}}$ are then drawn from either $\theta_{\text{ref}} \sim N(0,1)$ or $\theta_{\text{ref}} \sim N(0.5,1)$, depending on the setting.

[6] Note that, in this setting, the sample sizes correspond to the reference and focal group that are no longer directly defined by the group membership variable alone but by the combination of group members that have a true ability of $-0.5$ or higher.

ferent discrimination parameters in the reference and focal group. The DIF-items display a lower discrimination for the focal group, e.g. $\alpha_{\text{foc}} = 0.4$ than for the reference group $\alpha_{\text{ref}} = 1$ when one item has DIF. Then, in case of the 10% DIF-items, the difference in the discrimination parameters as simulated is $\delta_{\text{DIF}} = \alpha_{\text{foc}} - \alpha_{\text{ref}} = -0.6$, whereas in the case of 30% DIF-items, the difference is reduced to $\delta_{\text{DIF}} = -0.35$ where again $\alpha_{\text{ref}} = 1$ holds. Thus, in lower regions of the latent continuum, the focal group has an advantage, whereas the reference group has a higher probability of solving the item given the latent trait in upper regions of the latent continuum (cf. Figure 13, bottom-right, that displays the condition of one discrimination DIF-item).

The responses in the reference or focal group were simulated by means of a Rasch model (Rasch, 1960) when the discrimination parameters are set to 1, which is the case in the no DIF, the uniform DIF and the non-uniform jump condition. The 2pl model (Birnbaum, 1968) was used as the underlying model in the discrimination condition. The responses are generated in two steps. The probability of person $i$ solving item $j$ is computed by placing the item discrimination parameters $\alpha_j$, the item difficulty parameters $\beta_j$, and the person parameters $\theta_i$ in the 2pl model formula 16 that contains the Rasch model as a special case. The binary responses are then drawn from a binomial distribution with the resulting probabilities.

$$P(U_{ij} = 1 \,|\, \theta_i, \alpha_j, \beta_j) = \frac{\exp\{\alpha_j(\theta_i - \beta_j)\}}{1 + \exp\{\alpha_j(\theta_i - \beta_j)\}} \tag{16}$$

*Part II of the simulation study*
The second part addressing whether the Rasch trees can detect groups that display non-uniform DIF, is evaluated only for the discrimination condition in Section 4.4.6.

### 4.3.3 Outcome variables

The performance of the methods for DIF detection and also the ability to detect the DIF groups are evaluated by means of two main outcome variables.

- *DIF detection rate*

  The first outcome variable, the *DIF detection rate*, describes the proportion of replications where DIF is detected.

  Under the no DIF condition, it reflects the *false alarm* or *type one error rate*. Under the remaining conditions (the uniform, the jump, and the discrimination condition), this outcome variable corresponds to the *hit rate* or *power*.

- *Classification rate*

  Over all replications, the classification rate indicates the proportion of correct classifications of the type of DIF. When the classification rate is lower than the detection rate, it means that the correct type of DIF is not found.
  In the condition of no DIF, the classification is correct if no DIF is reported. Hence, the

classification rate can easily be calculated as one minus the false alarm rate and is not displayed here.

In the condition of uniform DIF, the classification rate represents the proportion of replications where uniform DIF is detected correctly. For the Rasch trees, the classification is correct if only the group membership variable – but not the sum score or the noise variable – is selected for splitting. For the i-logit or the v-logit method, the classification is correct if DIF is detected and the group membership variable is the only variable that obtains a significant post-hoc test result.

In the two non-uniform DIF settings, the jump and the discrimination condition, the classification of the type of DIF is correct if the variables that correspond to non-uniform DIF but not the noise variable indicate DIF. When the Rasch trees are investigated, both variables, the group membership and the sum score, have to occur in at least one split but the noise variable is not allowed to occur. For both extensions of the logistic regression, the classification is correct if DIF is detected and the interaction term shows a significant post-hoc test result (independent of the result for the group membership variable); at the same time the noise variable is not allowed to display a significant post-hoc test for the correct classification. In addition to the classification and the detection rate, information of the splitting variables and the post-hoc test results is stored for later analysis.

## 4.4 Results

In the following section, the results of the simulation study are presented. First, the results in the four main conditions including the no DIF, the uniform and both non-uniform DIF conditions (part one of the simulation study) are presented with regard to the DIF detection rate and the classification rate. Second, the selection of the splitting variables for the Rasch trees and the results of the post-hoc tests for the i-logit and v-logit method will be discussed in more detail. Finally, in Section 4.4.6 the ability of the Rasch trees to detect groups displaying non-uniform DIF is addressed (part two of the simulation study).

### 4.4.1 No DIF

In the no DIF condition only the false alarm rate was calculated. As can be seen in Figure 14 (left), all methods yielded a well-controlled false alarm rate when the reference and the focal group had the same mean abilities.

In the presence of ability differences, the Rasch trees again largely held the alpha level while inflated false alarm rates occurred for the i-logit as well as for the v-logit method (see Figure 14, right). Thus, these methods cannot fairly be compared (Li and Stout, 1996) with the Rasch trees. Similar findings occurred for the item-wise logistic regression (Narayanan and Swaminathan 1996, for an overview of related problems see DeMars 2010 and the references therein). The maximum of the observed false alarm rates was 0.07 for the Rasch trees, whereas for the logistic regression methods they were notably inflated to 0.11 for the i-logit, and 0.14 for the v-logit method. Such inflated false alarm rates no longer allow a correct statistical inference at the prespecified significance level, jeopardize the correct classification of DIF- and DIF-free items and "*can result in the inefficient use of testing resources, and [...] may interfere with the study of the underlying causes of DIF*" (Jodoin and Gierl, 2001, p. 329).

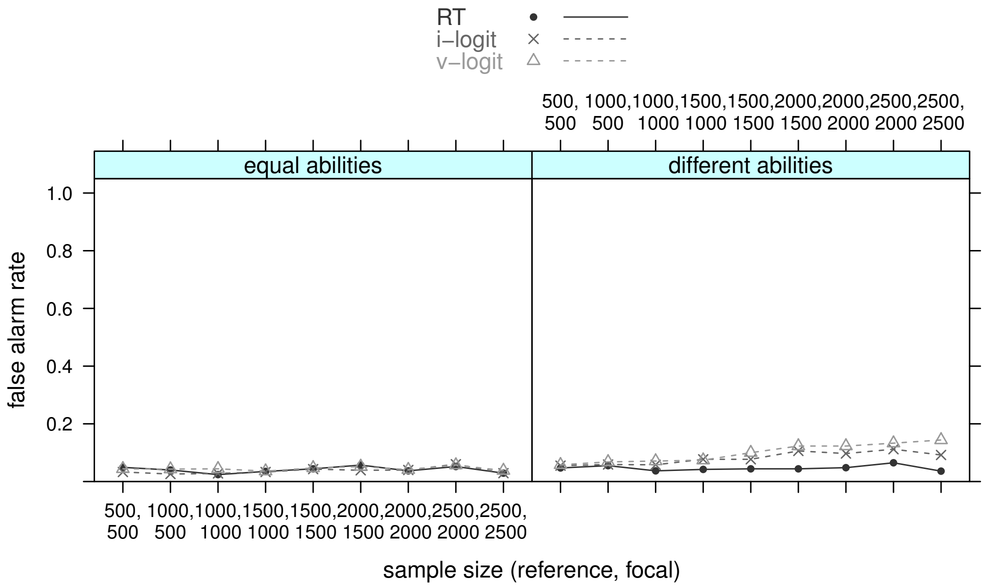

**Figure 14** – *False alarm rates under the null hypothesis of no DIF.*

### 4.4.2 Uniform DIF

In case of the uniform DIF condition, the detection rate in Figure 15 (left column) contains the proportion of replications where DIF was correctly detected and the classification rate (right column) indicates whether the correct type of DIF was found.

When the simulated data sets contained 10% uniform DIF-items, all methods showed a rapidly rising hit rate that converged against one in the simulated range of sample sizes (see Figure 15, row 1 left). The highest hit rate occurred for the i-logit method followed by the Rasch trees. In case of 30% uniform DIF-items (see Figure 15, row 2 left), the Rasch trees yielded the highest hit rate followed by the v-logit method. This reflects the fact discussed in Section 4.2.1 that the multiple testing i-logit procedure is likely to miss several smaller DIF effects. There was no visible difference between groups with equal abilities (see Figure 15, top-half left) and different abilities (see Figure 15, bottom-half left).

The classification rates are depicted in Figure 15 (right column). As opposed to the hit rates, the classification rates of the type of DIF were clearly lower for the extensions of the logistic regression. The best results occurred for the Rasch trees independent of the presence of ability differences or the proportion of DIF-items. Both extensions of the logistic regression led to notably lower classification rates.

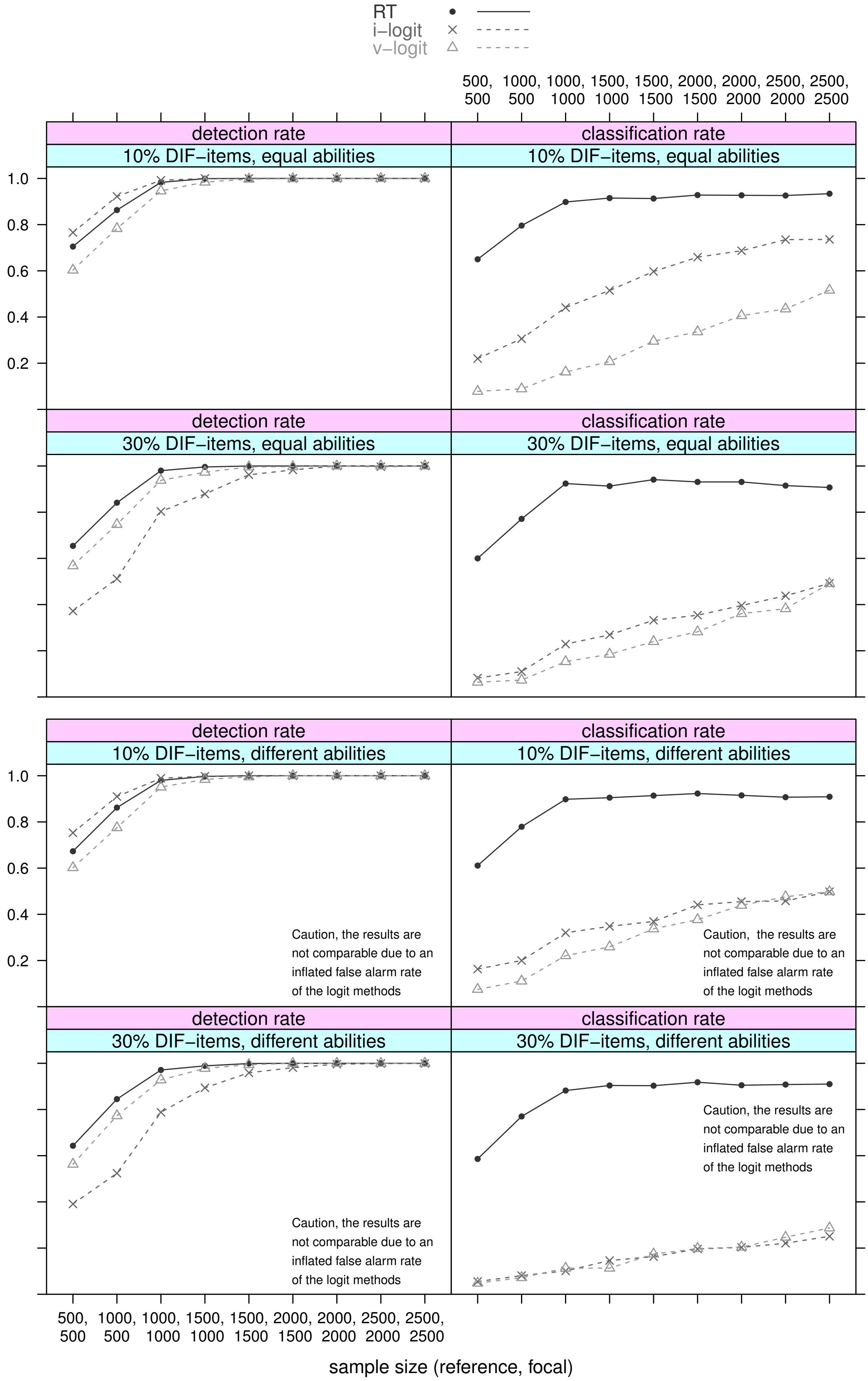

**Figure 15 –** *Detection and classification rates under the uniform DIF condition.*

In the setting of 10% DIF-items and equal abilities (see Figure 15, row 1 right), the i-logit method reached a maximum observed rate of 0.74 compared to 0.52 for the v-logit method and 0.93 for the Rasch trees. In all remaining settings of the uniform DIF condition, the difference between the logit methods was smaller. Generally, in conditions with different abilities (see Figure 15, bottom-half right), both extensions of the logistic regression performed poor with classification rates of the DIF type not exceeding 0.50 as opposed to the Rasch trees that reached true classification rates about 0.90 when the group size was 1000 and still held the alpha level (see again Figure 14, right).

### 4.4.3 Non-uniform DIF: Jump condition

The non-uniform jump condition is the first condition where non-uniform DIF was present and the focal group consisted of group members that had a true ability value of $-0.5$ or higher.

We first investigate the hit rates when no ability differences were present (see Figure 16, top-half left), since all methods held the significance level under the null hypothesis (see Section 4.4.1) and can, thus, be compared in a fair way. The hit rates varied over the manipulated variables. In the setting of 10% DIF-items and equal abilities (see Figure 16, row 1 left), the hit rate of the i-logit method outperformed the hit rate of the Rasch trees and the v-logit method in regions of medium sample sizes. In case of 30% DIF-items and equal abilities (see Figure 16, row 2 left), the multiple testing i-logit procedure again displayed a lower hit rate and the Rasch trees achieved the highest hit rate.

When ability differences were present (see Figure 16, bottom-half), again the i-logit method reached the highest hit rate for 10% DIF and the v-logit method displayed the highest hit rate for 30% DIF. Note that, in the case of different ability distributions, both logit methods yielded inflated false alarm rates (see again Figure 14, right) and can, thus, not fairly be compared (Li and Stout, 1996) with the Rasch trees.[7] When ability differences were present and the sample size was large, the Rasch trees, the i-logit and the v-logit method led to similar hit rates, but the Rasch trees held the alpha level and should, thus, be preferred in this case.

Figure 16 (right column) shows the classification rate for the non-uniform DIF jump condition. If no ability differences were present, the i-logit method had the highest classification rate when the sample size was small (see Figure 16, top-half right). When the sample size was medium or large, the Rasch trees led to the highest classification rate in the majority of the simulated settings.

In case of unequal abilities (see Figure 16, bottom-half right), the i-logit method reached the highest classification rate in the majority of the investigated settings, but remember again the logit extensions displayed an inflated false alarm rate when ability differences were present and cannot fairly be compared with the Rasch trees.

[7] Note, however, that the underlying data generating process here is different to the null hypothesis presented in Figure 14 (see again Section 4.3.2). Under the corresponding null hypothesis, the false alarm rates of the logit methods were slightly lower but still inflated (results not shown). Thus, our argument that the methods cannot be compared in a fair way still applies here.

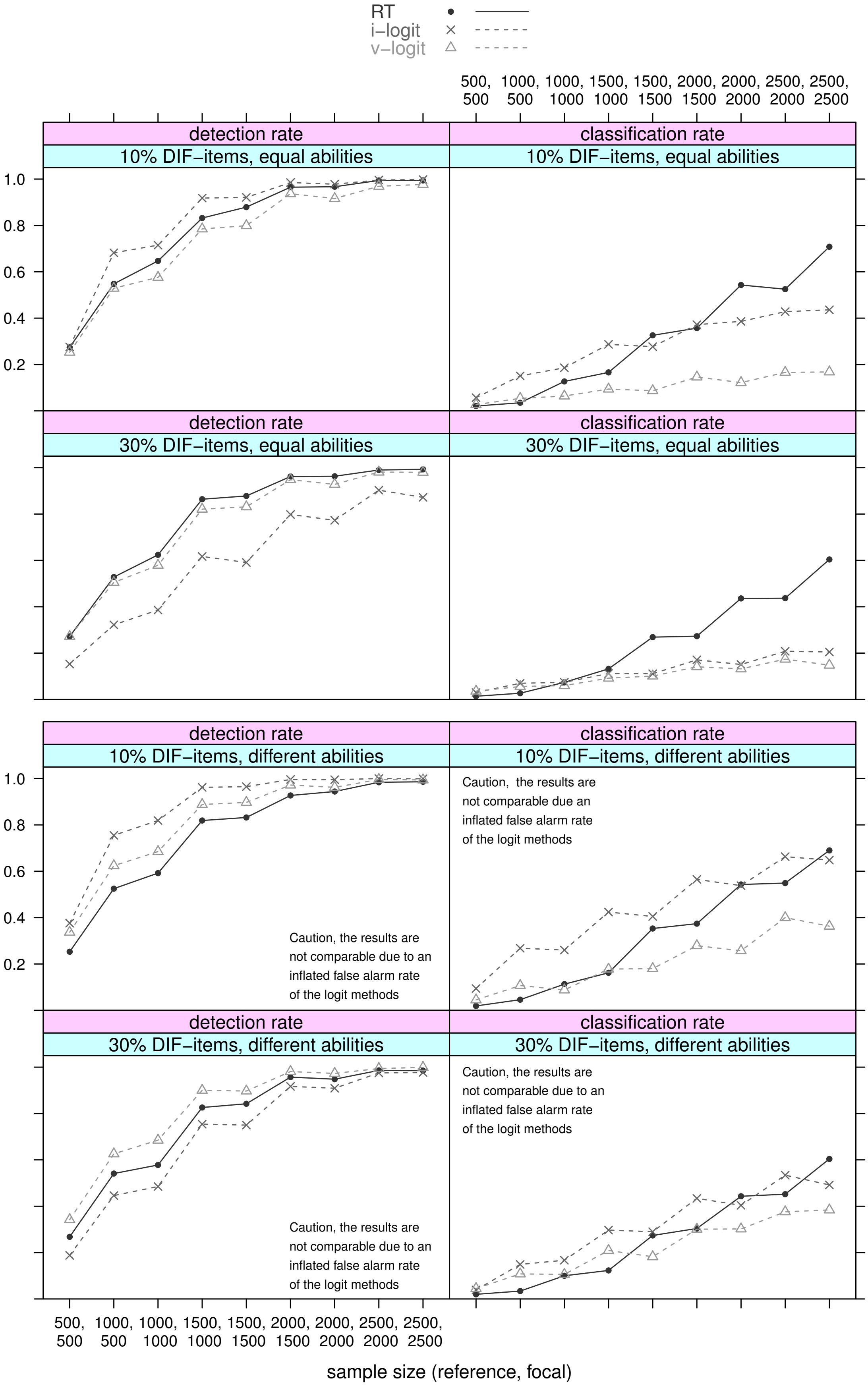

**Figure 16 –** *Detection and classification rates under non-uniform DIF jump condition.*

### 4.4.4 Non-uniform DIF: Discrimination condition

The non-uniform DIF discrimination condition reflects the situation where the DIF-items were assigned different discrimination parameters and the focal group consisted of members of a binary variable. As can be seen from Figure 17 (left column), the Rasch trees displayed lower hit rates than the extensions of the logistic regression, but also converged to one when the sample size was large enough. When 10% DIF-items were present (Figure 17, row 1 left and row 3 left), the i-logit method outperformed again the other methods. On the other hand, when 30% DIF-items were present (Figure 17, row 2 left and row 4 left), the v-logit method reached higher hit rates.

If ability differences were present (Figure 17, bottom-half left), the advantage of the extensions of the logistic regression seems larger, but again remember in that case both logit methods yielded inflated false alarm rates (see again Figure 14, right) and cannot fairly be compared with the Rasch trees.

One additional finding here is that the Rasch trees displayed lower hit rates when unequal sample sizes were present (the hit rates in Figure 17, left column, are zig-zagged rather than monotonically increasing), even though the absolute sample sizes increased. We will focus on the selection of the splitting variables for the Rasch trees to explain these findings in Section 4.4.5.

In case of the discrimination condition, the classification rates in Figure 17 (right column) behaved very similar to the hit rates (left column) but slightly lower. Here, again, the i-logit method was superior when only 10% DIF-items were present (Figure 17, row 1 and row 3). In case of 30% DIF-items (Figure 17, row 2 and row 4), the v-logit procedure reached the highest classification rate. The Rasch trees were clearly far behind the other procedures, but in principle also able to detect the differing discrimination parameters when the sample size was large enough. The reason for this is that Rasch trees rely on the item difficulty parameters of the Rasch model only and are, thus, less flexible in this case compared to the logistic methods that capture non-uniform DIF in one single parameter. The classification rate for the Rasch trees increased only slightly when the sample sizes increased only in the reference group. Hence, the selected splitting variables (and also the post-hoc test results) will be addressed in Section 4.4.5 to allow for a better understanding of the simulation results.

### 4.4.5 On the classification accuracy

In this section, the variables selected for splitting and the results of the post-hoc tests are discussed in more detail for two exemplary settings. We restrict the detailed analysis to one setting where the sample size in each group consists of 1500 examinees and one setting of unequal sample sizes $(n_{\text{ref}}, n_{\text{foc}}) = (2000, 1500)$ in order to save space.

Table 3 (upper part) contains the information of the classification rate (column ‘true’ typed in italic) and of the splitting variables or significant post-hoc test results for the setting of 1500 observations in each group.

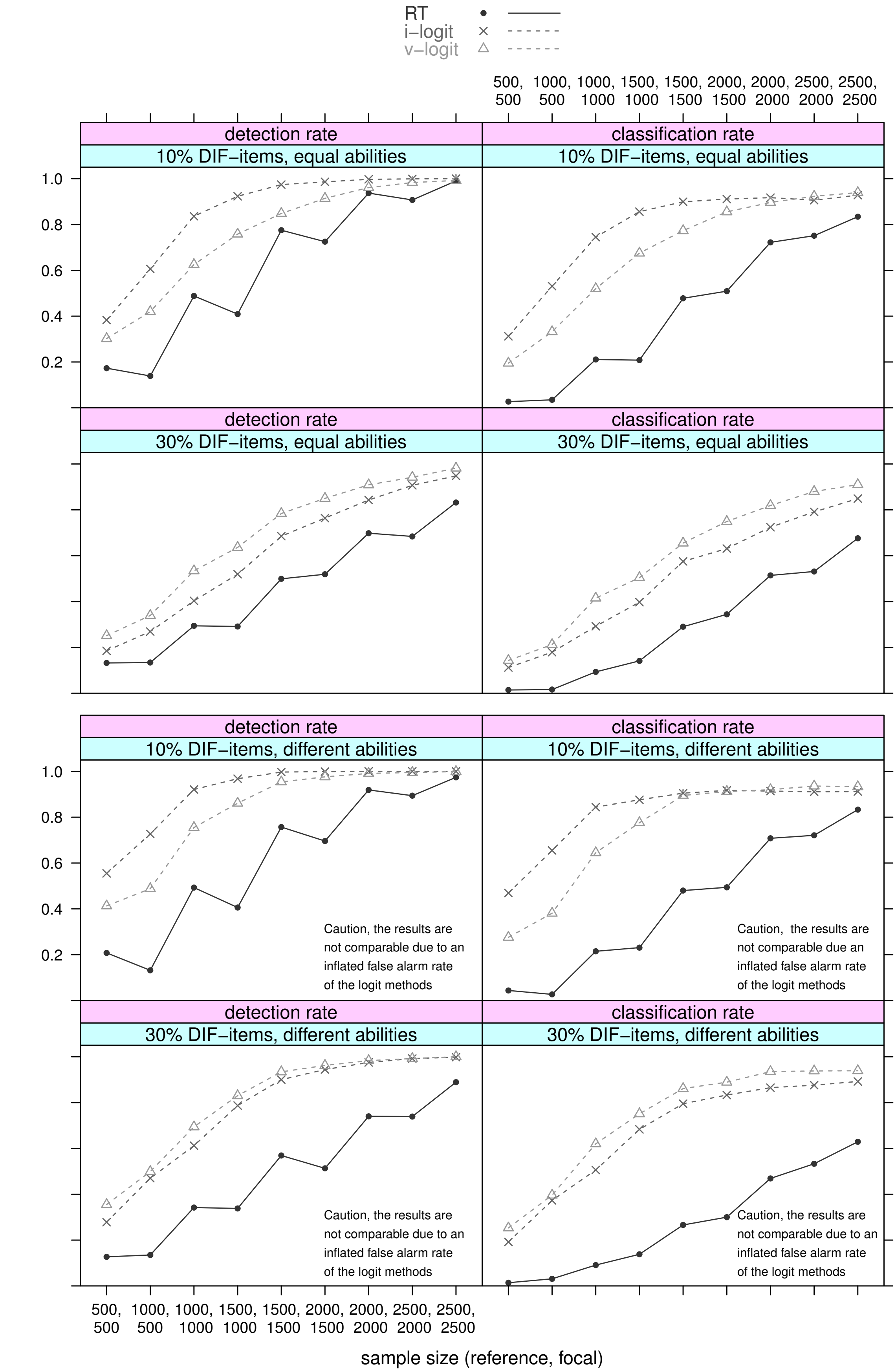

**Figure 17 –** *Detection and classification rates under the non-uniform discrimination condition.*

For the Rasch trees, the potential splitting variables were the noise $x_{\text{noise}}$, the group membership $x_{\text{group}}$, the sum score variable $x_{\theta}$, and the interaction $x_{\text{group},\theta}$ meaning that both types of variables, the group membership variable together with the sum score variable, were selected in one Rasch tree. The columns in Table 3 refer to the proportion of replications where the respective variables were selected for splitting. As opposed to the Rasch trees, the extensions of the logistic regression have no column regarding the main effect of the sum score since the sum score is used only as a conditioning variable in the logistic regression. Its effect is not tested post-hoc because it is not relevant for DIF classification suggested by Swaminathan and Rogers (1990). The columns indicate the proportion of significant post-hoc test results corresponding to the effects of the noise variable $x_{\text{noise}}$, the group membership variable $x_{\text{group}}$ and the interaction of the sum score and the group membership variable $x_{\text{group},\theta}$.

In the condition of no DIF-items, the true classification rate was defined as one minus the false alarm rate. What is interesting to note here again, is that all methods yielded similar results when no ability differences were present, but in the case of ability differences, the methods based on the logistic regression yielded inflated false alarm rates; especially the group membership and the interaction variable then displayed (erroneously) significant post-hoc tests (see the first two rows and the columns *true*, $x_{\text{group}}$ and $x_{\text{group},\theta}$ in Table 3).

When differing difficulty parameters were present in the uniform DIF condition, only the group variable should be used for splitting or obtain a significant post-hoc test result. As can be seen from Table 3, the Rasch trees almost always selected the group variable for splitting (see the bold marked column). Note, that the true classification rate was lower since it includes only those replications where the group variable was the only variable used for splitting. The remaining variables were rarely selected for splitting. In case of the i-logit method, the rate for the significant post-hoc tests for the group variable was far higher than the true classification rate that, again, includes only those replications where the group variable was the only variable with a significant post-hoc test result.

This can be explained through the high proportion of significant post-hoc tests for the remaining variables: the noise and the interaction variable. One possible explanation for this finding is that only those variables that were found to be significant in the global Wald test are included in the post-hoc tests for the i-logit method. Here, statistical tests were connected in a series that might deform the distribution of the test statistic as discussed by Leeb and Pötscher (2005) and Berk *et al.* (2010). To overcome this problem, the post-hoc tests might be based on new data or replaced by a descriptive analysis. For the v-logit method, the proportion of significant post-hoc test results for the group variable was low. The remaining variables obtained high proportions of erroneously significant results, but slightly lower proportions compared to the i-logit method. The Rasch trees were superior, since they rarely selected the noise and the sum score variable for splitting and, thus, led to a higher classification accuracy.

In the non-uniform DIF jump condition, the true classification rate represents the proportion where non-uniform DIF was detected correctly. For the Rasch trees, both variables, the group and the sum score, had to occur in at least one split (see the bold marked column) and the noise variable was not allowed to be used for splitting.

| | **Rasch trees** | | | | | **i-logit** | | | | **v-logit** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *true* | $x_{\text{noise}}$ | $x_{\text{group}}$ | $x_{\theta}$ | $x_{\text{group},\theta}$ | *true* | $x_{\text{noise}}$ | $x_{\text{group}}$ | $x_{\text{group},\theta}$ | *true* | $x_{\text{noise}}$ | $x_{\text{group}}$ | $x_{\text{group},\theta}$ |
| $(1500, 1500)$ | | | | | | | | | | | | | |
| **no DIF** | | | | | | | | | | | | | |
| equal 0% | *0.96* | 0.02 | 0.02 | 0.01 | 0.00 | *0.96* | 0.02 | 0.02 | 0.02 | *0.95* | 0.02 | 0.01 | 0.01 |
| unequal 10% | *0.96* | 0.02 | 0.02 | 0.01 | 0.00 | *0.92* | 0.03 | 0.05 | 0.05 | *0.90* | 0.03 | 0.05 | 0.04 |
| **uniform DIF** | | | | | | | | | | | | | |
| equal 10% | *0.91* | 0.06 | **1.00** | 0.03 | 0.03 | *0.60* | 0.11 | **0.73** | 0.09 | *0.30* | 0.06 | **0.34** | 0.03 |
| equal 30% | *0.94* | 0.04 | **1.00** | 0.02 | 0.02 | *0.33* | 0.10 | **0.44** | 0.10 | *0.24* | 0.06 | **0.29** | 0.03 |
| unequal 10% | *0.91* | 0.05 | **1.00** | 0.05 | 0.05 | *0.37* | 0.11 | **0.52** | 0.14 | *0.34* | 0.06 | **0.45** | 0.11 |
| unequal 30% | *0.90* | 0.06 | **1.00** | 0.04 | 0.04 | *0.16* | 0.12 | **0.26** | 0.13 | *0.17* | 0.05 | **0.21** | 0.05 |
| **jump** | | | | | | | | | | | | | |
| equal 10% | *0.33* | 0.06 | 0.82 | 0.41 | **0.36** | *0.28* | 0.09 | 0.07 | **0.31** | *0.09* | 0.06 | 0.04 | **0.10** |
| equal 30% | *0.27* | 0.06 | 0.84 | 0.34 | **0.30** | *0.11* | 0.11 | 0.10 | **0.13** | *0.10* | 0.07 | 0.04 | **0.11** |
| unequal 10% | *0.35* | 0.05 | 0.80 | 0.42 | **0.39** | *0.40* | 0.10 | 0.10 | **0.45** | *0.18* | 0.06 | 0.10 | **0.19** |
| unequal 30% | *0.27* | 0.05 | 0.82 | 0.32 | **0.30** | *0.29* | 0.10 | 0.14 | **0.33** | *0.18* | 0.06 | 0.07 | **0.19** |
| **discrimination** | | | | | | | | | | | | | |
| equal 10% | *0.48* | 0.05 | 0.52 | 0.77 | **0.52** | *0.90* | 0.07 | 0.95 | **0.97** | *0.77* | 0.06 | 0.71 | **0.83** |
| equal 30% | *0.29* | 0.03 | 0.35 | 0.45 | **0.30** | *0.57* | 0.09 | 0.64 | **0.65** | *0.66* | 0.06 | 0.74 | **0.71** |
| unequal 10% | *0.48* | 0.06 | 0.53 | 0.75 | **0.52** | *0.90* | 0.09 | 0.98 | **1.00** | *0.90* | 0.05 | 0.83 | **0.95** |
| unequal 30% | *0.27* | 0.04 | 0.32 | 0.53 | **0.29** | *0.80* | 0.10 | 0.87 | **0.90** | *0.86* | 0.07 | 0.93 | **0.93** |
| $(2000, 1500)$ | | | | | | | | | | | | | |
| **no DIF** | | | | | | | | | | | | | |
| equal 0% | *0.94* | 0.03 | 0.02 | 0.01 | 0.00 | *0.96* | 0.02 | 0.02 | 0.02 | *0.95* | 0.03 | 0.02 | 0.01 |
| unequal 10% | *0.96* | 0.01 | 0.03 | 0.01 | 0.00 | *0.89* | 0.04 | 0.07 | 0.07 | *0.88* | 0.03 | 0.06 | 0.05 |
| **uniform DIF** | | | | | | | | | | | | | |
| equal 10% | *0.93* | 0.05 | **1.00** | 0.02 | 0.02 | *0.66* | 0.09 | **0.78** | 0.08 | *0.34* | 0.06 | **0.39** | 0.04 |
| equal 30% | *0.93* | 0.05 | **1.00** | 0.02 | 0.02 | *0.35* | 0.11 | **0.47** | 0.11 | *0.28* | 0.06 | **0.35** | 0.06 |
| unequal 10% | *0.92* | 0.04 | **1.00** | 0.04 | 0.04 | *0.44* | 0.11 | **0.60** | 0.17 | *0.38* | 0.06 | **0.52** | 0.14 |
| unequal 30% | *0.92* | 0.04 | **1.00** | 0.05 | 0.05 | *0.20* | 0.11 | **0.30** | 0.15 | *0.20* | 0.07 | **0.26** | 0.08 |
| **jump** | | | | | | | | | | | | | |
| equal 10% | *0.36* | 0.06 | 0.94 | 0.40 | **0.39** | *0.37* | 0.10 | 0.08 | **0.41** | *0.15* | 0.07 | 0.04 | **0.16** |
| equal 30% | *0.27* | 0.07 | 0.95 | 0.32 | **0.31** | *0.17* | 0.13 | 0.11 | **0.20** | *0.14* | 0.06 | 0.04 | **0.15** |
| unequal 10% | *0.37* | 0.07 | 0.92 | 0.42 | **0.42** | *0.56* | 0.09 | 0.14 | **0.61** | *0.28* | 0.06 | 0.14 | **0.29** |
| unequal 30% | *0.30* | 0.06 | 0.95 | 0.33 | **0.33** | *0.43* | 0.11 | 0.19 | **0.48** | *0.30* | 0.05 | 0.10 | **0.32** |
| **discrimination** | | | | | | | | | | | | | |
| equal 10% | *0.51* | 0.05 | 0.55 | 0.72 | **0.55** | *0.91* | 0.08 | 0.98 | **0.99** | *0.86* | 0.05 | 0.82 | **0.91** |
| equal 30% | *0.34* | 0.04 | 0.43 | 0.45 | **0.36** | *0.63* | 0.11 | 0.73 | **0.73** | *0.75* | 0.06 | 0.81 | **0.80** |
| unequal 10% | *0.49* | 0.06 | 0.54 | 0.68 | **0.54** | *0.92* | 0.08 | 0.99 | **1.00** | *0.91* | 0.06 | 0.90 | **0.98** |
| unequal 30% | *0.30* | 0.04 | 0.36 | 0.46 | **0.32** | *0.83* | 0.11 | 0.94 | **0.94** | *0.89* | 0.07 | 0.96 | **0.96** |

**Table 3** – *Classification rate, splitting variables and post-hoc test results in all conditions, equal and unequal ability distributions as well as* 0%, 10% *or* 30% *DIF-items for 1500 observations in each group (upper part) and for 2000 observations in the reference and 1500 observations in the focal group (lower part).*

The combination of the group and the sum score variable was frequently found in case of equal abilities compared to the other methods. Generally, the group membership variable is more often selected for splitting compared to the sum score. The true classification for the i-logit and v-logit method is a significant post-hoc test for the interaction (regardless of the group variable) but with no significant result for the noise variable that was, again, often found to be significant for the i-logit method. Except those setting with unequal abilities (where the methods cannot be compared in a fair way due to the inflated false alarm rate of the extensions of the logistic regression), the Rasch trees outperformed the extensions of the logistic regression.

In the non-uniform DIF discrimination condition, the true classification is defined similarly to the non-uniform DIF jump condition (described above). Interestingly, the Rasch trees here displayed a different pattern. While the sum score variable was frequently used for splitting, the group variable was selected in fewer iterations. This setting was challenging for the Rasch trees since the item difficulty parameters for both groups were identical and only the discrimination parameters differed. In contrast to methods that catch non-uniform DIF in one single parameter, Rasch trees rely on the difference of the item difficulty parameters only. Since these item parameters appear similar for both groups for the Rasch tree, the method needs to first detect parameter instability with respect to the sum score. Regarding the sum score variable, different item difficulties should be detected since the different slopes of the ICCs of the DIF-items are translated to different item difficulties in groups defined by the observed sum score. The highest difference in the item difficulty parameters should occur in groups defined at the highest or lowest ability levels and these are the regions where splitting should occur (see, again, the example in Section 4.2.2). Then, the group variable needs to be selected in the resulting subsamples that consist of fewer observations and, thus, the power to detect instable item parameters decreases due to the reduced sample size. Consequently, the selection of both variables was found in a lower number of iterations than in the post-hoc tests of the i-logit and the v-logit method. Again, the i-logit method often displayed erroneously significant post-hoc tests for the noise variable.

Table 3 (second half) displays the results for DIF conditions where the reference group consisted of 2000 examinees and the focal group included 1500 examinees. Here we discuss only the discrimination condition since the Rasch trees displayed lower hit rates in this condition even if the total number of observations in the reference group increased.

In the non-uniform DIF discrimination condition, the Rasch trees displayed lower hit rates and only slightly increasing classification rates when only the sample size of the reference group increased (see again Figure 17, row 1 to row 4). As can be seen from Table 3 by comparing the discrimination condition in the upper part to the discrimination condition in the lower part, the rate of selecting the noise variable was constant in both conditions of the sample size. On the other hand, the rate of selecting the group variable was slightly higher in case of 2000 observations in the reference group. This was the reason for the slight increase of the true classification rate.

However, in all settings where the hit rate was zig-zagged in Figure 17 (left column, equal 10%, unequal 10%, unequal 30%), the rate of selecting the sum score variable was lower when the sample sizes of reference and focal group were unequal in Table 3 (lower part). This results in a lower hit rate, since it means that non-uniform DIF was detected less frequently. We attribute this fact to the following mechanism. To approximate varying slopes by only using item difficulty parameters, Rasch trees need to select the ability variable for splitting. In case of equal sample sizes in reference and focal group, the differences between the item parameters in regions of high or low ability levels (that occur due to the group membership) are easier to determine compared to situations where the reference group consists of more observations.

The reason for this is that when only the reference group obtains more observations, the differences in the item difficulty parameters associated with the sum score do not become more clear, as one might except, but become more indistinct. Different item discrimination parameters are simulated only for observations of the focal group. Thus, more observations in the reference group display the same item difficulties associated with the sum score instead of the differences that only occur together with observations of the focal group. When only the sample size of the reference group is increased, the differences in the item parameters associated with the sum score become more indistinct and the power of the underlying test decreases.

### 4.4.6 On the detection of DIF groups

In this second part of the simulation study, the composition of the focal group is varied. The focal group is defined either by a binary group membership variable (termed binary dummy), an interaction of two binary group membership variables (termed binary interaction) or by a u-shaped pattern in a numeric covariate (termed numeric u-shaped). The i-logit and the v-logit methods are now implemented in several ways. The “binary logistic” methods obtain only one dummy variable. In case of the numeric u-shaped pattern the binary variable is defined by a median split in the numeric covariate. The “saturated logistic” methods obtain all main and interaction effects that result when two binary covariates and the sum score variable are included in the model. The “numeric logistic” regression methods obtain the numeric covariate instead of the binary dummy variable.

*Binary group membership*

Table 4 contains the results. In case the focal group corresponded to the binary group membership (binary dummy), the logistic regression methods outperformed the Rasch trees by yielding a higher detection and classification rate as expected from the results presented in Section 4.4.4.

*Interaction of two binary group membership variables*

In case the focal group corresponded to the interaction of two binary group membership variables (binary interaction) the results were quite different. The i-logit binary and the v-logit binary methods were misspecified in this situation and were assigned a classification rate of zero. The Rasch trees outperformed the binary logistic regression methods by yielding both a higher detection and a higher classification rate. If, however, both binary group variables were handed over to the logit methods,[8] the saturated i-logit method reached the best results in case

[8] The classification was then defined correct if the interaction of both groups and the sum score obtained a significant post-hoc test, but the noise variable did not.

10% DIF items were present and the saturated v-logit method in case of 30%. Thus, the good performance of the logistic regression methods strongly depends on the correct knowledge of the affected DIF groups. Rasch trees may be useful to detect DIF even the discrimination condition when the group composition is unknown. Note however, that the classification rate of the Rasch trees is really small since the variable selection task is challenging for the Rasch trees in this situation as discussed in the previous section.

| | | **Rasch** | **i-logit** | | | **v-logit** | | |
|---|---|---|---|---|---|---|---|---|
| | | **trees** | binary | saturated | numeric | binary | saturated | numeric |
| $(1500, 1500)$ | | | | | | | | |
| **binary dummy** | | | | | | | | |
| equal 10% | detection | 0.78 | 0.97 | | | 0.84 | | |
| equal 10% | classification | 0.47 | 0.88 | | | 0.76 | | |
| equal 30% | detection | 0.48 | 0.66 | | | 0.76 | | |
| equal 30% | classification | 0.27 | 0.55 | | | 0.63 | | |
| **binary interaction** | | | | | | | | |
| equal 10% | detection | 0.75 | 0.62 | 0.87 | | 0.45 | 0.60 | |
| equal 10% | classification | 0.07 | 0.00 | 0.44 | | 0.00 | 0.15 | |
| equal 30% | detection | 0.41 | 0.24 | 0.37 | | 0.36 | 0.48 | |
| equal 30% | classification | 0.03 | 0.00 | 0.11 | | 0.00 | 0.09 | |
| **numeric u-shape** | | | | | | | | |
| equal 10% | detection | 0.79 | 0.04 | | 0.04 | 0.06 | | 0.06 |
| equal 10% | classification | 0.09 | 0.01 | | 0.01 | 0.02 | | 0.02 |
| equal 30% | detection | 0.38 | 0.05 | | 0.04 | 0.04 | | 0.05 |
| equal 30% | classification | 0.05 | 0.01 | | 0.01 | 0.00 | | 0.00 |

**Table 4** – *Detection and classification rate for different group compositions in case of* 10% *or* 30% *DIF-items for 1500 observations in each group.*

*Numeric u-shaped*

The numeric u-shaped condition is challenging for all logistic regression methods due to cancellation effects. All versions of the logistic regression methods had much lower detection rates compared to the Rasch trees. Again, the classification rates were really low, also for the Rasch trees, but the detection rates showed that the latter were at least able to detect that DIF is present in the test at all.

## 4.5 Discussion

To conclude, the results from the simulation study are briefly reviewed together with an outline of the implications for practical research. In our simulation study, we investigated the appropriateness of the Rasch trees in the detection of non-uniform DIF in the entire set of items. We presented two extensions of the logistic regression by using a correction for the multiple testing problem (i-logit) and by applying a global likelihood ratio test (v-logit) to allow for a direct comparison with the Rasch trees.

In the condition of the null hypothesis (no DIF), the Rasch trees always showed a well-controlled false alarm rate, while both extensions of the logistic regression displayed inflated false alarm rates when ability differences were present. Similar findings occurred for the item-wise logistic regression in the study of Narayanan and Swaminathan (1996). When using the extensions of the logistic regression in situations where ability differences are present, inflated false alarm rates are to be expected that jeopardize the results of the DIF analysis as well as the associated investigation of the causes of DIF (Jodoin and Gierl, 2001).

In case of the uniform DIF condition, where DIF in the difficulty parameters was simulated between two groups, the performance of the methods depended on the proportion of DIF-items. In case of 10% DIF-items, the i-logit method displayed the highest hit rate, whereas 30% DIF-items were most frequently detected by the Rasch trees. Opposed to the hit rate, the classification accuracy of the DIF methods was lower. In all situations of the uniform DIF condition, the Rasch trees reached far higher classification rates.

Non-uniform DIF was first investigated in the non-uniform DIF jump condition where the focal group consisted of binary group members with $\theta_{\text{foc}} > -0.5$ that obtained different item difficulty parameters. In case of 10% non-uniform DIF-items, the i-logit method displayed the highest hit rate in regions of medium sample sizes. Otherwise, the Rasch trees again outperformed the extensions of the logistic regression by yielding a high hit rate and holding the alpha level. When the sample sizes were in medium or large regions, the Rasch trees also yielded the highest classification rates.

In the non-uniform DIF discrimination condition, different discrimination parameters were assigned to the group members and the extensions of the logistic regression clearly outperformed the Rasch trees. In case of 10% DIF-items, the i-logit method reached the highest hit and classification rate; whereas, with 30% DIF-items the v-logit method was superior. Nevertheless, in case of large sample sizes, the Rasch trees also yielded a high hit and classification rate and were, thus, in principle able to also detect this type of non-uniform DIF.

In summary, the study showed that Rasch trees provide a flexible approach for the detection of uniform DIF and non-uniform DIF. In case of differing item difficulty parameters (in the uniform DIF and non-uniform jump condition), Rasch trees were often found to be superior while the alpha level was well-controlled. In the case of DIF in the discrimination parameters between two groups (non-uniform discrimination condition), the i-logit and v-logit method clearly outperformed the Rasch trees by using a single model coefficient.

Additionally, we investigated scenarios, where the true focal group consisted of an interaction between two binary variables or of the middle range of a numeric covariate. Rasch trees then displayed a comparable or higher hit rate compared to the (misspecified) binary versions of the logistic regression methods even though the discrimination condition was regarded. Thus, Rasch trees may prove useful to detect DIF in non-standard patterns, such as interactions of more than one grouping variable or u-shaped patterns, that would usually not be explicitly specified – and thus missed – in logistic regression DIF analysis. However, one drawback is that the classification rate may be rather low if DIF occurs only in the discrimination parameters.

For practical research, we recommend to use both methods, the logistic regression methods as well as the Rasch trees, and to compare the results. Rasch trees allow to assess non-uniform DIF in a flexible way when the data set is large enough. High sample sizes might often be found in psychological or educational testing situations, especially in large-scale assessments. In the case of prior information that the proportion of DIF-items is very small, e.g. one out of ten items displays DIF, and that no ability differences are present, the multiple comparison i-logit procedure is expected to reach satisfying results. But, again, these methods need the prior knowledge of the groups, whereas the Rasch trees automatically search for the DIF groups in the set of available covariates and are also applicable when the DIF groups are not defined prior to data analysis.

The simulation results also showed that the post-hoc inspections are important. In various situations, the true classification rate was far behind the rate of detecting DIF. This highlights that careful considerations of the type of DIF after the global test for (non-) uniform DIF are important to correctly assess the type of DIF. If the type of DIF is misclassified by the methods, wrong steps could be taken or the psychological source of DIF might be missed. Thus, the development and evaluation of post-hoc classifications of the type of DIF remain important tasks.

This study was limited to one type of DIF in the data set. In practical testing situations, both types of DIF can overlap. Therefore, the descriptive analysis of the type of DIF on the item level is important. Methods that do not rely on the post-hoc tests described here are the graphical exploration or the comparison of $R^2$ values (Gelin and Zumbo, 2003; Zumbo, 1999). Future research might compare those methods to the post-hoc Wald tests. Furthermore, the study was restricted to situations where the different data generating processes relied on the Rasch or the 2pl model. Future research may address further data generating processes such as the 3pl model or non-parametric ICCs. Moreover, we will try to address the generalization of the Rasch tree method to the 2pl or Birnbaum model (Birnbaum, 1968), that may prove helpful for the analysis of non-uniform DIF and the generalization to a 3pl model including a location and a guessing parameter.

# 5 A framework for anchor methods

***Summary:*** *In the analysis of differential item functioning (DIF) using item response theory (IRT), a common metric for the item parameters is necessary to compare item parameters between groups of test-takers. In the Rasch model, the same restriction is placed on the item parameters in each group in order to define a common metric for the item parameters. Several methods have previously been suggested to determine which items should be included in the restriction. These items are termed the anchor items. This chapter proposes a conceptual framework for categorizing the anchor methods: The anchor class to describe characteristics of the anchor methods and the anchor selection strategy to guide how the anchor items are determined.*

***Keywords:*** *Rasch model, anchoring, anchor selection, contamination, item response theory (IRT), differential item functioning (DIF), DIF analysis, item bias*

## 5.1 Introduction

The analysis of differential item functioning (DIF) in item response theory (IRT) research investigates the violation of the invariant measurement property among subgroups of examinees, such as male and female test-takers. Invariant item parameters are necessary to assess ability differences between groups in an objective, fair way. If the invariance assumption is violated, different item characteristic curves occur in subgroups. In this chapter, we focus on *uniform* DIF where one group has a higher probability of solving an item (given the latent trait) over the entire latent continuum and the group differences in the logit remain constant (Mellenbergh, 1982; Swaminathan and Rogers, 1990).

A variety of testing procedures for DIF on the item-level is available (as was discussed in Section 3.3, see, e.g., Lord 1980; Mellenbergh 1982; Holland and Thayer 1988; Thissen *et al.* 1988; Swaminathan and Rogers 1990; Shealy and Stout 1993a, for an overview see Millsap and Everson 1993). In the analysis of DIF using IRT, item parameters are to be compared across groups. Mostly, research focuses on the comparison of two pre-defined groups, the reference and the focal group. Thus, a common scale for the item parameters of both groups is required to assess meaningful differences in the item parameters. The minimum (necessary but not sufficient) requirement for the construction of a common scale in the Rasch model is to place the same restriction on the item parameters in both groups (Glas and Verhelst, 1995). The items included in the restriction are termed *anchor items*.

An anchor method determines how many items are used as anchor items and how they are located. Consistent with the literature, we use the term *locate* as a synonym for selecting anchor items. The choice of the anchor items has a high impact on the results of the DIF analysis: If the anchor includes one or more items with DIF, the anchor is referred to as *contaminated*. In this case, the scales may be biased and items that are truly free of DIF may appear to have DIF. Therefore, the false alarm rate may be seriously inflated – in the worst case all DIF-free items seem to display DIF (Wang, 2004) – and the results of the DIF analysis are doubtful, as various examples demonstrate (see Section 6.2).

In the next section, technical details of the anchor process for the Rasch model are explained and illustrated by means of an instructive example and the framework of anchor classes, anchor selection strategies and anchor methods is introduced in detail.

## 5.2 The anchor process for the Rasch model

In the following, the anchor process is technically described and analyzed for the Rasch model (see Section 3.1). The item parameter vector is $\beta = (\beta_1, \dots, \beta_k)^\top \in \mathbb{R}^k$, where $k$ denotes the number of items in the test. In the following, it is estimated using the conditional maximum likelihood (CML) estimation due to its unique statistical properties, its widespread application (Wang, 2004) and the fact that its estimation process does not rely on the person parameters (Molenaar, 1995a).

### 5.2.1 Scale indeterminacy

As the origin of the scale in the Rasch model can be arbitrarily chosen (Fischer, 1995) – what is often referred to as *scale indeterminacy* – one linear restriction of the form

$$\sum_{\ell=1}^{k} d_\ell \tilde{\beta}_\ell = 0\,, \tag{17}$$

with constants $d_\ell$ holding $\sum_{\ell=1}^{k} d_\ell \neq 0$ is placed on the item parameter estimates $\tilde{\beta}_\ell$ (Eggen and Verhelst, 2006). Thus, in the Rasch model only $k-1$ parameters are free to vary and one parameter is determined by the restriction. Note that equation 17 includes various commonly used restrictions such as setting one item parameter $\tilde{\beta}_\ell = 0$ or restricting all item parameters to sum zero $\sum_{\ell=1}^{k} \tilde{\beta}_\ell = 0$ (Eggen and Verhelst, 2006). Without loss of generality, we here estimate the item parameter vector $\tilde{\beta}$ with the employed restriction $\tilde{\beta}_1 = 0$. The corresponding covariance matrix $\widehat{\mathrm{Var}}(\tilde{\beta})$ then contains zero entries in the first row and in the first column. In the following, we discuss different restrictions for which the sum of the estimated item parameters of a selection of items is set to zero. These restrictions can be obtained by transformation using the equations

$$\begin{aligned} \hat{\beta} &= A\tilde{\beta} && (18)\\ \text{and}\quad \widehat{\mathrm{Var}}(\hat{\beta}) &= A\widehat{\mathrm{Var}}(\tilde{\beta})A^\top, && (19) \end{aligned}$$

where $A = I_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot a^\top$, $I_k$ denotes the identity matrix, $1_k$ denotes a vector of one entries and $a$ is a vector with one entries for those elements $a_\ell$ that are included in the restriction and zero entries otherwise (e.g., $a = (1, 0, 1, 0, 0, \dots)^\top$ including item 1 and item 3). Additionally, the entries of the rank deficient covariance matrix $\widehat{\mathrm{Var}}(\hat{\beta})$ in the row and in the column of the item that is first included in the restriction are set to zero.

While for the estimation itself, the choice of the restriction is arbitrary, for the anchor process a careful consideration of the linear restriction that is now employed in each group $g$ is necessary. A necessary but not sufficient requirement in order to build a common scale for the item parameters of two groups is that the same restriction is employed in both groups (Glas and Verhelst, 1995). Items in the restriction are termed *anchor items* and the restriction can be rewritten as

$$\sum_{\ell=1}^{k} a_\ell \hat{\beta}_\ell^g = \sum_{\ell \in \mathcal{A}} \hat{\beta}_\ell^g = 0, \tag{20}$$

where the set $\mathcal{A}$ is termed the *set of anchor items* or the *anchor*. The estimated and anchored item parameters are denoted $\hat{\beta}^g$. Equation 20 includes various commonly used anchor methods such as setting one estimated item parameter $\hat{\beta}_\ell^g$ to zero ($\hat{\beta}_\ell^g = 0$, for one $\ell \in \{1, 2, \ldots, k\}$) for the so called constant single-anchor method or restricting all items except the studied item $j$ to sum to zero in each group ($\sum_{\ell \neq j} \hat{\beta}_\ell^g = 0$) for the so called all-other anchor method. The item parameters, estimated separately in each group, are transformed to the respective anchor method by using equation 18, where all items are then shifted on the scale by $-\frac{1}{|\mathcal{A}|} \sum_{\ell \in \mathcal{A}} \tilde{\beta}_\ell^g$. The covariance matrices are transformed using equation 19.

### 5.2.2 Item-wise Wald test

As a statistical test for DIF, we will focus on the item-wise Wald test here (see, e.g., Glas and Verhelst, 1995). For this, the item parameters are estimated separately in each group using CML estimation. The anchor methods are then employed to obtain a common scale for the item parameters by using equation 18 and 19.

Lord (1980) had originally suggested the usage of the Wald test in combination with the joint estimation of item and person parameters, a proceeding that was found to yield inflated false alarm rates (McLaughlin and Drasgow, 1987; Lim and Drasgow, 1990). To make sure that the Wald test in combination with the consistent CML estimation of the Rasch model parameters (e.g., Molenaar, 1995a) is not affected by an inflated false alarm rate, we conducted a preliminary simulation study (the results are listed in the Appendix of this chapter). In this study, the Wald test was conducted using either one or four DIF-free items as anchor items. The Wald test displayed no inflated false alarm rate for any of the settings that were also included in our main simulation study (see Section 6.4), even if high proportions of DIF-items favored one group and ability differences were present. It is, thus, an appropriate test to compare the performance of different anchor methods for the Rasch model. In case other IRT models are regarded, the recent work of Woods *et al.* (2013) shows that an improved version of the Wald test, termed Wald-1 (see Paek and Han, 2013, and the references therein), also displays well-controlled false alarm rates if the anchor items are DIF-free. Since the Wald-1 test also requires anchor items, it can in principle be combined with the anchor methods discussed in this thesis as well.

The rationale behind the Wald test is that DIF is present if the item difficulties are not equal across groups. The test statistic $T_j$ for the null hypothesis $\text{H}_0 : \beta_j^{\text{ref}} = \beta_j^{\text{foc}}$, where $\beta_j^{\text{ref}}$ and $\beta_j^{\text{foc}}$ denote the item difficulties for reference and focal group for item $j$ and $\hat{\beta}_j^{\text{ref}}$ and $\hat{\beta}_j^{\text{foc}}$ the corresponding estimated item parameters using the anchor $\mathcal{A}^j$, has the following form:

$$T_j = \frac{\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}})}} = \frac{\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}^{\text{ref}})_{j,j} + \widehat{\text{Var}}(\hat{\beta}^{\text{foc}})_{j,j}}}. \tag{21}$$

Note that, the estimated and anchored item parameters $\hat{\beta}^g = \hat{\beta}^g(\mathcal{A}^j)$, which can be calculated using equation 18, depend on the anchor and, hence, so does the test statistic $T_j = T_j(\mathcal{A}^j)$. The anchor set $\mathcal{A}^j$ may depend on the studied item (as is the case for the all-other method). If the anchor is constant regardless which item is tested for DIF, it is denoted $\mathcal{A}$ in the following.

### 5.2.3 Illustration of artificial DIF

Now, we illustrate the requirement that the anchor items should be DIF-free by means of an instructive example. The data set from a general knowledge quiz was conducted by the weekly German news magazine SPIEGEL in 2009. A thorough discussion and analysis of the original data set are provided in Trepte and Verbeet (2010) including a global DIF analysis by means of model-based recursive partitioning by Strobl *et al.* (2010). From about 700,000 test-takers that answered each a total of 45 items from different domains, we select a subsample of 9,442 test-takers (that obtained their A-levels in Germany) and four items from politics (listed below together with the correct answers) for the illustration of the anchor problem:

✎ Item 1 Who determines the rules of action in politics according to the German Constitution? (The Bundeskanzler.)

Item 2 What is the role of the second vote in the elections for the German Bundestag? (It governs the seating in the German Bundestag.)

Item 3 How many people were killed by the RAF? (33)

Item 4 Indicate the location of Hessen on the German map.

As an exemplary illustration, let us suppose we want to test for DIF in the first item between the focal (foc) group of the test-takers that obtained their A-levels in the German federal state Hessen and the reference (ref) group of all remaining test-takers. Figure 18 displays three different restrictions: The second item as constant single-anchor, the fourth item as constant single-anchor and all other items (item 2 to item 4) as anchor. The points represent the estimated item parameters from the reference (light points) and the focal group (dark points). The rectangles surround the anchor item(s).

In Figure 18 (left), item 2 is used as constant single-anchor and, thus, both estimated item parameters are set to zero. The negligible difference in the item parameters of item 1, that we are currently interested in, suggests no DIF in this item. The item-wise Wald test (see equation 21) for item 1 does not display statistically significant DIF ($t = -.968$ with the corresponding p-value of .333). As a result, item 1 is classified as DIF-free. To understand the DIF test results for item 1 in the next scenarios, it is also important to note that the large difference in item 4 implies DIF in this item. Since item 4 was the question to indicate the location of Hessen on the German map, it is plausible that this item 4 is a true DIF-item since it was easier for test-takers that obtained their A-levels in Hessen.

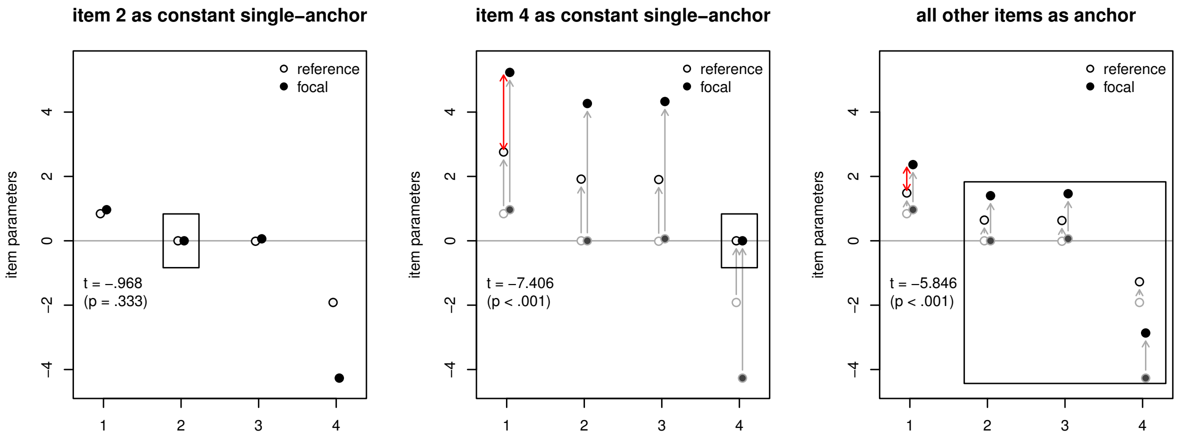

**Figure 18** – *Different restrictions placed on the item parameters that are estimated using the Rasch model in each group.*

In the next scenario in Figure 18 (middle), item 4 (that we found plausible to have true DIF) is used as a constant single-anchor. Compared to the first scenario, all item parameters are now shifted upwards by the estimated difficulties of item 4 and artificial differences occur for item 1, 2 and 3. This shows that the anchor items should be DIF-free to avoid the artificial differences, which are termed artificial DIF by Andrich and Hagquist (2012). The artificial DIF for item 1, that we are currently interested in, is statistically significant ($t = -7.406$ with the corresponding p-value < .001). Hence, item 1 is classified as a DIF-item.

In the last scenario in Figure 18 (right), all other items – except the currently studied item 1 – are used as anchor items. Compared to the second scenario, the scales are shifted apart less strongly since the scale shift is reduced from the estimated difficulties of the DIF-item 4 to the average over the estimated difficulties of item 2, 3 and 4 (including the apparently DIF-free items 2 and 3) as visible by the shorter arrows. However, the statistical test still classifies item 1 as a DIF-item ($t = -5.846$ with the corresponding p-value < .001). This example illustrates the major impact of the anchor method on the results of the DIF analysis, since – depending on the anchor set – three different test statistics result in the DIF tests for item 1.

From a theoretical perspective and from our instructive example, it is obvious that an appropriate anchor is crucial for the results of the DIF analysis. Previous simulation studies have compared different selections of anchor methods (an overview over the existing literature will be given in Section 6.3). Empirical findings also show that, ideally, the anchor items should be DIF-free. Unfortunately, since prior to DIF analysis it cannot be known which items are DIF-free, we face a somewhat circular problem, as pointed out by Shih and Wang (2009). If DIF-items are included in the anchor, this *contamination* may lead to seriously inflated false alarm rates in DIF detection (see, e.g., Wang and Yeh, 2003; Wang, 2004; Wang and Su, 2004; Finch, 2005; Stark *et al.*, 2006; Woods, 2009) that "*can result in the inefficient use of testing resources, and [...] may interfere with the study of the underlying causes of DIF*" (Jodoin and Gierl, 2001, p. 329). Naturally, the risk of contamination would suggest to use only few items in the restriction (i.e. a short anchor), but the simulation results also show that the statistical power increases with the length of a DIF-free anchor (Thissen *et al.*, 1988; Wang and Yeh, 2003; Wang, 2004; Shih and Wang, 2009; Woods, 2009).

## 5.3 A conceptual framework for anchor methods

The new conceptual framework distinguishes between the *anchor class* and the *anchor selection strategy*. Firstly, the *anchor class* describes the pre-specification of the anchor characteristics (such as a pre-defined anchor length). We review the all-other (used, e.g., by Cohen *et al.* 1996; Kim and Cohen 1998), the constant anchor (used, e.g., by Thissen *et al.* 1988; Wang 2004; Shih and Wang 2009) and the iterative anchor class (referred to here as iterative backward and used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002). Secondly, the *anchor selection strategy* determines which items are chosen as anchor items.

### 5.3.1 Anchor classes

In our conceptual framework *anchor classes* describe characteristics of the anchor that answer the following questions: Is the anchor length pre-defined? If so, how many items are included in the anchor? Is the anchor determined by the anchor class itself or is an additional anchor selection strategy necessary? Are iterative steps intended to define the anchor?

*The equal-mean and the all-other anchor class*

In the *equal-mean-difficulty* anchor class (see, e.g., Wang, 2004, and the references therein) all items are restricted to have the same mean difficulty (typically zero) in both groups, whereas in the *all-other* anchor class (used, e.g., by Cohen *et al.* 1996; Kim and Cohen 1998) the sum of all items – except the item currently tested for DIF – is restricted to be zero and the anchor set $\mathcal{A}^{\bar{j}} = \{1, \ldots, k\} \setminus j$ depends on the studied item $j = 1, \ldots, k$.

Both anchor classes have a pre-defined anchor length but no additional anchor selection is necessary as the items included in the restriction are already determined by the anchor class itself. The equal-mean-difficulty and the all-other class only differ in one anchor item and, therefore, essentially lead to similar results (cf. Wang, 2004) and, hence, only the all-other method is included in the following parts of this thesis.

*The constant anchor class*

The *constant* anchor class (used, e.g., by Thissen *et al.* 1988; Wang 2004; Shih and Wang 2009) includes a pre-defined number of the items (e.g., 1 or 4 items according to Thissen *et al.*, 1988) or a certain proportion of the items (e.g., 10% or 20% according to Woods, 2009) as anchor.

The term *constant* reflects the pre-defined, constant anchor length. The constant anchor class needs to be combined with an explicit anchor selection strategy. For the constant single anchor class, the first item of the ranking order of candidate anchor items is used as anchor, whereas for the constant four anchor class, the first four items of the ranking order of candidate anchor items are used as anchor.

*The iterative backward anchor class*

The *iterative backward* anchor class (used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002) includes a variety of iterative methods that have been suggested, discussed and combined with different statistical methods to assess DIF. Here, we focus on the commonly used re-linking procedure where one parameter estimation step suffices to conduct DIF analysis. Firstly, the scales of both groups are linked on (approximately) the same metric, e.g., by using the all-other anchor method. Then, the DIF-items are excluded from the current anchor, the scales are re-linked using the new current anchor, the DIF analysis is carried out and the steps are repeated until two steps reach the same results (e.g., Drasgow, 1987; Candell and Drasgow, 1988; Park and Lautenschlager, 1990; Kim and Cohen, 1995; Hidalgo-Montesinos and Lopez-Pina, 2002). This iterative procedure is referred to here as the *iterative backward* anchor class, since the method includes the majority of items in the anchor at the beginning. Then, it successively excludes items from the anchor.

A new anchor class will be suggested in Chapter 6 and systematically compared to the commonly used anchor classes presented above.

### 5.3.2 Anchor selection strategies

The anchor selection strategies discussed in this thesis are based on preliminary item analyses. This means that – before the final DIF test is done – DIF tests are conducted to locate (ideally) DIF-free anchor items. The (non-statistical) alternative relying on expert advice and certain prior knowledge of DIF-free anchor items (Wang, 2004; Woods, 2009) will not often be possible in practice (for a literature overview where this approach fails see Frederickx *et al.*, 2010).

*The all-other anchor selection*

An example of an anchor selection strategy is the rank-based strategy proposed by Woods (2009) that we term all-other (AO) anchor selection strategy. Initially, every item is tested for DIF using all other items as anchor. The ranking order of candidate anchor items is defined according to the lowest rank(s) of the resulting (absolute) DIF test statistics.

*Other suggestions*

Ideas on how to select anchor items without prior knowledge were also given by Wang (2004). His original suggestions and also an anchor selection strategy that simplifies his approach will be presented in Chapter 6. Another suggestion on how to locate anchor items by Shih and Wang (2009) will be discussed in Chapter 7 together with three new anchor selection strategies.

### 5.3.3 Anchor methods

An *anchor method* results as a combination of an anchor class with an anchor selection strategy (in cases where the latter is necessary). For example, the explicit anchor selection is necessary for the constant anchor class: Firstly, the anchor selection is carried out to determine a ranking order of candidate anchor items and the procedure defined by the anchor class is carried out to determine the final anchor. Secondly, the final anchor found in the first step is then used for the assessment of DIF. This procedure was termed DIF-free-then-DIF strategy by Wang *et al.* (2012).

Anchor classes that do not require an anchor selection strategy are the equal-mean, the all-other and the iterative backward anchor class. The next two chapters will illustrate that anchor methods that do not rely on an explicit anchor selection strategy are inadvisable for DIF analysis.

A variety of anchor methods – including the all-other, the constant, the iterative-backward and the newly suggested iterative forward anchor classes together with two anchor selection strategies – will be compared in the next chapter.

## 5.4 Appendix: Preliminary simulation study

To make sure the Wald test is appropriate for DIF detection, we conducted a preliminary simulation study with the Wald test based on a per definition DIF-free anchor. We implemented methods from the constant anchor class consisting of either one (termed *single-pure*) or four DIF-free items (termed *four-pure*). The settings are the same as for the simulation study in the next chapter where they are described in more detail (see Section 6.4).

- *Data generating process*

  The person parameters are generated from a normal ability distribution with a higher mean for the reference group (0.5) than for the focal group (0) and a variance of 1. In the following simulation studies, ability differences are simulated since this case is often found more challenging for the methods compared to a situation where no ability differences are present (see, e.g., Penfield, 2001). The item parameters are set to the replicated values $\beta_j^{\text{ref}} \in \{-1, -0.5, 0, 0.5, 1\}$. In case of DIF the difference in the item parameters is set to $\Delta_{\text{DIF}} = \beta_j^{\text{ref}} - \beta_j^{\text{foc}}$ to $+.5$ or $-.5$ (consistent with the intended direction of DIF). The responses in each group follow the Rasch model.

- *Sample sizes, directions and proportions of DIF*

  The sample sizes in reference and focal group are defined by the following pairs $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250, 250),\ (500, 250),\ (500, 500),\ (750, 500),\ (750, 750),\ \ldots,\ (1500, 1500)\}$ and represent both equal and different group sizes. In case of *balanced DIF*, each DIF-item favors either the reference or the focal group but no systematic advantage for one group remains because the effects cancel out. In case of *unbalanced DIF*, a systematic disadvantage for the reference group is generated such that every DIF-item favors the focal group. The proportion of DIF-items is set to $p \in \{15\%, 30\%, 45\%\}$.

- *Outcome variables*

  To evaluate whether the Wald test works appropriately, the *false alarm rate* of DIF-free items displaying DIF and the *hit rate* of correctly detecting DIF-items are calculated (for more details see Chapter 6).

The results showed – as already stated above – that the Wald test was not affected by an inflated false alarm rate. Figure 19 contains the false alarm rates and the hit rates for the balanced DIF condition, i.e. all DIF-effects canceled out and no group had an unfair advantage in the test. Figure 20 depicts both rates for the unbalanced DIF condition, i.e. all DIF-items favored the focal group. As can be seen from the top rows, the Wald test held the alpha level. Furthermore, the patterns of the hit rates in the bottom rows show that the power of the statistical test increased with the length of the DIF-free anchor since the constant anchor consisting of only one item resulted in a lower hit rate compared to the anchor including four anchor items.

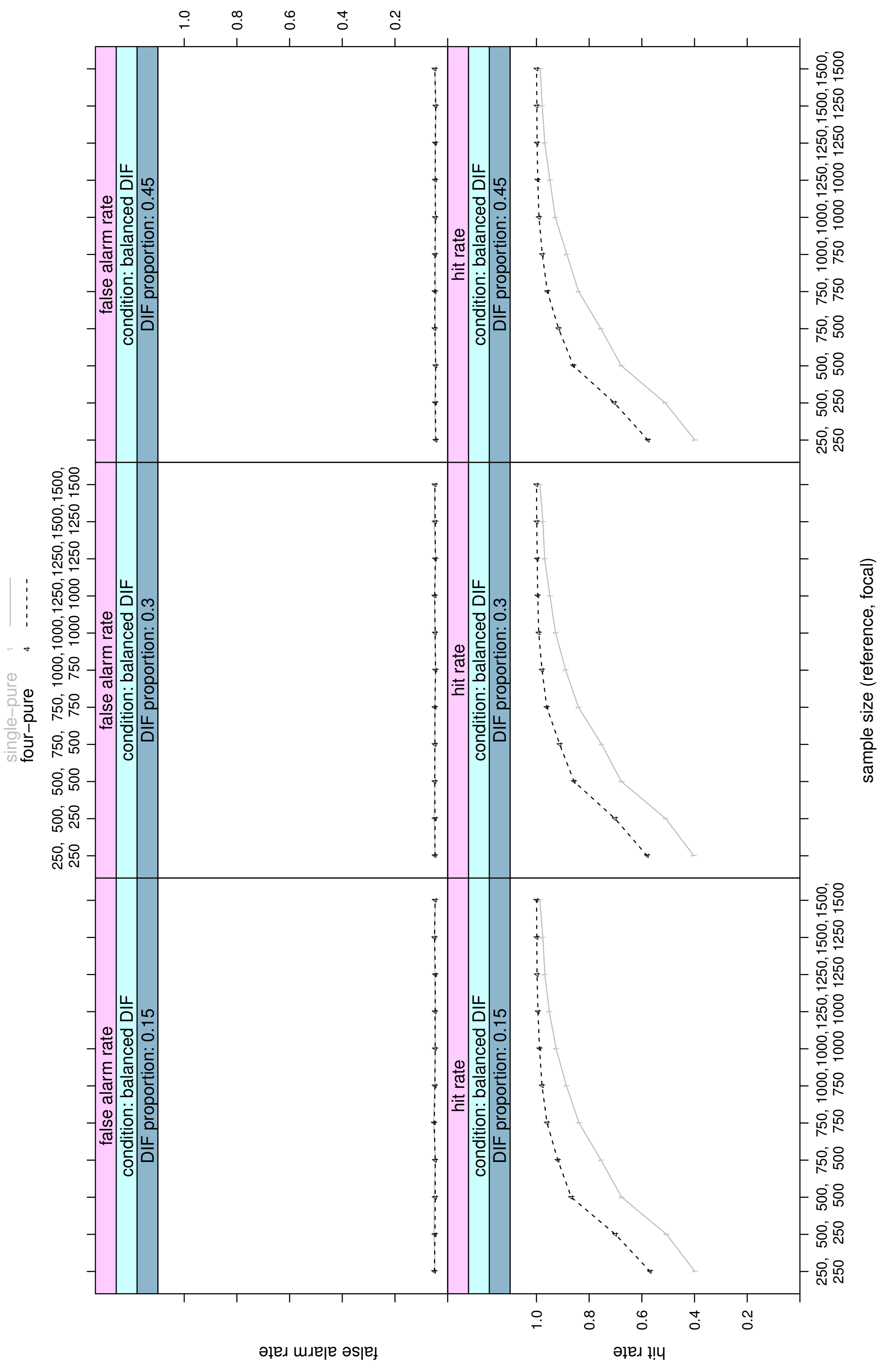

**Figure 19 –** *Balanced condition:* 15%, 30% *and* 45% *DIF-items with no systematic advantage for one group; sample size varied from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the balanced condition.*

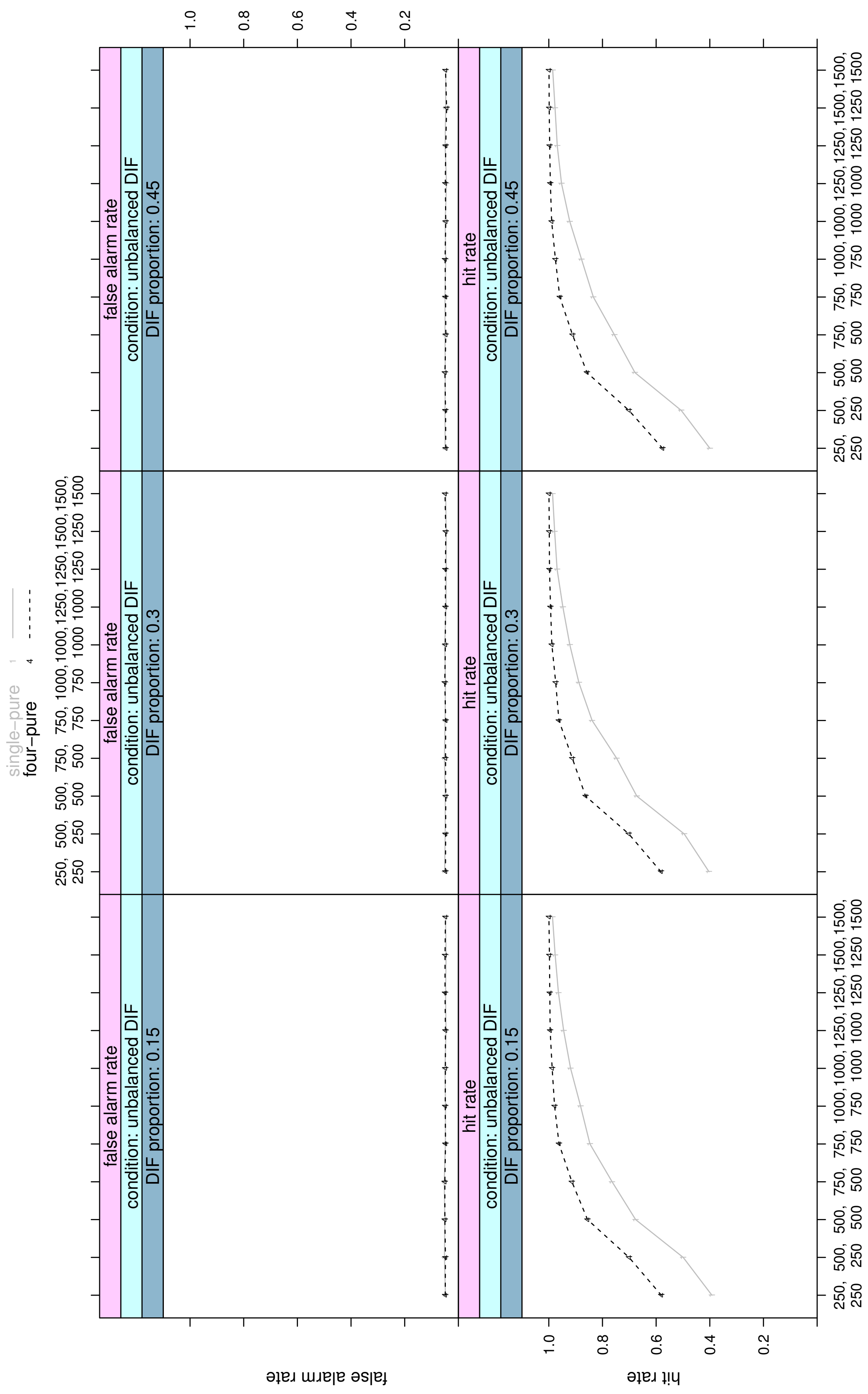

**Figure 20 –** *Unbalanced condition:* 15%, 30% *and* 45% *DIF-items favoring the reference group; sample size varied from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the unbalanced condition.*

# 6 Anchor classes for DIF detection in the Rasch model

***Summary:*** *A variety of anchor methods has recently been proposed for the analysis of differential item functioning (DIF). The choice of the anchor method is crucial for suitable DIF analysis because it can severely affect the results of the DIF analysis. If the anchor method does not work appropriately, items truely free of DIF appear to have DIF and, thus, inflated false alarm rates may result. The question how anchor items are selected appropriately is still a major challenge. In this chapter, a new anchor class termed the iterative forward anchor class is proposed. Several anchor classes are implemented with two different anchor selection strategies (the all-other and the number-of-significant anchor selection strategy) and are compared in an extensive simulation study. The results show that the newly proposed anchor class combined with the number-of-significant selection strategy is superior in situations where no prior knowledge about the direction of DIF is available. Moreover, it is shown that the proportion of DIF-items in the anchor – rather than the fact whether the anchor includes DIF-items at all (termed contamination in previous studies) – is crucial for suitable DIF analysis.*

***Keywords:*** *Rasch model, anchoring, anchor selection, contamination, item response theory (IRT), differential item functioning (DIF), DIF analysis, item bias*

## 6.1 Introduction

The previous chapter showed that the results of the DIF analysis strongly depend on the anchor items. Still, even though the importance of the anchor method is undeniable, Lopez Rivas *et al.* (2009, p. 252) claim that “[a]*t this point, little evidence is available to guide applied researchers through the process of choosing anchor items*”. Consequently, the aim of this chapter is to provide guidelines how to choose an appropriate anchor for DIF analysis in the Rasch model.

In the interest of clarity, we use the new conceptual framework presented in the previous chapter that distinguishes between the *anchor class* and the *anchor selection strategy*. Existing anchor classes are reviewed and a new anchor class named the iterative forward anchor class is introduced.

To derive guidelines which anchor method is appropriate for DIF detection in the Rasch model, we conduct an extensive simulation study. In our study, we compare the all-other, the constant, the iterative backward and the newly suggested iterative forward anchor class for the first time. Furthermore, our study is to our knowledge the first to systematically contrast different anchor selection strategies that are combined with the anchor classes. We discuss the all-other anchor selection strategy introduced by Woods (2009) and the number-of-significant anchor selection strategy (based on a suggestion by Wang 2004). Altogether, nine different anchor methods are evaluated regarding their appropriateness for DIF analysis. Finally, practical recommendations are given to facilitate the process of selecting anchor items for DIF analysis in the Rasch model.

In the next section, the new iterative forward procedure is introduced. The simulation study is presented in Section 6.4 and the results are discussed in Section 6.5. The problem of contamination and its impact on DIF analysis are addressed in Section 6.6. Characteristics of the selected anchor items are discussed in Section 6.7. A concluding summary of the simulation results and practical recommendations are given in Section 6.8.

## 6.2 The iterative forward class and comparison methods

In the interest of clarity, we use the classification of anchor methods introduced in the previous chapter that also includes the introduction of several previously suggested anchor classes.

### 6.2.1 The iterative forward anchor class and comparison anchor classes

The example in the previous chapter showed that the results of the DIF analysis strongly depend on the anchor method (see Section 5.2). Furthermore, the research of Wang and Yeh (2003), Wang (2004), Shih and Wang (2009) and Wang *et al.* (2012) made clear that the direction of DIF influences the results of the DIF analysis using all other items as anchor: If all items favor one group, what is referred to as *unbalanced* DIF, tests using all other items as anchor result in inflated false alarm rates. Hence, in complex DIF situations such as unbalanced DIF, the initial step of the iterative-backward anchor class, that includes all other items as anchor, may lead to biased test results.

*The iterative forward anchor class*

Inspired by the results of Wang and Yeh (2003), Wang (2004), Shih and Wang (2009) and Wang *et al.* (2012), we introduce another possible strategy to overcome the problem that the anchor selection is based on initially biased test results: the *iterative forward* anchor class. As opposed to the iterative backward class, we suggest to build the iterative anchor in a step-by-step forward procedure. Starting with the first candidate anchor item – determined by the anchor selection strategy – as single anchor item, we link the scales and estimate DIF. Then, iteratively, one item – located again by means of the respective anchor selection strategy – is added to the current anchor and DIF analysis is conducted using the new current anchor. These steps are repeated as long as the current anchor length is shorter than the number of non-significant test results in the current DIF test (in short the number of currently presumed DIF-free items). Unlike the iterative backward anchor class where items are successively excluded, now items are successively included in the anchor. An anchor selection strategy is again needed to guide which items are included in the anchor.

The rationale behind the *iterative forward* approach is similar to the common approach in confirmatory factor analysis (CFA) described by Stark *et al.* (2006), where one referent item is constrained to conduct DIF analysis. Here, we determine an order of candidate anchor items and guide the decision whether additional anchor items are included by DIF tests for all items (except the first anchor item, see Section 6.2.3) using the current anchor.

*Anchor classes for comparison*

To allow for a comparison, we include several previously suggested anchor classes in our simulation study. Since the equal-mean-difficulty and the all-other class, essentially lead to similar results (cf. Wang, 2004) we include only the all-other class in the following simulation study. We also implemented the constant anchor class with one single anchor item as well as the constant anchor including four items, which is supposed to assure sufficient power (cf. e.g., Shih and Wang, 2009; Wang *et al.*, 2012). The iterative backward anchor class that firstly uses all-other items as anchor and than systematically excludes items from the anchor (see again Section 5.3.1) is also included in the simulation study.

### 6.2.2 Anchor selection strategies

In our simulation study, we implemented different anchor selection strategies that provide a ranking order of candidate anchor items. One anchor selection strategy investigated in this chapter is the all-other (AO) strategy proposed by Woods (2009) that is often termed the rank-based approach (see again Section 5.3.2).

Originally, Wang (2004) suggested an anchor method that we refer to as the next candidate (NC) method. It includes both an anchor selection and an anchor class and is, thus, discussed in detail in the next section. Here, we simplify the suggestion of Wang (2004) for the anchor selection, that is applied in a modified version for the MIMIC procedure by Shih and Wang (2009), and call it the number-of-significant (NST) selection strategy. It is, to our knowledge, for the first time systematically compared with the all-other strategy using various anchor classes. With every item acting as single-anchor, every other item is tested for DIF. Again, the anchor sets $\mathcal{A}^j$ vary across the studied items and $k-1$ tests result for every item $j = 1, \ldots, k$ of the test. The ranking order of candidate anchor items is defined according to the smallest number of significant results. If more than one item displays the same number of significant results, one of the corresponding items is selected randomly. A similar approach was suggested by Cheung and Rensvold (1999) in the context of factorial invariance to concurrently select a set of non-variant items that also relies on results testing every item by using every other item in the restriction.

### 6.2.3 Anchor methods

The anchor methods to be investigated in this chapter are now presented and summarized in Table 5.

*The all-other anchor method*

The all-other anchor method (*all-other*) does not require an additional anchor selection. Every item is tested for DIF using all remaining items as anchor items. Here, the anchor set $\mathcal{A}^{\bar{j}} = \{1, \ldots, k\} \setminus j$ depends on the studied item $j = 1, \ldots, k$ and one re-linking step is necessary for each item.

As already stated in Chapter 5 all remaining anchor methods consist of two steps: Firstly, the anchor selection is carried out to determine a ranking order of candidate anchor items and the proceeding of the anchor class is carried out to determine the final anchor. Secondly, the final anchor found in the first step is then used for the assessment of DIF. This procedure was termed DIF-free-then-DIF strategy by Wang *et al.* (2012). For the following anchor methods,

| Class | Selection | Combination | Initial step and anchor selection strategy |
|---|---|---|---|
| **all-other** | none | all-other<br>cf. e.g., Woods (2009) | Initial step: Each item is tested for DIF using all remaining items as anchor.<br>Selection strategy: No additional selection strategy is required. |
| **constant** | AO | single-anchor-AO<br>Woods (2009) | Initial step: Each item is tested for DIF using all remaining items as anchor.<br>Selection strategy: The item with the lowest absolute DIF statistic (AO) is chosen. |
| | NST | single-anchor-NST<br>Wang (2004) | Initial step: Each item is tested for DIF using every other item as single-anchor.<br>Selection strategy: The item with the smallest number of significant DIF tests (NST) is chosen. |
| | AO | four-anchor-AO<br>Woods (2009); Wang *et al.* (2012) | Initial step: Each item is tested for DIF using all remaining items as anchor.<br>Selection strategy: The four anchor items corresponding to the lowest ranks of the absolute DIF statistics from the initial step (AO) are chosen. |
| | NST | four-anchor-NST<br>Wang (2004) | Initial step: Each item is tested for DIF using every other item as single-anchor.<br>Selection strategy: The four anchor items corresponding to the smallest number of significant DIF tests (NST) are chosen. |
| | NC | four-anchor-NC<br>proposed by Wang (2004) | Initial step: Each item is tested for DIF using every other item as single-anchor.<br>Selection strategy: The first anchor is found as in single-anchor-NST; the next candidate anchor item (up to three) is found from tests using the current anchor if its result corresponds to the lowest non-significant absolute test statistic and is then added to the current anchor. |
| **iterative backward** | AO | iterative-backward-AO<br>e.g., Drasgow (1987) | Initial step: Each item is tested for DIF using all remaining items as anchor.<br>Selection strategy: Iteratively, all items displaying DIF are excluded from the anchor and the next DIF test with the current anchor is conducted. |
| **iterative forward** | AO | iterative-forward-AO | Initial step: Each item is tested for DIF using all remaining items as anchor.<br>Selection strategy: As long as the current anchor is shorter than the number of currently presumed DIF-free items, the next item with the lowest rank in the initial step (AO) is added to the anchor. |
| | NST | iterative-forward-NST | Initial step: Each item is tested for DIF using every other item as single-anchor.<br>Selection strategy: As long as the current anchor is shorter than the number of currently presumed DIF-free items, the next item with the smallest number of significant test results in the initial step (NST) is added to the anchor. |

**Table 5** – *Classification and nomenclature of the investigated anchor methods.*

the final anchor $\mathcal{A}$ is independent of which item is studied. Since $k - 1$ parameters are free in the estimation, only $k - 1$ estimated standard errors result (Molenaar, 1995a), the k-th standard error is determined by the restriction and, hence, only $k - 1$ tests can be carried out and one item in the final assessment of DIF obtains no DIF test statistic. Thus, for the following anchor methods, the first item selected as anchor item is presumed DIF-free in the final DIF test and all remaining items are tested for DIF using the final anchor $\mathcal{A}$.

### *Methods form the constant anchor class*

The constant anchor class consisting of one anchor item or four anchor items can be combined with the all-other selection strategy (*single-anchor-AO*, *four-anchor-AO*). The anchor is selected according to the lowest rank(s) of the DIF test statistics resulting from the initial DIF tests for every item using the all-other method (Woods, 2009). The final DIF test is conducted after re-linking the parameters using the final anchor.

The constant anchor class can also be combined with the NST-selection strategy (*single-anchor-NST*, *four-anchor-NST*). In this case, initially every item is tested for DIF by using every other item as single-anchor. The final anchor (consisting of either one item for the single-anchor-NST or four items for the four-anchor-NST method) is selected from the set of items displaying the smallest number of significant test results.

Furthermore, we implemented a suggestion of Wang (2004) that we refer to as the four-anchor next candidate (NC) method. In the *four-anchor-NC* method, the item that is selected by the NST-selection strategy functions as the current single-anchor and DIF tests are conducted (see Wang, 2004, p. 249) similar to the single-anchor-NST method. In this step, one DIF test statistic results for every item except for the anchor. The next candidate anchor item is the item that displays "*the least magnitude of DIF*" (Wang, 2004, p. 250) among all remaining items that we defined as lowest absolute DIF test statistic from the tests using the current single-anchor item. The candidate item is added to the current anchor only if its DIF test result is not significant (Wang, 2004).[9] The next DIF test is conducted using the new current anchor and the next candidate item is selected again if it has the lowest absolute DIF test statistic among all remaining items and displays no significant DIF. These steps are repeated until either the next candidate anchor item displays DIF or the maximum anchor length (of four items in our implementation) is reached.

Technically speaking, this procedure is a combination of the constant and the iterative anchor class because it allows a varying anchor length but its length is limited to a pre-specified number of items. However, since in our simulation always four anchor items were selected for the final anchor, here we classify the anchor class as constant.

### *The iterative backward method*

The iterative backward class is implemented using all-other items as anchor in the initial step and then excluding DIF-items from the anchor (*iterative-backward-AO*) as it is widely used

[9] We employed a significance level of .05, but future research may investigate the additional proposition of Wang (2004) augmenting it to e.g., .30.

in practice (e.g., Bolt *et al.*, 2004; Edelen *et al.*, 2006). When items are excluded from the anchor, the scales are re-linked with the new current anchor and DIF tests are conducted. Items displaying DIF are again excluded from the anchor and the steps are repeated until the final anchor is found.[10] Note that the iterative backward class is not combined with the NST-selection since the latter provides only a ranking order of candidate anchor items, but no information which set of items should be used in the initial step.

*Methods form the iterative forward anchor class*

The newly suggested iterative forward class can be combined with the all-other anchor selection strategy (*iterative-forward-AO*). In this case, DIF analysis for every item is conducted using the all-other method. The DIF test statistics are again ranked by their absolute value. The item corresponding to the lowest DIF test statistic is the first anchor item. As long as the anchor length does not exceed the number of currently presumed DIF-free items, the item with the next lowest absolute DIF test statistic from the initial step is included in the current anchor and DIF analysis is carried out using the new current anchor to check whether the anchor length exceeds the number of presumed DIF-free items. When the final iteration is reached, again, DIF analysis is conducted with the new final anchor, where, again, the first anchor item is presumed DIF-free.

Furthermore, the proposed iterative forward class can be also combined with the NST-selection strategy (*iterative-forward-NST*). In this case, initially every item is tested for DIF using every other item as single-anchor. The items are ranked according to the number of significant test results. The item with the smallest number of significant test results is chosen as anchor. As long as the anchor length does not exceed the number of currently presumed DIF-free items, the next (new) item displaying the smallest number of significant DIF tests from the initial step is included in the anchor and DIF analysis is carried out using the new current anchor to check whether the current anchor length exceeds the number of currently presumed DIF-free items. When the final iteration is reached, the DIF test results of the DIF analysis using the final anchor are returned and the first anchor item is presumed DIF-free.

## 6.3 Background of the simulation study

In the next section, a simulation study is described that investigates the trade-off between the false alarm rate and the hit rate of DIF tests using the anchor methods introduced in the previous section. The results are used to develop guidelines which anchor methods should be used for DIF analysis in the Rasch model.

If no DIF is present in the test, we expect all anchor methods to yield well-controlled false alarm rates, since no DIF-items and, therefore, no risk of contamination exists (Wang and Yeh, 2003; Stark *et al.*, 2006; Woods, 2009; González-Betanzos and Abad, 2012).

If DIF is balanced, i.e. the DIF-items favor either the reference or the focal group and no systematic disadvantage exists, previous simulation studies showed that the all-other class yielded

[10] In case all items were excluded from the anchor (which happened in only 2 out of 154,000 replications), one single anchor item was chosen randomly in our simulation study.

a well-controlled false alarm rate and a high hit rate (Wang and Yeh, 2003; Wang, 2004). However, if DIF is unbalanced i.e. all DIF-items are simulated to favor one group, an inflated false alarm rate for the all-other method was reported (Wang and Yeh, 2003; Wang, 2004).

In accordance with Thissen *et al.* (1988) and Woods (2009), we anticipate the constant anchor class to show an increase in the false alarm and the hit rate when the anchor length rises from one to four items and the proportion of DIF-items is high. Wang *et al.* (2012) also found that four anchor items combined with the IRTLRDIF procedure (Thissen, 2001) yielded low power rates as might also be the case in this simulation with the Wald test.

González-Betanzos and Abad (2012) compared an iterative backward two-step procedure based on the AO-selection strategy to specific constant single-anchors, to a purification procedure based on a DIF-free constant single anchor and to the all-other method. The constant single-anchor items were selected from the set of known a priori DIF-free items. The iterative backward two-step procedure showed slightly inflated false alarm rates. Due to the fact that one additional purification step improved the test results, the authors assumed improvements when further purification steps are added as we have implemented in this chapter. Accordingly, we expect the iterative backward anchor class to achieve high hit rates as they allow for a long anchor, but at the expense of an inflated false alarm rate especially in settings where the proportion of DIF-items is high and DIF is unbalanced.

Little information is available on how well the anchor selection strategies perform, as Wang and Yeh (2003), Wang (2004) and Thissen *et al.* (1988) included only DIF-free items in the constant anchor class. This approach is only possible in simulation studies, however, where it is known by design which items are DIF-free. In practice, on the other hand, a set of DIF-free items prior to DIF analysis is usually not available (González-Betanzos and Abad, 2012). Including only DIF-free items avoids the risk of contamination (for the consequences of contamination see Section 6.2) and, thus, leads to an advantage for the methods from the constant anchor class. However, in order to compare the anchor classes under realistic conditions where it is not known a priori which items are DIF-free, the methods from the constant anchor class should be investigated together with an anchor selection strategy.

Woods (2009) investigated the AO-selection strategy to locate a set of constant anchor items and found results suitable for DIF analysis and superior to the all-other method. However, Wang *et al.* (2012) investigated the constant anchor method based on the selection of four anchor items using the AO-selection strategy (here referred to as the four-anchor-AO method) and found that the anchors were often contaminated and showed an inflated false alarm rate when DIF was unbalanced and no additional purification step was used. Therefore, we expect the four-anchor-AO method to perform well only in the condition of balanced DIF and poorly in the condition where DIF is unbalanced (Wang and Yeh, 2003; Wang, 2004; Shih and Wang, 2009; Wang *et al.*, 2012).

The NST-selection strategy that is a simplification of the proposition by Wang (2004) is (to our knowledge) implemented and combined with several anchor classes here for the first time. Since the NST-selection strategy relies on DIF tests using every item as single anchor, we anticipate the NST-selection strategy to outperform the AO-selection strategy if the sample size is large and DIF is unbalanced. When DIF is balanced, we expect the AO-selection strategy to be superior.

The newly suggested iterative forward class builds the anchor in a step-by-step forward procedure. In comparison with the iterative backward method, we expect the forward procedure to be superior when the NST-selection strategy is used and DIF is unbalanced since the initial step of the iterative backward procedure is built on biased test results. In comparison with methods from the constant anchor class, we anticipate higher hit rates because the anchor of the iterative forward procedure grows as long as the current anchor is shorter than the number of currently presumed DIF-free items and should, thus, include more than four items. As a drawback, we also expect higher false alarm rates since the risk of contamination increases with the anchor length. Furthermore, we anticipate the methods from the iterative forward class to show lower hit rates than the all-other method in the balanced case, because the latter uses all items – except the studied item – as anchor.

## 6.4 Simulation study

To evaluate which of the anchor methods presented in the previous section (for a brief description and nomenclature see again Table 5) are best suited to correctly classify items with and without DIF, an extensive simulation study is conducted. 2000 data sets are generated from each of 77 different simulation settings.

For every data set, the item-wise Wald test (Lord 1980, see Section 6.2) – based on the currently investigated anchor method – is conducted at the significance level of .05 in the free R system for statistical computing (R Development Core Team, 2011). A short description of the study design is given in the following paragraphs. Parts of the simulation design were inspired by the settings used by Woods (2009) and Wang (2004).

### 6.4.1 Data generating process

Each data set, that represents one of 2000 replications from one simulation setting, corresponds to the simulated responses of two groups of subjects (the *reference* (ref) and the *focal* (foc) group) in a test with $k = 40$ items.

- *Person and item parameters*

  The person parameters are generated from a normal ability distribution with a higher mean for the reference group $\theta^{\text{ref}} \sim N(0.5, 1)$ than for the focal group $\theta^{\text{foc}} \sim N(0, 1)$. Values assigned to the item parameters are replicated from the sequence of $\beta_j^{\text{ref}} \in$ {-1, -0.5, 0, 0.5, 1} in equal proportions.

- *DIF-items*

  In case of DIF, the affected DIF-items are chosen randomly to display uniform DIF by setting the difference in the item parameters of reference and focal group $\Delta_{\text{DIF}} = \beta_j^{\text{ref}} - \beta_j^{\text{foc}}$ to +.5 or −.5 (consistent with the intended direction of DIF).

- *IRT model*

  The responses in each group follow the Rasch model. They are generated in two steps: The probability of person $i$ solving item $j$ is computed by placing the corresponding item and person parameters in the Rasch model formula 22. The binary responses are then drawn from a binomial distribution with the resulting probabilities.

$$P(U_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{22}$$

### 6.4.2 Manipulated variables

Three main conditions determine the specification of the manipulated variables: One condition under the null hypothesis where no DIF is present and two conditions under the alternative where DIF is present.

- *Sample sizes*

  The sample sizes in reference and focal group are defined by the following pairs $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250, 250), (500, 250), (500, 500), (750, 500), (750, 750), \ldots, (1500, 1500)\}$. Thus, both equal and different group sizes are considered.

- *Directions and proportions of DIF*

  Under the condition of the null hypothesis (*no DIF*), only the sample sizes are varied. The two remaining conditions represent the alternative hypothesis where DIF is present, but they differ with respect to the direction of DIF: The second condition represents *balanced DIF*. Here, each DIF-item favors either the reference or the focal group but no systematic advantage for one group remains because the effects cancel out. For the third *unbalanced DIF* condition a systematic disadvantage for the reference group is generated such that every DIF-item favors the focal group. In addition to the sample size, also the proportion of DIF is manipulated including the following percentages $p \in \{15\%, 30\%, 45\%\}$.

### 6.4.3 Outcome variables

To allow for a comparison of the anchor methods, the classification accuracy of the DIF tests is evaluated by means of false alarm rate and hit rate.

- *False alarm rate*

  For a single replication the *false alarm rate* is defined as the proportion of DIF-free items that are (erroneously) diagnosed with DIF. The estimated false alarm rate for each experimental setting is computed as the mean over all 2000 replications and, thus, corresponds to the *type I error rate*. Similarly, the standard error is estimated as the square root of the unbiased sample variance over all replications.

- *Hit rate*

  Analogously, for a single replication the *hit rate* is computed as the proportion of DIF-items that are (correctly) diagnosed with DIF. The hit rate is only defined in conditions that include DIF-items, namely in the balanced and unbalanced condition. The estimated hit rate and the standard error are again computed as mean and standard error over all 2000 replications and correspond to the *power* of the statistical test and its variation.

- *Further outcome variables*

  Moreover, the percentage of replications where at least one item in the anchor is a DIF-item (*risk of contamination*) is computed over all replications of one setting. The average proportion of DIF-items as compared to the overall number of anchor items (*degree of contamination*) is computed, too, for replications where the anchor is contaminated. Average false alarm rates are also computed separately for the tests based on a contaminated and for the tests based on a pure (not contaminated) anchor to allow for a more detailed interpretation of the results.

In summary, nine anchor methods are compared for eleven different sample sizes in one condition of the null hypothesis (no DIF) and two conditions of the alternative (balanced and unbalanced DIF) with three different proportions of DIF-items (15%, 30% or 45%). The aim of this simulation study is to assess their appropriateness to identify DIF and DIF-free items and to evaluate the trade-off between the false alarm rate and the hit rate. Altogether, 77 different simulation settings result. Each setting is replicated 2000 times to ensure reliable results.

## 6.5 Results

### 6.5.1 Null hypothesis: No DIF

In the first condition, all items were truly DIF-free. Therefore, only the false alarm rate (proportion of DIF-free items that are diagnosed with DIF) was computed.

*False alarm rates*

The estimated false alarm rates are depicted in Figure 21 and (only for equal sample sizes to save space) also reported together with their standard errors in Table 6 in the Appendix.

As shown in Figure 21, all anchor methods held the 5% level. While methods from the all-other, the iterative backward (iterative-backward-AO) and the iterative forward class (iterative-forward-NST, iterative-forward-AO) were near the significance limit, methods from the constant anchor class (constant single-anchors: single-anchor-AO and single-anchor-NST; constant four-anchors: four-anchor-AO, four-anchor-NST and four-anchor-NC) remained below that level. The constant single-anchors – that consisted of an anchor with the constant length of only one item – displayed false alarm rates not exceeding 0.01, whereas the constant four-anchors displayed slightly higher false alarm rates (approximately 0.03 for the constant four-anchor-AO as well as for the four-anchor-NST method and about 0.045 for the constant four-anchor-NC method). Hence, DIF tests with an anchor method from the constant anchor class – especially the constant single-anchor methods – were over-conservative.

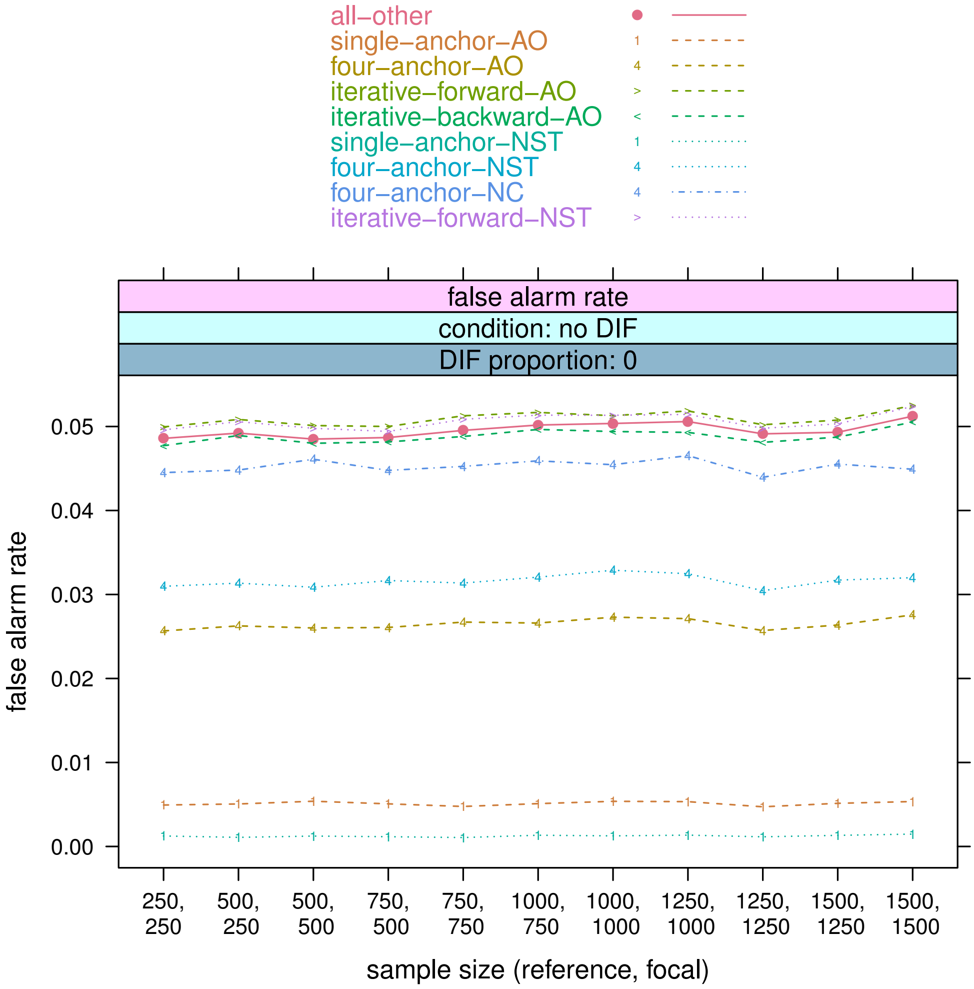

**Figure 21** – *False alarm rates under the null hypothesis of no DIF.*

*Summary*

All anchor methods held the alpha-level. Over-conservative test results were found for the constant anchor methods, whereas the newly suggested iterative-forward-AO and iterative-forward-

NST methods together with the all-other and the iterative-backward-AO methods were near the alpha level.

### 6.5.2 Balanced DIF: No advantage for one group

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

*False alarm rates*

Figure 22 (top row) contains the false alarm rates for the balanced condition, reported also for equal sample sizes together with the standard errors in Table 7 in the Appendix. Most methods displayed well-controlled false alarm rates – similar to the null condition – with the following exceptions: The constant four-anchor-NST and the constant four-anchor-NC method showed a false alarm rate that first increased but then decreased again with growing sample size. The same inverse u-shaped pattern occurred in case of unbalanced DIF and will be explained in more detail in Section 6.7.

Both constant single-anchor methods (single-anchor-AO and -NST) as well as the four-anchor-AO method, again, remained below the significance level. Hence, DIF tests based on the single-anchor-AO, the single-anchor-NST and the four-anchor-AO method were over-conservative.

*Hit rates*

For DIF test results, hit rates specify how likely true DIF is detected. Figure 22 (bottom row) depicts the hit rates in the balanced condition that increased monotonically with the sample size (for standard errors see also Table 8 in the Appendix). The hit rates with the slowest increase corresponded to the constant single-anchor methods, but also to the constant four-anchor methods. The methods from the constant anchor class that were combined with the AO-selection (single-anchor-AO, four-anchor-AO) achieved higher hit rates than those combined with the NST-selection (single-anchor-NST, four-anchor-NST) or its modification (four-anchor-NC). All anchor methods reached a hit rate of almost one, when the sample size was sufficiently high. In terms of hit rates, all iterative procedures (iterative-forward-AO, iterative-forward-NST and iterative-backward-AO) as well as the all-other method showed rapidly increasing hit rates that converged to one for sample sizes above 750 in each group.

*Summary*

In the balanced condition, the AO-selection strategy outperformed the NST-selection by yielding higher hit rates as expected. The difference was large for methods from the constant anchor class, but negligible for methods from the iterative forward anchor class.

All anchor methods showed a well-controlled false alarm rate, except the constant four-anchor-NST and the four-anchor-NC method. All iterative methods (from the forward and backward class) and the all-other method displayed the most rapidly rising hit rates. The newly suggested

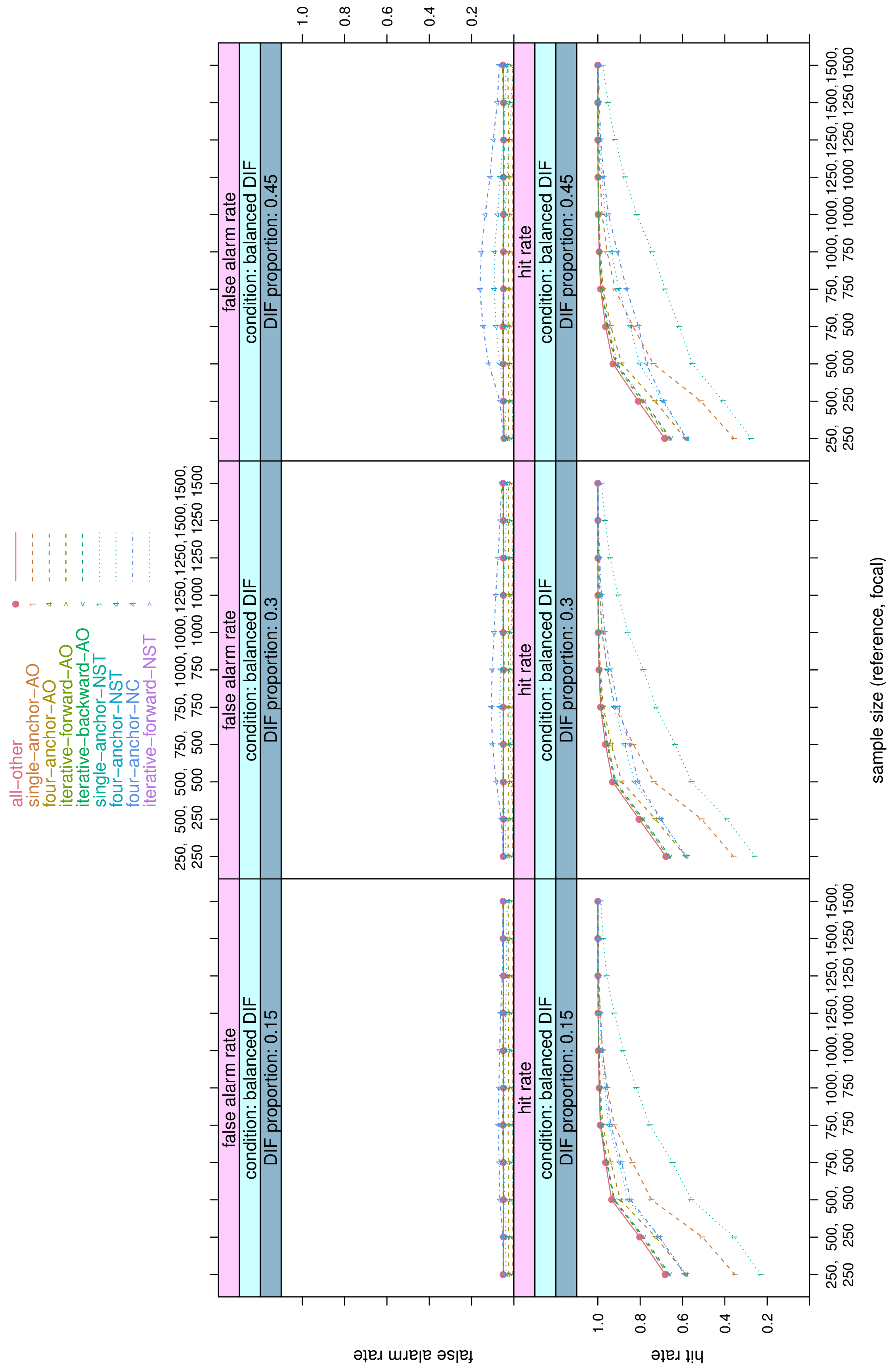

**Figure 22 –** *Balanced condition:* 15%, 30% *and* 45% *DIF-items with no systematic advantage for one group; sample size varies from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the balanced condition.*

iterative-forward-AO and iterative-forward-NST method enabled a high rate of correctly classified DIF-items and simultaneously maintained the alpha level in the balanced condition.

### 6.5.3 Unbalanced DIF: Advantage for the focal group

In the unbalanced condition, all items simulated with different item parameters favored the focal group. False alarm rates for the unbalanced condition are shown in Figure 23 (top row) and in Table 9 in the Appendix together with the standard errors (again only for settings with equal sample sizes in reference and focal group to save space).

*False alarm rates*

As opposed to the previous results, in this condition the majority of the anchor methods produced inflated false alarm rates: When the proportion of DIF-items increased, the false alarm rates rose as well. Moreover, for most anchor methods, the false alarm rates increased with growing sample size. The settings from the unbalanced condition – especially with 30% and 45% DIF-items – are now discussed in more detail in groups of anchor classes.

The all-other method yielded the highest false alarm rate in the majority of the simulation settings. The reason for this is that the all-other method was always contaminated in situations where more than one item had DIF. On average, the mean item parameters of the focal group were lower than the mean item parameters of the reference group. These mean differences in the item parameters shifted the scales of focal and reference group apart when the all-other method defined the restriction (similar to the instructive example in Section 5.2). These artificial differences became significant when the sample size increased and, thus, resulted in an inflated false alarm rate.

For methods from the constant anchor class, the selection strategy explained the false alarm rates: The strategy of selecting anchors based on the DIF tests with all-other items as anchor yielded biased DIF test results that induced a high false alarm rate when the sample size was large (as illustrated and discussed in more detail regarding the impact of contamination in Section 6.6).

Constant anchors selected by the number-of-significant strategy produced lower false alarm rates in regions of medium or large sample sizes. Here, again, an inversely u-shaped form was visible. After a certain point, the false alarm rates decreased again (a detailed explanation will be given in Section 6.7). The constant single-anchor methods showed lower false alarm rates than the corresponding constant four-anchor methods. For all constant methods, the single-anchor-NST method had the lowest false alarm rate when the sample size was large. Only in the condition with 45% DIF-items, it systematically exceeded the alpha level at medium sample sizes.

The method from the iterative backward anchor class, which started the initial step by using the all-other method, also led to inflated false alarm rates (with a maximum observed rate of 0.31 for the highest sample size in case of 45% DIF) that rose when sample size increased.

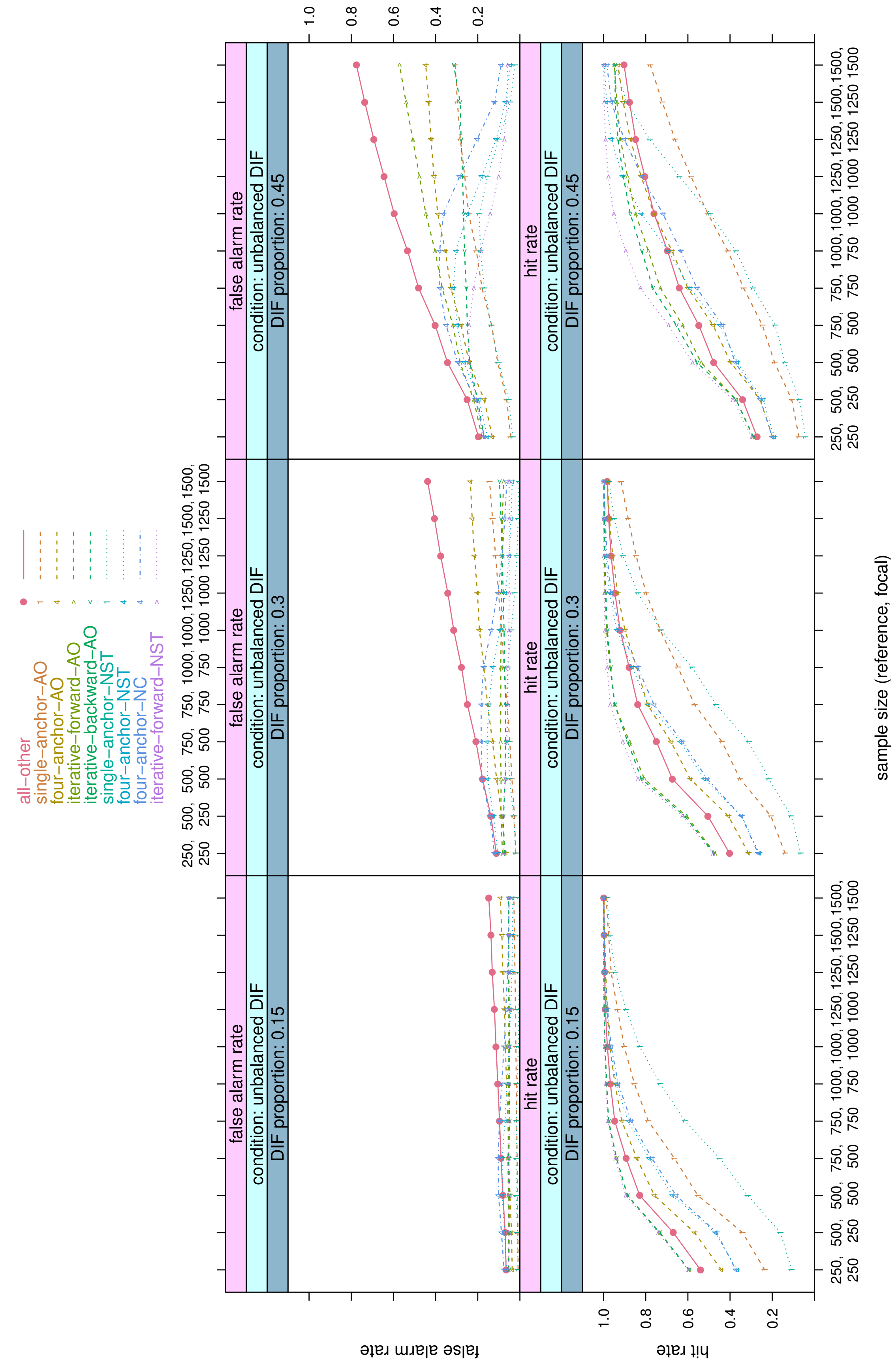

**Figure 23 –** *Unbalanced condition:* 15%, 30% *and* 45% *DIF-items favoring the focal group; sample size varies from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the unbalanced condition.*

Methods from the iterative forward class displayed heterogeneous false alarm rates. The iterative-forward-AO method led to increased false alarm rates – similar to the constant methods with the AO-selection criterion – in the setting with 30% or 45% DIF (up to 0.57 for the highest sample size). The clearly best iterative method in terms of a low false alarm rate was the new iterative-forward-NST method (yielding a false alarm rate of 0.06 for the highest sample size in case of 45% DIF and an observed maximum of 0.24).

*Hit rates*

The hit rate in the setting with 15% DIF in the unbalanced condition was relatively similar to the corresponding setting in the balanced condition (cf. Figure 23 (bottom row) and Table 10 in the Appendix). The hit rate increased with growing sample size. But in regions of small sample sizes especially the hit rates of the all-other and of methods from the constant anchor class were lower.

The results in the settings of larger proportions of DIF-items were different: Generally, the overall level of the hit rate was lower. Methods from the constant anchor class as well as the all-other method showed the slowest increase with the sample size. These methods also had lower hit rates compared to the methods from the iterative forward or backward class that were the only methods that displayed rapidly increasing and high hit rates. The new iterative forward-NST method provided the highest hit rate and a rapid rise of the hit rate with increasing sample size. The NST-selection strategy in combination with methods from the constant anchor class was more suitable than the AO-selection strategy regarding the hit rates when the sample size was large. The simplified four-anchor-NST method outperformed the original suggestion of Wang (2004), the four-anchor-NC method, in terms of higher hit rates (and lower false alarm rates). The iterative forward procedure with the NST-selection was equal or superior to the iterative-forward-AO method over the entire range of simulated sample sizes.

*Summary*

In the unbalanced condition, the NST-selection strategy was superior to the AO-selection strategy when the sample size and the DIF proportion were high as expected, since it not only allowed a higher hit rate but it also corresponded to a lower false alarm rate.

In the condition of unbalanced DIF, the false alarm rates were no longer well-controlled. When the DIF proportion was high, only the single-anchor-NST and the iterative-forward-NST method had low false alarm rates in regions of large sample sizes. Both constant single-anchor methods yielded low false alarm rates – but also low hit rates – when the sample size was small. All methods from the constant anchor class, especially in regions of small sample sizes, showed poor hit rates. The highest hit rate – in all settings from the unbalanced condition – corresponded to the newly proposed iterative-forward-NST method.

## 6.6 The impact of anchor contamination

As discussed in Section 5.2 the contamination of the anchor may induce artificial DIF and thus induce a seriously inflated false alarm rate. Short anchors are preferred in the constant anchor class in order to minimize the risk of contamination that includes only the information whether all anchor items are DIF-free or not (e.g. Shih and Wang, 2009). When new anchor methods are proposed, they are judged by their ability to correctly locate a completely DIF-free anchor in order to avoid anchor contamination (e.g. Wang *et al.*, 2012). Since contamination is an important argument in the construction of new anchor methods and since it affects the results of DIF detection (see Section 5.2), we will take a deeper look at the simulation results focussing on the aspect of anchor contamination.

For one exemplary simulation setting, the extreme condition of unbalanced DIF where 45% of the items favored the focal group will now be discussed in more detail in groups of anchor classes to illustrate the impact of contamination.

Figure 24 (top row) depicts the proportion of replications where at least one item of the anchor was a DIF-item (top-left) – this is referred to as *risk of contamination* – and the proportion of DIF-items in the anchor when the anchor was contaminated (top-right) – this is referred to as *degree of contamination*. The false alarm rates including only the replications that resulted in a contaminated anchor are displayed in Figure 24 (bottom-left) next to the false alarm rates including only the replications that resulted in a pure anchor (bottom-right).

The results show the following: For the all-other method all items functioned as anchor items. Due to the fact that more than one item has DIF in our simulation design, the anchor was contaminated in 100% of the replications. The proportion of DIF-items in the anchor was 45% as simulated. With increasing sample size, the power of detecting artificial DIF (DIF-free items that display DIF due to the anchor method chosen) increased and, thus, the false alarm rate rose.

Methods from the constant anchor class are investigated next. The risk of contaminated anchors decreased when the sample size increased. The anchor selection strategy determined how rapidly: When the NST-selection strategy chose the anchor, the convergence to zero percent contamination was clearly visible in the simulated range of the sample size. The all-other selection strategy showed a decreasing tendency, but even when 1500 observations for each group were available, 70% of the anchors were still contaminated for the four-anchor-AO and 17% for the single-anchor-AO method.

If the constant single-anchors were contaminated, this means that the single anchor item had DIF and, inevitably, the false alarm rates exploded when the sample size was large enough to detect significant artificial DIF (in case of 1500 observations in each group: single-anchor-AO: 0.70, single-anchor-NST: 0.54).

Surprisingly, there was a large gap between the degree of contamination for the constant four-anchor methods: When the AO-strategy or the NST-strategy were directly chosen and the sample size was large, on average about one out of four anchor items had DIF. In contrast to this, about three out of four anchor items had DIF when the four-anchor-NC method was chosen. In contaminated situations, consequently, the four-anchor-NC method corresponded to a larger false alarm rate (observed maximum: 0.75) than the four-anchor-AO (observed maximum: 0.54) or the four-anchor-NST method (observed maximum: 0.36). Therefore, the four-anchor-NC method corresponded to larger false alarm rates compared to the four-anchor-NST method

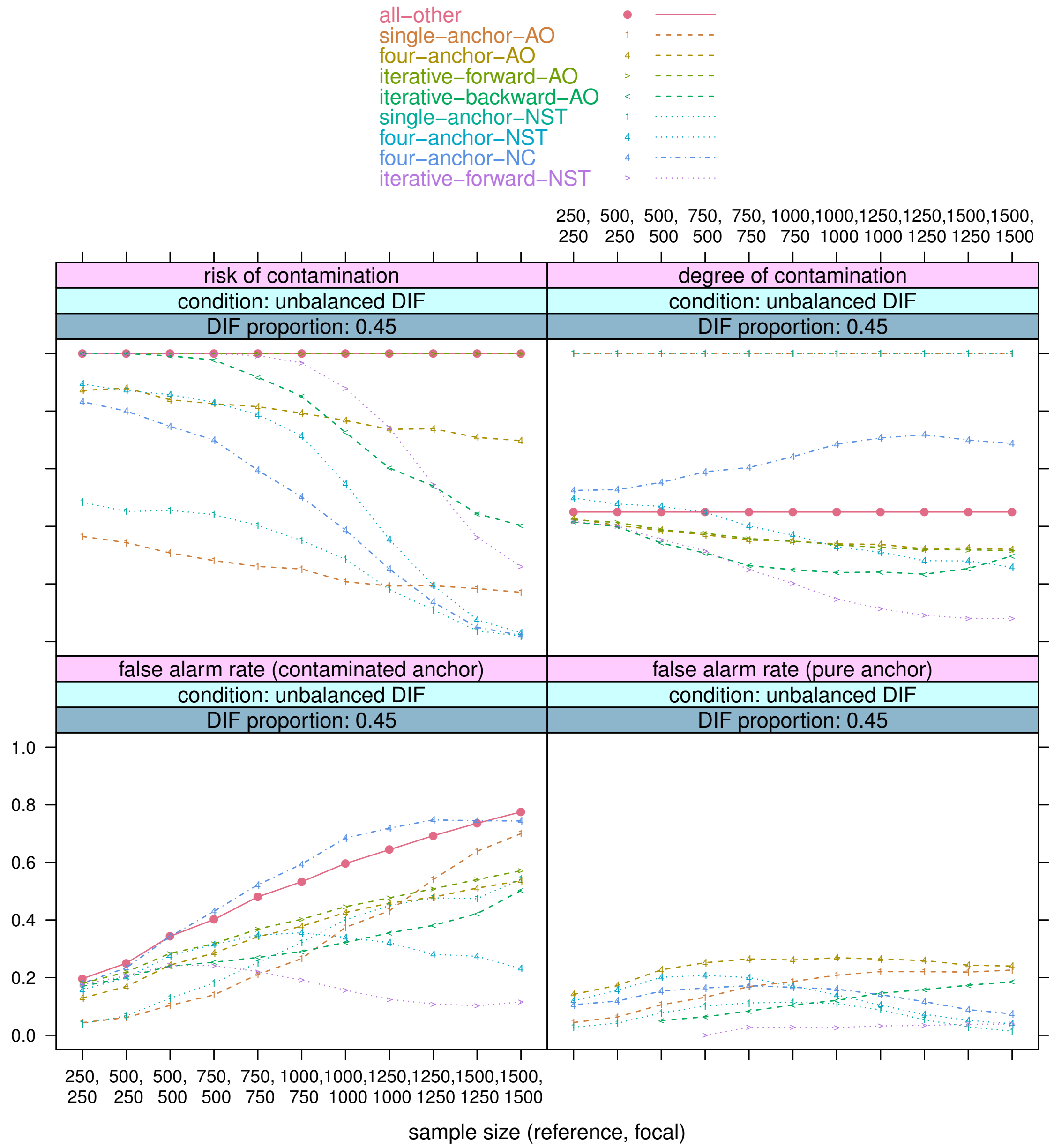

**Figure 24 –** *Condition of unbalanced DIF with* 45% *DIF-items favoring the focal group; sample size ranges from* (250, 250) *to* (1500, 1500)*; top-left: risk of contaminated anchors (at least one DIF-item included in the anchor); top-right: degree of contamination (proportion of DIF-items in contaminated anchors); bottom-left: false alarm rates when the anchor is contaminated; bottom-right: false alarm rates when the anchor is pure.*

over all unbalanced conditions with 45% DIF-items (see again Figure 23, top row), even though it had a lower risk of contamination. Hence, the degree of contamination is important for the results of the DIF assessment.

If the anchor was pure, the false alarm rates for the constant four-anchor method located by the NST-selection strategy (or the NC-selection) were lower. Note, however, that even if the anchor was pure, the false alarm rates of the constant anchor methods exceeded the alpha level. To clarify this fact, we will present an additional simulation study in the next section.

The methods from the iterative forward and backward anchor class were more often contaminated than the constant single-anchors (see Figure 24). This is not surprising, as the anchors included more items. Similar to the all-other method, all replications of the iterative-forward-AO method were contaminated. The iterative-forward-NST and the iterative-backward-AO method yielded a risk of contamination that decreased with the sample size (minimum 0.26 and 0.40 in the simulated range). In case of contaminated anchors, the methods from the iterative forward and backward class also produced inflated false alarm rates. When the sample size in each group exceeded 750, the iterative-forward-NST method definitely had the lowest false alarm rate. When the sample size was 1000 in each group, the mean false alarm rate of the iterative-forward-NST method including all replications did not even exceed the mean false alarm rate of all three constant four-anchor methods including only the pure replications while simultaneously the hit rate was higher for the iterative-forward-NST method.

Our findings clarify that it is not the risk of contamination alone that explains the inflated false alarm rates. The best method in the unbalanced condition when the sample size was large was the iterative-forward-NST method even if it had a high risk of contamination. Nevertheless, the iterative-forward-NST method displayed a low false alarm rate and also a high hit rate. Therefore, the consequences of contamination depended on the degree of contamination which was low for this method due to the NST-selection strategy that was suitable for unbalanced DIF. Research on DIF analysis and anchor selection procedures should, thus, not only concentrate on the risk of contamination, but also focus on the consequences, which strongly depend on the proportion of contaminated items in the anchor.

## 6.7 Characteristics of the anchor items inducing artificial DIF

In our simulation study, several anchor methods – especially the four-anchor-NST and the four-anchor-NC method – displayed inversely u-shaped false alarm rates that are yet to be explained. There are two mechanisms at work here: On one hand, the risk and the degree of contamination decrease with increasing sample size when the anchor selection strategy works appropriately and, thus, the extent of artificial DIF decreases. On the other hand, the power of detecting artificial DIF increases with growing sample size. One possible explanation for the inversely u-shaped pattern is the interaction between the decreasing extent of artificial DIF induced by anchor contamination and the increasing power of detecting statistically significant artificial DIF. In the beginning the false alarm rate increases due to the increasing power for detecting artificial DIF but at some point the false alarm rate decreases again as the risk of contamination decreases. This explanation is consistent with the findings from Section 6.6, where the anchor was contaminated: The four-anchor-NST method, for example, displayed a degree of contamination that decreased with sample size (see again Figure 24, top-right) and an inversely u-shaped false alarm rate when the anchor was contaminated (see again Figure 24, bottom-left).

However, with this argument we cannot yet explain why the false alarm rates showed the same pattern for pure (uncontaminated) replications (see again Figure 24, bottom-right). Here, the single-anchor-NST, the four-anchor-NST as well as the four-anchor-NC method displayed inversely u-shaped false alarm rates even though the anchor was pure. Therefore, the presence of artificial DIF that is induced by contamination alone cannot explain this finding. To understand this finding, it is important to note that artificial DIF can also be caused by special characteristics of the anchor items that were located by an anchor selection strategy.

To clarify how artificial DIF is related to the observed patterns of the false alarm rates, we conducted an additional simulation study focussing again on the extreme condition of unbalanced DIF where 45% of the items favored the focal group. Here, we examined the difference in the sum of the estimated anchor item parameters between focal and reference group that we termed *scale shift* (because it measures how far both scales are shifted apart during the construction of the common scale for the item parameters and this shift can cause artificial DIF as illustrated in Figure 18, right) for all constant four-anchor methods (the four-anchor-NST and the four-anchor-NC method that displayed inversely u-shaped patterns as well as the four-anchor-AO method that displayed an increasing false alarm rate, see again Figure 23, top-right). To assess reliable estimates of the scale shift, we used all items that are DIF-free by design as anchor items to build the ideal common scale for the item parameters. The scale shift represents how far the scales are shifted apart and reflects the extent of artificial DIF. The scale shift may be caused by contamination, as discussed in the previous sections, or by special characteristics of the anchor items in particular when the selection strategies locate anchor items that show relatively high empirical differences in the estimated item parameters due to random sampling fluctuation even if the located anchor items were simulated to be DIF-free.

To determine whether anchor items found by a selection strategy display this characteristic, we included a benchmark method of four anchor items that were randomly selected from the set of all DIF-free items. The benchmark method, thus, represents the ideal four-anchor method that does not select items with high differences more often than others. The results, separated for contaminated and pure replications, are depicted in Figure 25.

In case of contaminated anchors (Figure 25, left), the four-anchor methods displayed positive scale shifts. Even though the scale shifts were almost constant over the sample size in regions of small to medium sample sizes for the four-anchor-AO and four-anchor-NC or even slightly decreasing for the four-anchor-NST method, the false alarm rates rose with growing sample size in the respective range of the sample sizes (Figure 24, bottom-left). We attribute this fact to the increasing power of detecting artificial DIF. This also explains the increasing false alarm rates of the four-anchor-AO and the four-anchor-NC methods: The scale shifts were almost constant over the simulated range of the sample size but the false alarm rates increased (Figure 24, bottom-left).

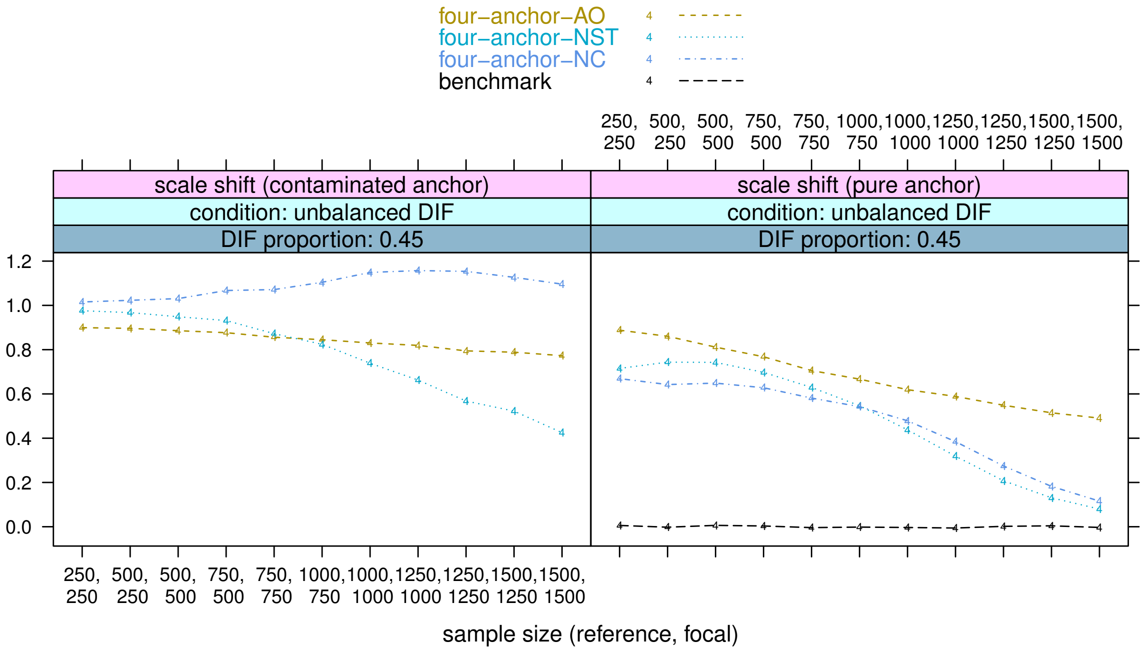

**Figure 25 –** *Condition of unbalanced DIF with* 45% *DIF-items favoring the focal group; sample size ranges from* (250, 250) *to* (1500, 1500)*; left: the scale shift when the anchor is contaminated; right: the scale shift in case of pure anchors.*

For the four-anchor-NST method the scale shift also decreased with increasing sample size in regions of medium or large sample sizes (Figure 25, left) and so did the false alarm rate in the respective range of the sample sizes (Figure 24, bottom-left).

Our second argument now becomes important in the case of pure anchors: The scale shift for the benchmark method of randomly chosen DIF-free anchor items (Figure 25, right) fluctuated around zero and displayed no systematic shift in one direction. However, the scale shift of all remaining constant four-anchor methods was positive. This represents the fact that the supposedly pure items chosen by an anchor selection strategy displayed different characteristics than randomly chosen pure anchor items. From all items that were "pure" by definition (i.e. were drawn from distributions with no parameter difference) the anchor selection strategies selected not the ones with the lowest empirical difference (due to random sampling), as one might hope, but those with a large empirical difference which again induced artificial DIF for the other items.

As can be seen from Figure 25 (right), the scale shift for the four-anchor methods reduced with increasing sample size. In regions of large sample sizes, the scale shift is directly related to the false alarm rate: When the scale shift was high (as was the case for the four-anchor-AO method), the false alarm rate was high as well (Figure 24, bottom-right). When the scale shift decreased with growing sample size (e.g. for the four-anchor-NST method), the corresponding false alarm rate decreased as well. In regions of smaller sample sizes, the scale shift of all four-anchor methods was high, but the false alarm rates were low at the beginning and then increased with growing sample size. Here, again, the interaction between the extent of artificial DIF – now induced by large empirical differences in the anchor items – and the power of detecting artificial DIF was visible.

These findings explain why the u-shaped patterns occur for the four-anchor-NST and the four-anchor-NC method: These methods are able to reduce the scale shift with increasing sample size because the scale shift in pure settings reduces and the risk of contamination reduces as well (i.e. the number of pure settings increases). Taking the increasing power of detecting artificial DIF with growing sample size into account, an inversely u-shaped pattern results for the false alarm rates. In contrast to this, the four-anchor-AO method always displayed a relatively high scale shift (that only reduces slightly when the anchor was pure). The power of detecting artificial DIF increased with growing sample size and, therefore, the false alarm rate showed an increase and no considerable decrease.

In summary, the interaction between a decreasing extent of artificial DIF and an increasing statistical power to detect artificial DIF with growing sample size results in an inversely u-shaped false alarm rate. The risk and degree of contamination alone cannot explain the presence of artificial DIF. The anchor items selected by certain anchor selection strategies differed systematically from randomly chosen pure anchor items even if the located anchor items were by definition DIF-free. Counterintuitively instead of items with small differences, these methods tended to select exactly those items with large differences. Therefore, the anchor items found by the NST-, the NC- or the AO-selection strategy displayed a positive scale shift in the additional simulation study and, thus, shifted the scales apart and induced artificial DIF. This implies that not only the risk and the degree of contamination but also the scale shift in by definition pure replications should be regarded when anchor methods are developed and investigated in simulation studies. Otherwise, inflated false alarm rates might occur even if the anchor is pure.

## 6.8 Summary and discussion

To conclude, the results from the simulation study are briefly reviewed. Thereafter, practical guidelines on how to choose the anchor for DIF analysis in the Rasch model are given. Finally, future research questions are addressed.

*Conclusions from the simulation study*

The assessment of differential item functioning for the Rasch model was investigated. The results of the Wald test were compared by means of hit and false alarm rates under three main conditions: The null hypothesis of no DIF, the balanced condition where no group had a systematic advantage and the unbalanced condition where all items favored the focal group.

Under the null hypothesis, all methods from the iterative forward and backward class as well as the all-other method were near the alpha-level, while methods from the constant anchor class remained below that level.

When DIF was balanced, the all-other method and also methods from the iterative forward and backward class yielded high hit rates while simultaneously exhausted the alpha-level. As expected, the all-other selection strategy outperformed the number-of-significant selection strategy.

In case of unbalanced DIF, the NST-selection procedure was superior to the AO-selection strategy when the sample size was large. In this case, the newly suggested iterative-forward-NST method yielded the highest hit rate and a low false alarm rate and was, thus, the best performing anchor method.

The constant four-anchor class was not only combined with the AO-selection and the NST-selection strategy, but also in the way Wang (2004) originally suggested. Even though the four-anchor-NC method led to a low risk of contamination (see Section 6.6), it yielded higher false alarm rates and lower hit rates compared to the four-anchor-NST method. The latter was directly build with the NST-selection that simplifies the suggestion of Wang (2004). Thus, the four-anchor-NST method was superior.

Based on these results, when no reliable prior knowledge about the DIF situation exists, as will be the case in most real data analysis settings (as opposed to simulation analysis where the true DIF pattern is known), we thus recommend to use the iterative-forward-NST method. When the sample size was large enough, the false alarm rates were low in any condition even if the anchor was contaminated. Hit rates rapidly grew with the sample size and converged to one. The forward item-wise selection of anchor items outperformed the iterative-backward-AO, iterative-forward-AO, the all-other as well as anchor methods from the constant anchor class.

There are several reasons that explain the superior performance of the iterative-forward-NST method. Firstly, the method has a head start compared to the methods that rely on DIF tests using the all-other items as anchor (e.g. the classical iterative procedures, such as the iterative-backward-AO). While the latter start with a criterion that is severely biased when DIF is unbalanced, the iterative-forward-NST method does not require that DIF effects almost cancel out (for a detailed discussion of this assumption for the all-other method see Wang, 2004). Secondly, the NST-selection strategy combined with the iterative forward anchor class also performs well in case of balanced DIF. While the AO-selection strategy performed better than the NST-selection strategy when it was combined with the methods from the constant anchor class, the advantage in combination with the iterative forward class appeared negligible. Thirdly, our study showed that the consequences of contamination depend on the proportion of contaminated items rather than on the risk of contamination itself. Therefore, the iterative-forward-NST method yielded better results in DIF analysis even though the anchor was long and, thus, often contaminated. The risk of contamination decreases with increasing sample size and, beyond that, the proportion of DIF-items in contaminated situations decreased. Fourthly, the iterative forward anchor class adds items to the anchor as long as the number of anchor items is smaller than the set of presumed DIF-free items. If the sample size is large enough, this leads to the desirable property, that it produces a longer anchor when the proportion of DIF-items is low and a shorter anchor if the proportion of DIF-items is high, similar to the iterative backward method. It may appear as a drawback that the iterative forward anchor class uses a short anchor in the initial steps, beginning with only one anchor item. The resulting DIF tests may lack statistical power due to fact that the anchor is short. However, this does not affect the performance of the new iterative forward anchor methods since the test results are only used for the decision whether the anchor should include one more anchor item and, thus, a small statistical power of the DIF tests in the first iterations automatically yields to a longer anchor that is expected to increase the power of the DIF test itself. Another astounding finding of our simulations was that anchor items located by an anchor selection strategy displayed different characteristics compared to randomly chosen DIF-free items and may be exactly those items that again induce artificial DIF. Including more anchor items (than e.g. four anchor items) reduces the artificial scale shift that is induced by anchor items with empirical group differences and, thus, can also occur when the anchor is (by definition) pure.

In addition to the item parameter values presented in Chapter 6.4, we have replicated the main results of our simulation study with the item parameter values $\beta$ = (-2.522, -1.902, -1.351, -1.092, -0.234, -0.317, 0.037, 0.268, -0.571, 0.317, 0.295, 0.778, 1.514, 1.744, 1.951, -1.152, -0.526, 1.104, 0.961, 1.314, -2.198, -1.621, -0.761, -1.179, -0.610, -0.291, 0.067, 0.706, -2.713, 0.213, 0.116, 0.273, 0.840, 0.745, 1.485, -1.208, 0.189, 0.345, 0.962, 1.592) used by Wang *et al.* (2012). Furthermore, the direction of DIF was changed so that – in case of unbalanced DIF – all items favored the reference group. The results were similar to the results presented in this chapter and, thus, indicate the stability of our simulation study, are depicted at the end of the Appendix of this chapter (see Figure 26 and Figure 27). Therefore, we are confident that the different behavior of the anchor methods is not limited to the settings investigated here.

*Practical recommendations*

Our simulation study highlights the importance of the anchor selection for the correct classification of DIF and DIF-free items. A careful consideration of the anchor method used is necessary to avoid high misclassification rates and doubtful test results.

In case of balanced DIF, the all-other method was slightly better than the iterative-forward-NST strategy. However, due to the fact that the all-other method resulted in seriously inflated false alarm rates when the situation was unbalanced – and that it is doubtful whether the situation of balanced DIF is ever met in practice (Wang and Yeh, 2003; Wang *et al.*, 2012) – the usage of this anchor method is inadvisable. Thus, the iterative-forward-NST strategy is recommended. When the sample size was large enough, the false alarm rates were low in any condition even if the anchor was contaminated and the hit rates grew rapidly.

The adequacy of the selection strategies – by the number-of-significant or by the all-other approach – depends on the DIF situation. In the balanced condition, the AO-selection strategy performed suitable, whereas in the unbalanced condition the NST-selection strategy was more appropriate. But when the iterative-forward class was used, the advance of the AO-selection strategy was marginal. Therefore, we recommend the iterative-forward-NST method over the iterative-forward-AO method.

*Related research questions*

The simulation study presented here was limited to DIF analysis using the Wald test in the

Rasch model. Thus, future research may investigate the usefulness of the iterative-forward-NST method for other IRT models and combine it with other methods that test for DIF.

Furthermore, the iterative forward anchor class with the NST-selection may be compared with modifications of the anchor selection strategy. For example, Shih and Wang (2009) suggest to use the items corresponding to the lowest rank of the mean absolute DIF statistics similar to the all-other strategy of Woods (2009). Then items are anchor candidates if they display the lowest mean DIF test statistic when every item is tested for DIF using every other item as constant single-anchor. This modification may be less affected by sample size.

Wang *et al.* (2012) established an improvement of the AO-selection strategy by incorporating additional iterations. Firstly, every item is tested for DIF using the all-other method. Then, iteratively, DIF-items are excluded from the anchor candidates and a new DIF analysis using the current anchor is conducted until two steps reach the same results. Finally, the anchor items are selected from the remaining candidates using the all-other strategy.

Moreover, the DIF test results may also be improved by the construction of new anchor selection strategies. Ideally, the anchor items are DIF-free and induce no artificial scale shift. Furthermore, the impact of the degree of contamination is important for the appropriateness of the results in DIF detection.

Therefore, improving the anchor selection strategies with the aim to locate anchors with a small degree of contamination is the aim of the next chapter. Therein, the all-other selection, the number-of-significant selection, the suggestion of Shih and Wang (2009), the improved version of the all-other selection by Wang *et al.* (2012) and three new anchor selection strategies are systematically compared.

## 6.9 Appendix

| **false alarm** | all-other | single-anchor-AO | four-anchor-AO | iterative-forward-AO | iterative-backw.-AO | single-anchor-NST | four-anchor-NST | four-anchor-NC | iterative-forward-NST |
|---|---|---|---|---|---|---|---|---|---|
| **no DIF** | | | | | | | | | |
| 250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.04 | 0.05 |
| | (0.03) | (0.01) | (0.03) | (0.03) | (0.03) | (0.01) | (0.03) | (0.04) | (0.03) |
| 500 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.03) | (0.01) | (0.03) | (0.04) | (0.03) |
| 750 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.03) | (0.01) | (0.03) | (0.04) | (0.04) |
| 1000 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 |
| | (0.03) | (0.01) | (0.03) | (0.03) | (0.03) | (0.01) | (0.03) | (0.04) | (0.03) |
| 1250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.04 | 0.05 |
| | (0.03) | (0.01) | (0.03) | (0.03) | (0.03) | (0.01) | (0.03) | (0.04) | (0.03) |
| 1500 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.04 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.03) | (0.01) | (0.03) | (0.04) | (0.04) |

**Table 6** – *False alarm rates and standard errors under the null hypothesis (no DIF) with equal sample sizes in reference and focal group.*

| **false alarm** | all-other | single-anchor-AO | four-anchor-AO | iterative-forward-AO | iterative-backw.-AO | single-anchor-NST | four-anchor-NST | four-anchor-NC | iterative-forward-NST |
|---|---|---|---|---|---|---|---|---|---|
| **15% DIF** | | | | | | | | | |
| 250 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.04 | 0.05 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) |
| 500 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.01 | 0.04 | 0.07 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.06) | (0.04) |
| 750 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.01 | 0.05 | 0.07 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.06) | (0.04) |
| 1000 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.04 | 0.07 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.06) | (0.04) |
| 1250 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.04 | 0.06 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.05) | (0.04) |
| 1500 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) |
| **30% DIF** | | | | | | | | | |
| 250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.04 | 0.05 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.05) | (0.04) |
| 500 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.01 | 0.05 | 0.08 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.02) | (0.05) | (0.07) | (0.04) |
| 750 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.02 | 0.07 | 0.11 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.03) | (0.05) | (0.08) | (0.04) |
| 1000 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.01 | 0.06 | 0.09 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.02) | (0.05) | (0.08) | (0.04) |
| 1250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.01 | 0.04 | 0.08 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.02) | (0.04) | (0.07) | (0.04) |
| 1500 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.00 | 0.03 | 0.06 | 0.05 |
| | (0.04) | (0.01) | (0.03) | (0.04) | (0.04) | (0.01) | (0.04) | (0.05) | (0.04) |
| **45% DIF** | | | | | | | | | |
| 250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.00 | 0.04 | 0.05 | 0.05 |
| | (0.05) | (0.02) | (0.03) | (0.05) | (0.05) | (0.01) | (0.04) | (0.06) | (0.05) |
| 500 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.02 | 0.07 | 0.12 | 0.05 |
| | (0.05) | (0.01) | (0.03) | (0.05) | (0.05) | (0.04) | (0.06) | (0.11) | (0.05) |
| 750 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.05 | 0.09 | 0.16 | 0.05 |
| | (0.05) | (0.02) | (0.03) | (0.05) | (0.04) | (0.06) | (0.07) | (0.13) | (0.05) |
| 1000 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.04 | 0.08 | 0.14 | 0.05 |
| | (0.05) | (0.02) | (0.03) | (0.05) | (0.05) | (0.06) | (0.07) | (0.12) | (0.05) |
| 1 250 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 | 0.02 | 0.05 | 0.10 | 0.05 |
| | (0.05) | (0.01) | (0.03) | (0.05) | (0.04) | (0.05) | (0.05) | (0.09) | (0.05) |
| 1500 | 0.05 | 0.01 | 0.03 | 0.05 | 0.05 | 0.01 | 0.04 | 0.07 | 0.05 |
| | (0.05) | (0.02) | (0.04) | (0.05) | (0.05) | (0.02) | (0.04) | (0.07) | (0.05) |

**Table 7** – *False alarm rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.*

| **hit rate** | all-other | single-anchor-AO | four-anchor-AO | iterative-forward-AO | iterative-backw.-AO | single-anchor-NST | four-anchor-NST | four-anchor-NC | iterative-forward-NST |
|---|---|---|---|---|---|---|---|---|---|
| **15% DIF** | | | | | | | | | |
| 250 | 0.68 | 0.35 | 0.58 | 0.67 | 0.66 | 0.23 | 0.58 | 0.59 | 0.67 |
| | (0.19) | (0.20) | (0.20) | (0.19) | (0.19) | (0.15) | (0.20) | (0.20) | (0.19) |
| 500 | 0.93 | 0.75 | 0.89 | 0.92 | 0.92 | 0.56 | 0.86 | 0.85 | 0.92 |
| | (0.10) | (0.18) | (0.12) | (0.11) | (0.11) | (0.15) | (0.13) | (0.14) | (0.11) |
| 750 | 0.99 | 0.92 | 0.98 | 0.99 | 0.98 | 0.75 | 0.95 | 0.94 | 0.98 |
| | (0.04) | (0.11) | (0.06) | (0.05) | (0.05) | (0.15) | (0.09) | (0.10) | (0.05) |
| 1000 | 1.00 | 0.98 | 0.99 | 1.00 | 1.00 | 0.88 | 0.98 | 0.98 | 1.00 |
| | (0.02) | (0.06) | (0.03) | (0.03) | (0.03) | (0.13) | (0.05) | (0.06) | (0.03) |
| 1250 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 | 1.00 | 1.00 | 1.00 |
| | (0.01) | (0.03) | (0.01) | (0.01) | (0.01) | (0.08) | (0.02) | (0.03) | (0.01) |
| 1500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 |
| | (0.00) | (0.01) | (0.01) | (0.00) | (0.00) | (0.05) | (0.01) | (0.02) | (0.00) |
| **30% DIF** | | | | | | | | | |
| 250 | 0.68 | 0.36 | 0.58 | 0.67 | 0.66 | 0.26 | 0.58 | 0.58 | 0.66 |
| | (0.13) | (0.14) | (0.14) | (0.13) | (0.13) | (0.11) | (0.14) | (0.13) | (0.13) |
| 500 | 0.93 | 0.74 | 0.89 | 0.92 | 0.92 | 0.56 | 0.83 | 0.81 | 0.92 |
| | (0.08) | (0.13) | (0.09) | (0.08) | (0.08) | (0.10) | (0.11) | (0.11) | (0.08) |
| 750 | 0.99 | 0.92 | 0.97 | 0.98 | 0.98 | 0.72 | 0.92 | 0.90 | 0.98 |
| | (0.03) | (0.08) | (0.05) | (0.04) | (0.04) | (0.11) | (0.08) | (0.09) | (0.04) |
| 1000 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 | 0.86 | 0.98 | 0.97 | 1.00 |
| | (0.01) | (0.04) | (0.02) | (0.01) | (0.01) | (0.10) | (0.04) | (0.05) | (0.02) |
| 1250 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.94 | 0.99 | 0.99 | 1.00 |
| | (0.01) | (0.02) | (0.01) | (0.01) | (0.01) | (0.08) | (0.02) | (0.03) | (0.01) |
| 1500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 |
| | (0.00) | (0.01) | (0.00) | (0.00) | (0.00) | (0.04) | (0.01) | (0.01) | (0.00) |
| **45% DIF** | | | | | | | | | |
| 250 | 0.68 | 0.36 | 0.59 | 0.67 | 0.66 | 0.28 | 0.58 | 0.58 | 0.66 |
| | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.09) | (0.11) | (0.11) | (0.11) |
| 500 | 0.93 | 0.74 | 0.89 | 0.92 | 0.91 | 0.55 | 0.80 | 0.77 | 0.91 |
| | (0.06) | (0.10) | (0.07) | (0.07) | (0.07) | (0.07) | (0.10) | (0.11) | (0.07) |
| 750 | 0.99 | 0.92 | 0.97 | 0.98 | 0.98 | 0.68 | 0.90 | 0.86 | 0.98 |
| | (0.03) | (0.06) | (0.04) | (0.03) | (0.03) | (0.09) | (0.09) | (0.10) | (0.04) |
| 1000 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 | 0.82 | 0.97 | 0.95 | 1.00 |
| | (0.01) | (0.03) | (0.02) | (0.01) | (0.01) | (0.10) | (0.05) | (0.07) | (0.02) |
| 1250 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.92 | 0.99 | 0.99 | 1.00 |
| | (0.00) | (0.02) | (0.01) | (0.01) | (0.01) | (0.09) | (0.02) | (0.04) | (0.01) |
| 1500 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 |
| | (0.00) | (0.01) | (0.00) | (0.00) | (0.00) | (0.04) | (0.01) | (0.02) | (0.00) |

**Table 8** – *Hit rates and standard errors in the balanced condition with equal sample sizes in reference and focal group.*

| **false alarm** | all-other | single-anchor-AO | four-anchor-AO | iterative-forward-AO | iterative-backw.-AO | single-anchor-NST | four-anchor-NST | four-anchor-NC | iterative-forward-NST |
|---|---|---|---|---|---|---|---|---|---|
| **15% DIF** | | | | | | | | | |
| 250 | 0.07 | 0.01 | 0.04 | 0.06 | 0.05 | 0.01 | 0.06 | 0.07 | 0.06 |
| | (0.04) | (0.02) | (0.03) | (0.04) | (0.04) | (0.02) | (0.04) | (0.06) | (0.04) |
| 500 | 0.08 | 0.01 | 0.05 | 0.05 | 0.05 | 0.02 | 0.08 | 0.10 | 0.05 |
| | (0.04) | (0.02) | (0.04) | (0.04) | (0.04) | (0.03) | (0.05) | (0.07) | (0.04) |
| 750 | 0.10 | 0.02 | 0.06 | 0.05 | 0.05 | 0.02 | 0.07 | 0.10 | 0.05 |
| | (0.04) | (0.02) | (0.04) | (0.04) | (0.04) | (0.03) | (0.05) | (0.07) | (0.04) |
| 1000 | 0.11 | 0.02 | 0.07 | 0.05 | 0.05 | 0.01 | 0.05 | 0.08 | 0.05 |
| | (0.05) | (0.02) | (0.04) | (0.04) | (0.04) | (0.02) | (0.04) | (0.06) | (0.04) |
| 1250 | 0.13 | 0.02 | 0.08 | 0.05 | 0.05 | 0.00 | 0.04 | 0.06 | 0.05 |
| | (0.05) | (0.03) | (0.04) | (0.04) | (0.04) | (0.01) | (0.04) | (0.05) | (0.04) |
| 1500 | 0.15 | 0.03 | 0.09 | 0.05 | 0.05 | 0.00 | 0.03 | 0.05 | 0.05 |
| | (0.05) | (0.03) | (0.05) | (0.04) | (0.04) | (0.01) | (0.03) | (0.05) | (0.04) |
| **30% DIF** | | | | | | | | | |
| 250 | 0.11 | 0.02 | 0.07 | 0.08 | 0.07 | 0.02 | 0.10 | 0.12 | 0.08 |
| | (0.05) | (0.03) | (0.04) | (0.06) | (0.06) | (0.03) | (0.06) | (0.08) | (0.06) |
| 500 | 0.18 | 0.04 | 0.11 | 0.09 | 0.07 | 0.05 | 0.15 | 0.18 | 0.07 |
| | (0.06) | (0.03) | (0.05) | (0.06) | (0.06) | (0.05) | (0.07) | (0.12) | (0.05) |
| 750 | 0.25 | 0.06 | 0.15 | 0.09 | 0.07 | 0.06 | 0.15 | 0.18 | 0.06 |
| | (0.07) | (0.05) | (0.07) | (0.06) | (0.05) | (0.06) | (0.08) | (0.14) | (0.05) |
| 1000 | 0.31 | 0.09 | 0.19 | 0.09 | 0.08 | 0.04 | 0.10 | 0.14 | 0.05 |
| | (0.07) | (0.05) | (0.08) | (0.06) | (0.05) | (0.06) | (0.07) | (0.11) | (0.04) |
| 1250 | 0.38 | 0.12 | 0.22 | 0.08 | 0.08 | 0.01 | 0.05 | 0.08 | 0.05 |
| | (0.07) | (0.06) | (0.09) | (0.06) | (0.06) | (0.03) | (0.05) | (0.08) | (0.04) |
| 1500 | 0.44 | 0.14 | 0.24 | 0.08 | 0.10 | 0.00 | 0.04 | 0.06 | 0.05 |
| | (0.07) | (0.08) | (0.10) | (0.06) | (0.06) | (0.02) | (0.04) | (0.06) | (0.04) |
| **45% DIF** | | | | | | | | | |
| 250 | 0.20 | 0.04 | 0.13 | 0.18 | 0.17 | 0.03 | 0.15 | 0.17 | 0.18 |
| | (0.07) | (0.04) | (0.07) | (0.09) | (0.10) | (0.04) | (0.08) | (0.11) | (0.10) |
| 500 | 0.34 | 0.11 | 0.24 | 0.28 | 0.24 | 0.10 | 0.26 | 0.29 | 0.24 |
| | (0.08) | (0.07) | (0.09) | (0.12) | (0.15) | (0.09) | (0.11) | (0.18) | (0.14) |
| 750 | 0.48 | 0.18 | 0.33 | 0.37 | 0.25 | 0.17 | 0.32 | 0.38 | 0.22 |
| | (0.09) | (0.10) | (0.12) | (0.13) | (0.18) | (0.13) | (0.14) | (0.27) | (0.16) |
| 1000 | 0.60 | 0.24 | 0.39 | 0.45 | 0.27 | 0.19 | 0.25 | 0.36 | 0.14 |
| | (0.09) | (0.15) | (0.16) | (0.15) | (0.20) | (0.19) | (0.16) | (0.32) | (0.15) |
| 1250 | 0.69 | 0.28 | 0.42 | 0.51 | 0.28 | 0.10 | 0.11 | 0.20 | 0.07 |
| | (0.08) | (0.20) | (0.18) | (0.16) | (0.20) | (0.17) | (0.13) | (0.26) | (0.09) |
| 1500 | 0.78 | 0.31 | 0.45 | 0.57 | 0.31 | 0.02 | 0.05 | 0.09 | 0.06 |
| | (0.08) | (0.25) | (0.21) | (0.17) | (0.23) | (0.09) | (0.06) | (0.13) | (0.06) |

**Table 9** – *False alarm rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.*

| **hit rate** | all-other | single-anchor-AO | four-anchor-AO | iterative-forward-AO | iterative-backw.-AO | single-anchor-NST | four-anchor-NST | four-anchor-NC | iterative-forward-NST |
|---|---|---|---|---|---|---|---|---|---|
| **15% DIF** | | | | | | | | | |
| 250 | 0.54 | 0.24 | 0.44 | 0.59 | 0.59 | 0.11 | 0.37 | 0.37 | 0.60 |
| | (0.20) | (0.17) | (0.20) | (0.22) | (0.22) | (0.13) | (0.19) | (0.21) | (0.22) |
| 500 | 0.83 | 0.55 | 0.76 | 0.89 | 0.89 | 0.32 | 0.67 | 0.66 | 0.89 |
| | (0.15) | (0.20) | (0.17) | (0.14) | (0.14) | (0.22) | (0.19) | (0.21) | (0.14) |
| 750 | 0.95 | 0.79 | 0.91 | 0.98 | 0.98 | 0.61 | 0.88 | 0.87 | 0.98 |
| | (0.09) | (0.16) | (0.11) | (0.06) | (0.06) | (0.25) | (0.14) | (0.16) | (0.06) |
| 1000 | 0.98 | 0.90 | 0.97 | 0.99 | 0.99 | 0.83 | 0.97 | 0.96 | 0.99 |
| | (0.05) | (0.12) | (0.07) | (0.03) | (0.03) | (0.20) | (0.07) | (0.08) | (0.03) |
| 1250 | 0.99 | 0.96 | 0.99 | 1.00 | 1.00 | 0.94 | 1.00 | 0.99 | 1.00 |
| | (0.03) | (0.07) | (0.04) | (0.01) | (0.02) | (0.11) | (0.03) | (0.04) | (0.01) |
| 1500 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 |
| | (0.01) | (0.05) | (0.02) | (0.01) | (0.01) | (0.06) | (0.01) | (0.02) | (0.01) |
| **30% DIF** | | | | | | | | | |
| 250 | 0.40 | 0.14 | 0.31 | 0.47 | 0.48 | 0.06 | 0.26 | 0.27 | 0.48 |
| | (0.13) | (0.10) | (0.13) | (0.17) | (0.18) | (0.08) | (0.13) | (0.15) | (0.18) |
| 500 | 0.67 | 0.35 | 0.59 | 0.81 | 0.82 | 0.22 | 0.52 | 0.51 | 0.84 |
| | (0.12) | (0.13) | (0.14) | (0.13) | (0.14) | (0.15) | (0.16) | (0.20) | (0.13) |
| 750 | 0.84 | 0.57 | 0.79 | 0.95 | 0.95 | 0.47 | 0.79 | 0.76 | 0.97 |
| | (0.10) | (0.14) | (0.12) | (0.07) | (0.07) | (0.21) | (0.14) | (0.19) | (0.06) |
| 1000 | 0.92 | 0.73 | 0.90 | 0.99 | 0.98 | 0.73 | 0.94 | 0.92 | 0.99 |
| | (0.08) | (0.13) | (0.09) | (0.04) | (0.04) | (0.21) | (0.09) | (0.11) | (0.02) |
| 1250 | 0.96 | 0.85 | 0.96 | 1.00 | 0.99 | 0.91 | 0.99 | 0.98 | 1.00 |
| | (0.05) | (0.11) | (0.06) | (0.02) | (0.02) | (0.13) | (0.03) | (0.05) | (0.01) |
| 1500 | 0.98 | 0.92 | 0.98 | 1.00 | 1.00 | 0.97 | 1.00 | 1.00 | 1.00 |
| | (0.04) | (0.08) | (0.04) | (0.01) | (0.02) | (0.06) | (0.01) | (0.02) | (0.00) |
| **45% DIF** | | | | | | | | | |
| 250 | 0.27 | 0.07 | 0.20 | 0.29 | 0.29 | 0.04 | 0.19 | 0.20 | 0.30 |
| | (0.09) | (0.06) | (0.09) | (0.13) | (0.14) | (0.05) | (0.09) | (0.12) | (0.14) |
| 500 | 0.48 | 0.19 | 0.40 | 0.54 | 0.56 | 0.14 | 0.37 | 0.38 | 0.58 |
| | (0.10) | (0.10) | (0.13) | (0.16) | (0.20) | (0.11) | (0.14) | (0.20) | (0.18) |
| 750 | 0.64 | 0.34 | 0.60 | 0.73 | 0.77 | 0.29 | 0.59 | 0.56 | 0.83 |
| | (0.10) | (0.15) | (0.15) | (0.14) | (0.18) | (0.20) | (0.18) | (0.28) | (0.15) |
| 1000 | 0.76 | 0.51 | 0.76 | 0.85 | 0.87 | 0.49 | 0.82 | 0.72 | 0.95 |
| | (0.09) | (0.20) | (0.15) | (0.11) | (0.14) | (0.28) | (0.17) | (0.32) | (0.08) |
| 1250 | 0.85 | 0.66 | 0.87 | 0.92 | 0.93 | 0.78 | 0.96 | 0.90 | 0.99 |
| | (0.07) | (0.23) | (0.13) | (0.08) | (0.11) | (0.26) | (0.08) | (0.23) | (0.03) |
| 1500 | 0.90 | 0.78 | 0.93 | 0.95 | 0.95 | 0.94 | 1.00 | 0.98 | 1.00 |
| | (0.06) | (0.23) | (0.10) | (0.07) | (0.11) | (0.13) | (0.02) | (0.10) | (0.01) |

**Table 10** – *Hit rates and standard errors in the unbalanced condition with equal sample sizes in reference and focal group.*

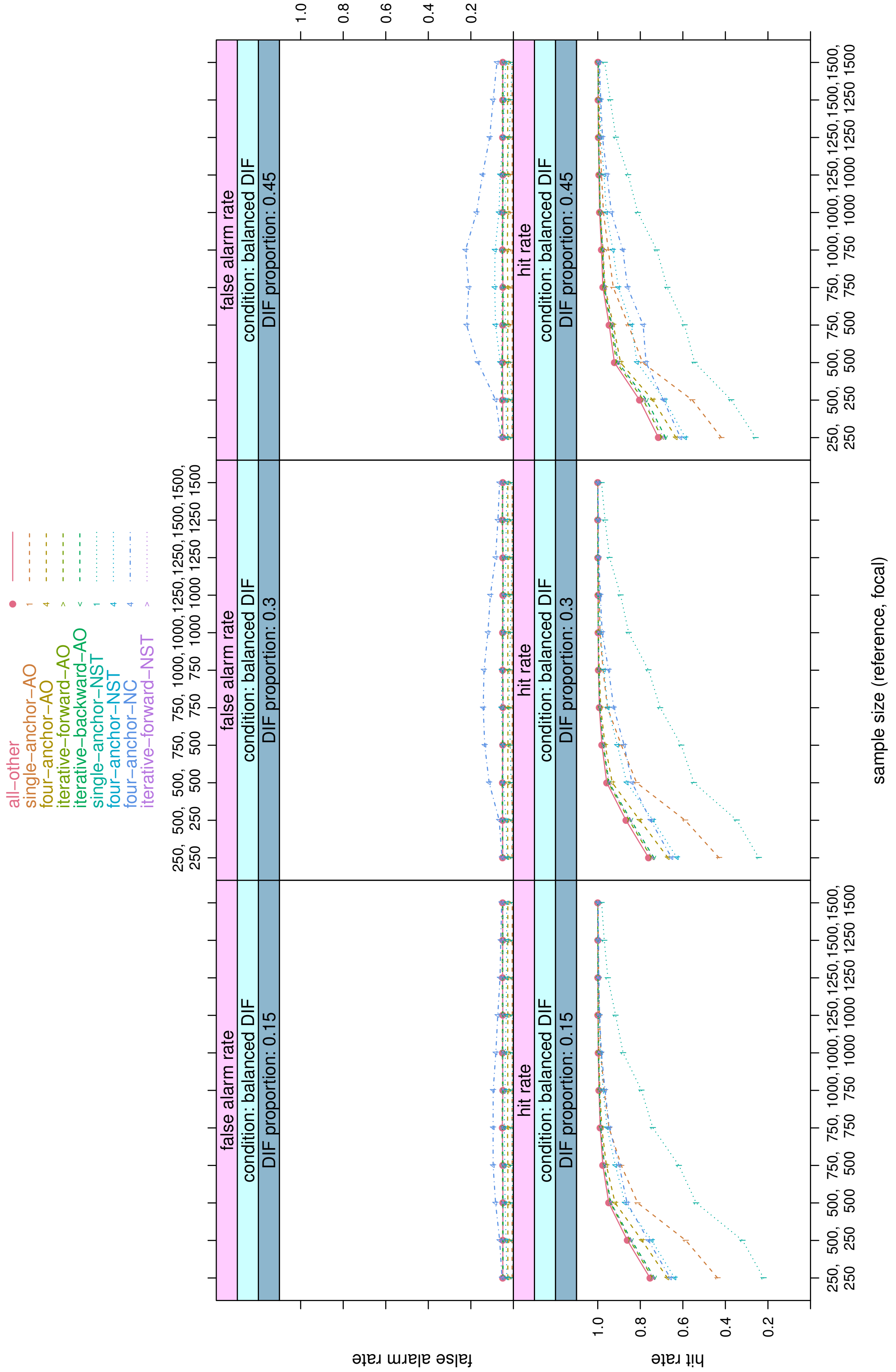

**Figure 26 –** *Balanced condition:* 15%, 30% *and* 45% *DIF-items with no systematic advantage for one group; sample size varies from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the balanced condition.*

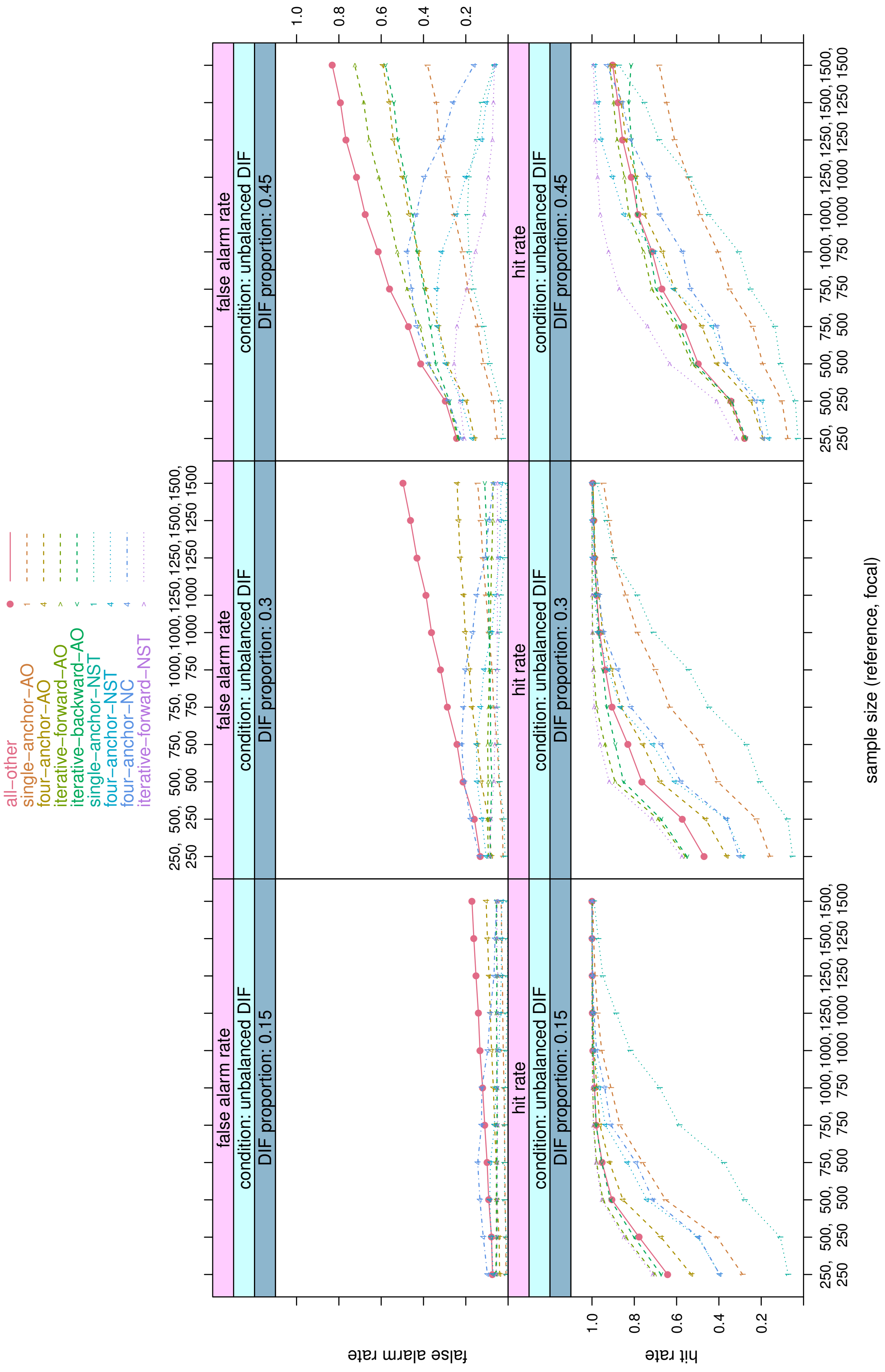

**Figure 27 –** *Unbalanced condition:* 15%, 30% *and* 45% *DIF-items favoring the reference group; sample size varies from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates in the unbalanced condition.*

# 7 Anchor selection strategies for DIF analysis in the Rasch model

***Summary:*** *Chapter 5 illustrated that a common metric for the item parameters of the groups that are to be compared (e.g. for the reference and the focal group) is necessary for DIF analysis in the Rasch model. Therefore, the same linear restriction is imposed in both groups. Items in the restriction are termed the anchor items. Ideally, these items are DIF-free or display a low proportion of DIF-items to avoid artificially augmented false alarm rates as discussed in Chapter 6. The question how DIF-free anchor items are selected appropriately remains a major challenge. Furthermore, various authors point out the lack of new anchor selection strategies and the lack of a comprehensive study especially for dichotomous IRT models. This chapter reviews existing anchor selection strategies that do not require any knowledge prior to DIF analysis, offers a straightforward notation and proposes three new anchor selection strategies. An extensive simulation study is conducted to compare the performance of the anchor selection strategies. The results show that an appropriate anchor selection is crucial for suitable DIF analysis. The newly suggested anchor selection strategies outperform the existing strategies and can reliably locate a suitable anchor when the sample sizes are large enough.*

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

## 7.1 Introduction

DIF tests based on item response theory (IRT) such as the item-wise Wald test (see, e.g., Glas and Verhelst, 1995) rely on the comparison of the estimated item parameters of the underlying IRT model. For this purpose, *anchor methods* are employed to place the estimated item parameters onto a common scale (see Chapter 5).

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

In the literature, one can find both methods that do and methods that don't require an explicit anchor selection. While at first sight it may seem that methods that do not require an anchor selection strategy have an advantage, it has been shown that there are situations where these methods are not suitable for DIF detection. The *all-other anchor method*, for example, uses all items except for the currently studied item as anchor (see, e.g., Cohen *et al.*, 1996; Kim and Cohen, 1998) and, thus, requires no anchor selection strategy. However, the method was shown to be inadvisable for DIF detection when the test contains DIF-items that favor one group (Wang and Yeh, 2003; Wang, 2004). Excluding DIF-items from the anchor by using iterative steps does not solve the problem when the test contains many DIF-items (Wang *et al.*, 2012).

In practice, there is usually no prior knowledge about the exact composition of the DIF effects and, thus, it is advisable to use an anchor method that relies on an objective anchor selection strategy such as an anchor of the constant length of four items (used, e.g., by Thissen *et al.* 1988; Wang 2004; Shih and Wang 2009). An anchor selection strategy then guides the decision which particular items are used as anchor items.

Several anchor selection strategies have already been proposed. Here, only those strategies that do not require any information prior to data analysis, such as the knowledge of certain DIF-free items, will be reviewed and presented in a straightforward notation in Section 7.2.3. The reason for excluding strategies that require prior knowledge about DIF-free items from this review is that in practical testing situations sets of truly DIF-free items are most likely unknown and even the judgment of content experts is unreliable (Frederickx *et al.*, 2010). New suggestions of anchor selection strategies are often only compared to few alternative strategies or in situations of only a limited range of the sample size and "*have not been exhaustively compared for the dichotomous case*" (González-Betanzos and Abad, 2012, p. 2). Therefore, in this chapter, we systematically evaluate the performance of the existing anchor selection strategies for DIF analysis in the Rasch model by conducting an extensive simulation study.

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

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

The chapter is organized as follows. A brief summary of the technical aspects of the anchor process in the Rasch model as well as details of the anchor classes, of the existing and of the newly suggested anchor selection strategies are given in Section 7.2. The simulation design is addressed in Section 7.3 and the results are discussed in Section 7.4. A concluding summary and practical recommendations are presented in Section 7.5.

## 7.2 Anchor methods

In this section, the anchor process is briefly summarized (for details see again Chapter 5). After that, anchor classes and existing anchor selection strategies are reviewed and new anchor selection strategies are proposed. We, again, focus on the Rasch model where the item parameter vector is $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$ (where $k$ denotes the number of items in the test). It is estimated using the CML approach similar to Chapter 6. To solve the *scale indeterminacy* of the item parameters, we impose the restriction $\tilde{\beta}_1 = 0$. The corresponding covariance matrix $\widehat{\text{Var}}(\tilde{\beta})$ then contains zero entries in the first row and in the first column.

In accordance to Chapter 5, other restrictions, where the sum of a selection of items is set to zero, are built using an indicator vector $a$ that indicates those elements $a_\ell$ that are included in the restriction with entries of ones and contains zero entries otherwise (e.g., $a = (1, 1, 0, 0, 0, \ldots)^\top$ including item 1 and item 2). In order to obtain these other restrictions, the estimated item parameters and the covariance matrix can be transformed by using the equations

$$\begin{aligned} \hat{\beta} &= A\tilde{\beta} && (23) \\ \text{and} \quad \widehat{\text{Var}}(\hat{\beta}) &= A\widehat{\text{Var}}(\tilde{\beta})A^\top, && (24) \end{aligned}$$

where $A = I_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot a^\top$ is a contrast matrix, $I_k$ denotes the identity matrix and $1_k$ denotes a vector of one entries of length $k$.

### 7.2.1 The anchor process

We focus here on the situation where two groups are compared: the reference and the focal group. In order to build a common scale for the item parameters, the same linear restriction

$$\sum_{\ell=1}^{k} a_\ell \hat{\beta}_\ell^g = \sum_{\ell \in \mathcal{A}^j} \hat{\beta}_\ell^g = 0 \tag{25}$$

is now imposed in each group $g$ (Glas and Verhelst, 1995). $\mathcal{A}^j \subseteq \{1, \ldots, k\}$ denotes the set of *anchor items*. Note that the set of items may depend on the item $j$ currently investigated for DIF. When the *all-other method* is used, all items except for the currently studied item $j$ are used as anchor items. The anchor set is then denoted as $\mathcal{A}^{\bar{j}} = \{1, \ldots, k\} \setminus j$. In contrast to this, when e.g. four anchor items are selected as the anchor set, the anchor is independent of the currently studied item, may include it and is denoted as $\mathcal{A}$. If the anchor set contains only a single anchor item $\ell$, the anchor set is denoted as $a_\ell$.

The estimated item parameters $\hat{\beta} = \hat{\beta}(\mathcal{A}^j)$ depend on the choice of the anchor $\mathcal{A}^j$ (what will be suppressed in the following notation) and can again be obtained by transformation (using equation 23). The item parameter estimates are then used to calculate DIF indices or test statistics $T_j = T_j(\mathcal{A}^j)$ which also depend on the anchor. In this chapter, we, again, focus on the item-wise Wald test to assess DIF in item $j$ between the reference (ref) and the focal (foc) group that display equal item parameters under the null hypothesis of no DIF $H_0 : \beta_j^{\text{ref}} = \beta_j^{\text{foc}}$ (for more details, again, Chapter 5 and the references therein).

The results of the DIF analysis strongly depend on the choice of the anchor items, as previous studies illustrated. If the anchor contains at least one DIF-item, it is referred to as *contaminated* (see, e.g., Finch, 2005; Woods, 2009; Wang *et al.*, 2012). The scales may be artificially shifted apart and the false alarm rates of the DIF tests may be seriously inflated (see, e.g., Wang and Yeh 2003, Wang 2004, Wang and Su 2004, Finch 2005, Stark *et al.* 2006, Woods 2009). Instructive examples that illustrate the artificial scale shift are provided by Wang (2004) and in Chapter 5 of this thesis.

### 7.2.2 Anchor classes

For distinguishing between the different approaches, we employ the framework for anchor methods introduced in Chapter 5 where the *anchor class* determines characteristics of the anchor methods, such as a predefined anchor length, and the *anchor selection strategy* guides the decision which items are used as anchor items. The combination of an anchor class together with an anchor selection strategy is then termed an *anchor method*. Different anchor classes are now briefly reviewed.

The *constant anchor class* consists of an anchor with a predefined, constant length. Usually, it is claimed that a constant anchor of four items assures sufficient power (cf. e.g., Shih and Wang, 2009; Wang *et al.*, 2012). An anchor selection strategy is needed to guide the decision which items are used as anchor items. The *all-other anchor class* uses all items except for the currently studied item as anchor and the *equal-mean difficulty anchor class* uses all items as anchor (see, e.g., Wang, 2004, and the references therein). These latter two anchor classes do not require an additional anchor selection strategy. Furthermore, iterative anchor classes build the anchor in an iterative manner. The *iterative backward class* (used, e.g., by Drasgow 1987; Candell and Drasgow 1988; Hidalgo-Montesinos and Lopez-Pina 2002) starts with all other items as anchor and excludes DIF-items from the anchor, whereas the *iterative forward anchor class* starts with a single anchor item and then, iteratively, includes items in the anchor (see Chapter 6).

Wang (2004) compared the *constant anchor class* implemented with 1, 2, 4, 10 and all DIF-free items as anchor to the *all-other* and the *equal-mean difficulty anchor class* for the Rasch model and other models and found appropriate results for DIF detection with the *constant anchor class* in any condition as opposed to the *all-other* and the *equal-mean difficulty anchor class*. When the anchor length of the DIF-free constant anchor was increased, the power increased as well. However, note that in reality it is unknown which items are DIF-free and the anchor may become contaminated. Both for the *all-other* and the *equal-mean difficulty anchor classes*, the type one error rate was seriously inflated when the direction of DIF was unbalanced (i.e. the DIF effects did not cancel out between groups and one group was favored in the test). This is of utmost importance for practical testing situations since items truly free of DIF display artificial DIF and may be eliminated by mistake. Since in Wang's study the anchor items for the *constant anchor class* were DIF-free by definition, it is yet to be explained how DIF-free anchor items can be selected appropriately in practice.

Wang and Yeh (2003) compared the *all-other anchor class* with different anchor methods from the *constant anchor class* including 1, 4 or 10 DIF-free anchor items. The analysis was carried out for the two-parameter logistic model, the three-parameter logistic model and the Graded Response model. The authors again found the direction of DIF (together with the proportion of unbalanced DIF-items) to strongly affect the results of the DIF analysis using the *all-other class*. Since the latter was only suited for balanced DIF, the authors discourage from the application of the *all-other class* since it is doubtful whether the situation of balanced DIF is met in practice (Wang and Yeh, 2003; Wang *et al.*, 2012). The *constant anchor class* yielded well-controlled type one error rates. Four anchor items were found to assure sufficient power in the condition of 1000 observations in each group similar to the findings of Shih and Wang (2009) and Wang *et al.* (2012). Longer constant anchors achieved higher power but the location of a DIF-free anchor is harder for longer anchors, too.

Addressing problems of different anchor methods, González-Betanzos and Abad (2012) investigated modified two-step versions of the *constant single anchor class* and the *all-other class*. For the former, the anchor set is formed by all items that were found DIF-free with a pure single anchor item, while for the latter, items displaying DIF using all other items as anchor were excluded from the anchor set. These procedures were compared with pure single anchor items – that differed in their discrimination and difficulty parameters – and the *all-other class*. Again, the *all-other class* was inadvisable but all remaining methods displayed well-controlled type one error rates and a high power when the effect of DIF was large. The two-step extensions then found DIF-free anchor sets in the majority of the settings.

All three studies showed that the direction of DIF has a major impact on the results of the DIF analysis for the *all-other* and the *equal-mean difficulty anchor class* as opposed to the *constant anchor class*. The latter displays appropriate false alarm rates when the anchor items are DIF-free. Thus, the *constant anchor class* is in principle able to yield appropriate results for the DIF analysis when the anchor is DIF-free. However, since Wang and Yeh (2003), Wang (2004) and González-Betanzos and Abad (2012) used prior knowledge of the set of DIF-free items (that is unknown in real DIF analysis – as opposed to simulation analysis, where the true DIF pattern is known) to select the constant anchor items, no information how well anchor selection strategies without prior knowledge perform is yet available and "*[f]urther research is needed to investigate how to locate anchor items correctly and efficiently*" (Wang and Yeh, 2003, p. 496).

Another anchor class was suggested in Chapter 6 of this thesis. Instead of a predefined anchor length, the *iterative forward anchor class* builds the anchor in a step-by-step procedure. First, one anchor item is used for the initial DIF test. As long as the current anchor length is shorter than the number of currently presumed DIF-free items, one item is added to the current anchor and DIF analysis is conducted using the new current anchor. The sequence which items are added to the anchor is determined by the anchor selection strategy. In the simulation study in Chapter 6, the *iterative forward anchor class* and the *constant anchor class* were combined with two different anchor selection strategies and compared to the *all-other class* and the *iterative backward anchor class* (that initiates with all other items as anchor and successively excludes items from the anchor). The *iterative forward anchor class* was found to be superior since it yielded high hit rates and, simultaneously, low false alarm rates when the sample size was large in any studied condition of balanced or unbalanced DIF if the *number of significant threshold anchor selection strategy* (see Section 7.2.3) was employed.

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

### 7.2.3 Anchor selection strategies

Several authors have proposed anchor selection strategies, some of which rely on prior knowledge of a set of DIF-free items or on the advise of content experts (for a literature overview where this approach fails see Frederickx *et al.*, 2010), while others are based on preliminary item analysis (for a overview see Wang, 2004). This means that DIF tests are conducted to locate (ideally DIF-free) anchor items and has been referred to as the DIF-free-then-DIF strategy by Wang *et al.* (2012). We focus here on anchor selection strategies based on preliminary item analysis that do not require any prior information of DIF-free items since this will be most common in practice.

The ranking order resulting from the anchor selection strategies is used within the anchor classes to conduct the final DIF analysis. For the *constant four anchor class*, the items with the lowest four ranks are selected as the final anchor set $\mathcal{A}_{\text{final}}$. Then, the DIF tests are carried out using the anchor set $\mathcal{A}_{\text{final}}$. Since $k - 1$ parameters are free in the estimation, only $k - 1$ estimated standard errors result (Molenaar, 1995a), the k-th standard error is determined by the restriction and, hence, only $k - 1$ tests can be carried out. To overcome the problem that the classification of an item as a DIF- or a DIF-free item is intended for each of the $k$ items, we classify the first final anchor item with the lowest rank to be DIF-free – a decision that may be false if even the item with the lowest rank does indeed have DIF, but in this case this would be noticeable in the final test results. Note that this decision is by no means as drastic as choosing the anchor items only from the set of items that are known to be DIF-free in a simulation, as was done by Wang and Yeh (2003) and Wang (2004) but cannot be done in any real study where the true DIF- and DIF-free items are unknown.

For the *iterative forward anchor class*, items are selected into the anchor as long as the anchor is shorter than the number of currently presumed DIF-free items. In this anchor class, anchor items are selected in a step-by-step procedure following the ranking order that results from the anchor selection. When the stopping criterion is reached, the final anchor set $\mathcal{A}_{\text{final}}$ is found. The final DIF analysis is carried out but, again, the first anchor item is classified DIF-free to account for the fact that only $k-1$ tests can be carried out.

In the following, first, the strategies that are built on DIF tests using all other items as anchor are reviewed. Second, selection strategies that rely on DIF tests with every item acting as constant single anchor item for every other item are discussed. Third, three new strategies are suggested at the end of the section. Note that the DIF tests mentioned in the next paragraphs are only used as preliminary steps to assess the ranking order of candidate anchor items for the final DIF analysis.

*All-other selection*

The *all-other selection strategy* (AO-selection) was proposed by Woods (2009) as discussed in Chapter 5. Every item is tested for DIF using all remaining items as anchor items, yielding the observed test statistic $t_j(\mathcal{A}^{\bar{j}})$ for every currently studied item $j$ ($j = 1, \ldots, k$) with the all-other anchor set $\mathcal{A}^{\bar{j}} = \{1, \ldots, k\} \setminus j$. A predefined number of anchor items is then chosen according to the lowest ranks of the resulting absolute DIF test statistics:

$$\operatorname{rank}\left(|t_j(\mathcal{A}^{\bar{j}})|\right).$$

The item displaying the lowest rank is the first candidate anchor item, whereas the item corresponding to the highest rank is the last candidate anchor item. (Note that, originally, Woods (2009) suggested to use the ratios of the test statistics and the degrees of freedom that may vary across items if the items display a different number of response categories. However, this is not discussed here since the responses are always dichotomous in the Rasch model.)

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

*All-other purified selection*

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

$$\operatorname{rank}\left(|t_j(\mathcal{A}_{\text{purified}})|\right)$$

(or if necessary the lowest ranks of the ratios of the test statistics and the degrees of freedom).

In a simulation study, Wang *et al.* (2012) found the modified AOP-selection to be superior to the AO-selection since both methods displayed comparable results when DIF was balanced but the AOP-selection yielded more often a DIF-free anchor set when DIF was unbalanced. Still, there were conditions where the proportions of replications yielding a DIF-free anchor set were far away from 100%, e.g. 14% for the AO- and 20% for the AOP-selection when the sample size in each group was 250 in a test with 40% of 40 items favoring the reference group and when ability differences were present.

*Number of significant threshold selection*

An anchor selection strategy that is a simplified version of the proposition of Wang (2004) is here called *number of significant threshold* (NST) selection strategy. Every item $j = 1, \ldots, k$ is tested for DIF using every other item $\ell \neq j$ as constant single anchor, yielding $k \cdot (k-1)$ test statistics $t_j(a_\ell)$ with the corresponding p-values $p_j(a_\ell)$. Note that the test statistics and p-values display the following symmetry properties $|t_j(a_\ell)| = |t_\ell(a_j)|$ and $p_j(a_\ell) = p_\ell(a_j)$ since the constant scale shift of one single anchor item is reflected in the test statistic of the item currently investigated and vice versa. The number of significant DIF tests defines the ranking order of the candidate anchor items

$$\text{rank}\left( \sum_{\ell \in \{1,\ldots,k\} \setminus j} \mathbb{1}\left\{ p_j(a_\ell) \leq \alpha \right\} \right)$$

that is written as the number of p-values that do not exceed the threshold 0.05. Thus, the item displaying the least number of significant DIF tests is chosen as the first anchor item. If more than one item displays the same number of significant results, one of the corresponding items is selected randomly.

Furthermore, Wang (2004) originally suggested the next candidate (NC) modification: The item that was selected by the NST-selection strategy functions as the current single anchor item and DIF tests are again carried out (see Wang, 2004, p. 249) where one DIF test statistic results for every item except for the anchor item. The candidate is then included in the anchor if it displays "*the least magnitude*" (Wang, 2004, p. 250) of (non-significant) DIF and the steps are repeated until either the pre-defined anchor length (of e.g. four items) is reached or the candidate item displays significant DIF. Since in Chapter 6 the NST-selection was found superior to the original NC-strategy, only the former is investigated in this chapter.

*Mean test statistic selection*

In the study of Shih and Wang (2009), first, pure constant anchors of 1, 2, 4 and 10 items were employed for DIF tests that yielded well-controlled type one error rates in any condition with varying sample sizes, DIF proportions, DIF directions, test lengths and underlying IRT models. To reach an ideally pure set of anchor items, the authors introduced the following anchor selection procedure: For every item, a DIF test statistic is calculated using every other item once as single anchor, yielding again $k \cdot (k-1)$ test statistics $t_j(a_\ell)$. After that, every item is assigned the mean absolute DIF test statistic and the predefined number of anchor items is selected according to the lowest ranks

$$\operatorname{rank}\left(\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\setminus j}\left|t_j(a_\ell)\right|\right).$$

We abbreviate this method MT-selection (for *mean test statistic selection*). Shih and Wang (2009) found high rates of correctly locating one or four DIF-free anchor items when the sample size was high.

*Mean p-value selection*

We propose three alternative anchor selection strategies. First, we suggest an idea similar to the MT-strategy of Shih and Wang (2009). Instead of the lowest mean absolute DIF test statistic $t_j(a_\ell)$ resulting from tests for every item using every remaining item as constant single anchor, items are here chosen that display the highest mean p-value $p_j(a_\ell)$ from the resulting significance tests and, thus, (for easier comparability with the previous methods) the lowest rank of the negative mean p-values

$$\operatorname{rank}\left(-\frac{1}{k-1}\sum_{\ell\in\{1,\ldots,k\}\setminus j}p_j(a_\ell)\right).$$

This strategy is abbreviated MP-strategy (for *mean p-value selection*). Even though the p-values represent a monotone decreasing transformation of the absolute test statistics, the means of both measures may yield different ranking orders.

The next two suggestions were inspired by the threshold approach of the NST-selection of Wang (2004). The author proposed to choose those items as anchor items that display the least number of significant DIF test results. In Chapter 6 is was shown that this strategy was superior to the AO-selection when the DIF direction was unknown prior to data analysis. The major drawback using the NST-selection was that it was strongly affected by the sample size. The reason for this is that the selection is based on the decisions of statistical significance tests which are strongly influenced by the sample size. Therefore, the next two newly suggested anchor selection strategies rely on a different criterion and both methods assume – similar to the MT- and the MP-selection – that the majority of items is DIF-free, an assumption that is often found in the construction of anchor or DIF methods (see, e.g., Shih and Wang, 2009; Magis and De Boeck, 2011).

*Mean test statistic threshold selection*

For every item the absolute mean of the test statistics $t_j(a_\ell)$ resulting from tests for every item using every other item as single anchor is calculated and the resulting values are ordered. The threshold for the MTT-selection (for *mean test statistic threshold*) is the ($\lceil 0.5 \cdot k\rceil$)-th ordered value, which is indicated by the index in parenthesis, for an even number of items or the next larger whole number in case of an odd number of items (indicated by the ceiling function $\lceil\,\rceil$).

The number of absolute test statistics exceeding this threshold determines the ranking order of candidate anchor items:

$$\operatorname{rank}\left(\sum_{\ell\in\{1,\dots,k\}\setminus j} \mathbb{1}\left\{\left|t_j(a_\ell)\right| > \left(\left|\frac{1}{k-1}\sum_{\ell\in\{1,\dots,k\}\setminus j} t_j(a_\ell)\right|\right)_{(\lceil 0.5\cdot k\rceil)}\right\}\right).$$

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

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

*Mean p-value threshold selection*

Similar to the MTT-selection, the threshold of the MPT-selection (for *mean p-value threshold*) relies again on DIF tests for every item using every other item as constant single anchor. Now, the ($\lceil 0.5 \cdot k\rceil$)-th ordered (from large to small) value (or the next larger whole number) of the mean of the resulting p-values $p_j(a_\ell)$ is used as the threshold. The ranking order is defined by the number of tests per item that yields p-values exceeding the threshold p-value

$$\operatorname{rank}\left(-\sum_{\ell\in\{1,\dots,k\}\setminus j} \mathbb{1}\left\{p_j(a_\ell) > \left(\frac{1}{k-1}\sum_{\ell\in\{1,\dots,k\}\setminus j} p_j(a_\ell)\right)_{(\lceil 0.5\cdot k\rceil)}\right\}\right).$$

We, again, assume the classification of DIF-free anchor items to be more accurate compared to the MP-selection, since the individual test information is used and not only the mere mean values. The method is again relying on the assumption that the majority of items is DIF-free and designed for both situations, balanced and unbalanced DIF, and all sample sizes.

In summary, the newly suggested methods are developed for balanced and unbalanced DIF situations and should, thus, outperform not only the AO-selection that initiates with the potentially biased DIF test results using the all-other method, but also the AOP-selection that may not be able to exclude all DIF-items from the anchor set when the proportion of DIF-items is high (Wang *et al.*, 2012). In comparison with the NST-selection, which uses the binary decisions of the significance tests (Woods, 2009), the newly suggested methods should be less affected by sample size.

While the MT- and the MP-selection use mere mean values, the MPT- and the MTT-selection use all individual test results and are, therefore, expected to better distinguish between DIF- and DIF-free anchor items. By employing a threshold, the new methods should select those items as anchor that display little artificial DIF which can be caused by contamination (see, e.g., Finch, 2005; Woods, 2009) or by random sampling fluctuation (Kopf *et al.*, 2013b).

## 7.3 Simulation study

In order to evaluate the performance of the newly suggested anchor selection strategies, we conducted an extensive simulation study in the free R system for statistical computing (R Development Core Team, 2011).

Parts of the simulation design were inspired by the settings used by Wang *et al.* (2012). Each setting from the simulation study is replicated 1000 times to ensure reliable results.

### 7.3.1 Data generating processes

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

- *Test characteristics*

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

- *IRT model*

  The responses follow again the Rasch model

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

  with the difficulty parameters $\beta$ = (-2.522, -1.902, -1.351, -1.092, -0.234, -0.317, 0.037, 0.268, -0.571, 0.317, 0.295, 0.778, 1.514, 1.744, 1.951, -1.152, -0.526, 1.104, 0.961, 1.314, -2.198, -1.621, -0.761, -1.179, -0.610, -0.291, 0.067, 0.706, -2.713, 0.213, 0.116, 0.273, 0.840, 0.745, 1.485, -1.208, 0.189, 0.345, 0.962, 1.592) used by Wang *et al.* (2012).

The first 10, 20, 30, 40 or 45 percent of the items (cf. paragraph DIF proportion in the next section) are simulated as the DIF-items.

- *Ability distribution*

  The ability parameters $\theta_i$ follows a standard normal distribution for the reference group $\theta^{\mathrm{ref}} \sim N(0,1)$ and a normal distribution with a lower mean for the focal group $\theta^{\mathrm{foc}} \sim N(-1,1)$. Thus, ability differences are present in this simulation study.

### 7.3.2 Manipulated variables

In addition to the anchor selections investigated by Wang *et al.* (2012), namely the AO- and the AOP-selection, five other anchor selection strategies and also the perfect selection of DIF-free items that serves as a benchmark method are included (for a summary see Table 11).

- *Sample size*

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

- *DIF proportion, DIF magnitude and DIF direction*

  The *proportion* of simulated DIF-items is set to $\%DIF \in \{0, 0.1, 0.2, 0.3, 0.4, 0.45\}$. In order to save space, we will not discuss all DIF proportions in detail, but only the most challenging situation with 45% DIF-items. An outlook on the remaining results with other DIF proportions is presented in Section 7.4.5.

  If an item $j$ is a simulated DIF-item, the *magnitude* of DIF is set to the constant value of $\Delta_{\mathrm{DIF}} = \beta_{\mathrm{ref}} - \beta_{\mathrm{foc}} = \pm 0.6$ in the main simulation study. An outlook on results when other magnitudes are present is given in Section 7.4.5.

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

- *Anchor methods*

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

  *Anchor selections*: Eight different anchor selection strategies (for a brief summary see Table 11) are compared across the simulated settings: The **AO**-, **AOP**[11]-, **NST**- and **MT**-

[11] In case all items were excluded from the anchor in the initial step (which happened in only 2 out of 121,000 replications), here, one single anchor item was chosen using the AO-strategy. When the constant four anchor class was combined with the AO-strategy, four anchor items were selected according to the lowest ranks of the resulting DIF test statistics using the AO-selected anchor item. Similarly, when the iterative forward anchor class was investigated, the ranking order was then built from the resulting DIF test statistics using the AO-selected anchor item and the iterative procedure was conducted normally.

selection as well as the newly suggested **MP**-, **MTT**- and **MPT**-selection and the **perfect**-selection that serves as the benchmark condition: The perfect selection for the *four anchor class* includes four randomly chosen DIF-free items. For the *iterative forward anchor class*, a random ranking order that includes the DIF-free items first, followed by the DIF-items is handed to the procedure. The remaining steps of the iterative procedure are carried out as usual. Thus, for the *'perfect' forward method*, it may happen that DIF-items occur in the anchor because the length of the iteratively selected anchor may exceed the length of the sequence of DIF-free items, which is not the case for the *perfect four anchor method*.

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

### 7.3.3 Outcome variables

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

- *False alarm rate*

  For a single replication the *false alarm rate* is defined as the proportion of DIF-free items that are (erroneously) diagnosed with DIF in the final DIF test. The estimated *false alarm rate* for each simulated setting is computed as the mean over all 1000 replications and, hence, represents the type one error rate of the final DIF test.

- *Hit rate*

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

## 7.4 Results

In the following, we restrict the presentation of the main simulation study to the extreme condition where 45% of the items display DIF. The reason for this is that large proportions of DIF-items may indeed occur in practical testing situations (examples can be found in Shih and Wang 2009, p. 186 and Allalouf *et al.* 1999, p. 185) and the anchor selection strategies should be compared in a situation where the classification of DIF- and DIF-free items is rather challenging.

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

**Table 11 –** *A short summary of the anchor selection strategies that are investigated in this chapter.*

### 7.4.1 Anchor selection for the constant four anchor class

In this section, the anchor selection strategies combined with the *constant anchor class* are regarded. Thus, four anchor items were selected by the respective strategy and the results of the final DIF tests are discussed. Figure 28 contains the results of the false alarm rate (top row) and the hit rate (bottom row) in case of 45% DIF-items that did not systematically favor one group (balanced DIF pattern, left column) or that systematically favored the reference group (unbalanced DIF pattern, right column).

In the balanced condition, almost all anchor methods displayed false alarm rates holding the 5% level in the observed range of the sample size. The only exception was the method relying on the NST-selection with the maximum observed false alarm rate of 0.09 that occurred at the sample size of 750 observations in each group. The corresponding false alarm rate for the NST-selection displayed an inversely u-shaped pattern that was also found and discussed in detail in Chapter 6. The perfect selection was near the significance level, while the remaining methods (except for the constant4-NST method) stayed below 0.05.

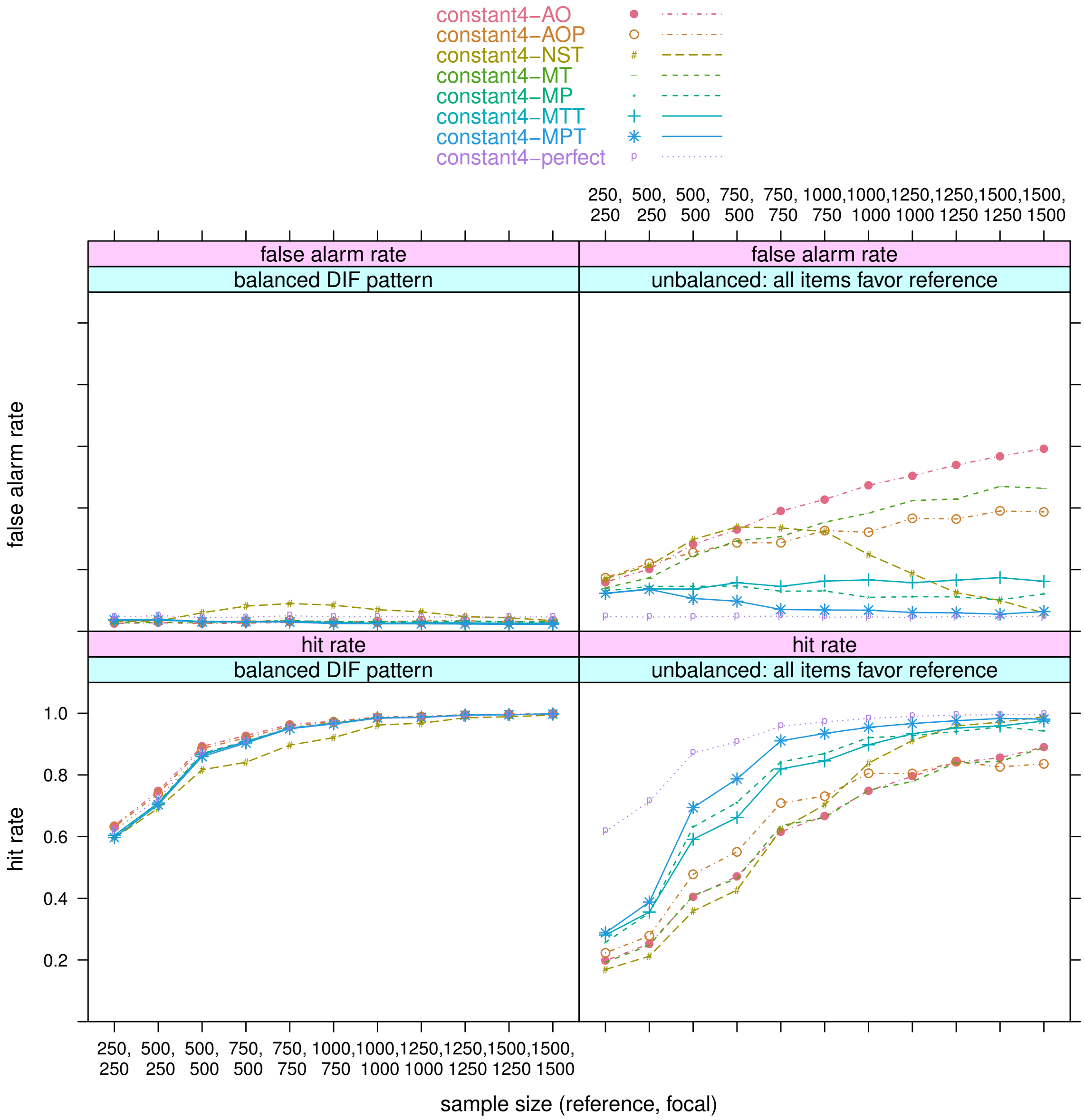

**Figure 28 –** *Balanced condition:* 45% *DIF-items with no systematic advantage for one group; unbalanced condition:* 45% *DIF-items favoring the reference group; sample size varied from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates.*

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

In the unbalanced condition, the anchor selections strongly differed regarding their ability to select four anchor items that yielded appropriate final DIF tests. Both selections based on DIF tests using all other items as anchor were not appropriate for the unbalanced DIF condition: The

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

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

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

### 7.4.2 Anchor selection for the iterative forward anchor class

In the next section, we investigate the combination of the anchor selection strategies with the *iterative forward anchor class* that was designed to specify a longer anchor (see Chapter 6). Similar to Section 7.4.1, Figure 29 includes the results for the false alarm rate (top row) and the hit rate (bottom row) in case of 45% DIF-items that did not systematically favor one group (balanced DIF pattern, left column) or that systematically favored the reference group (unbalanced DIF pattern, right column).

In case of balanced DIF, there was neither a visible difference in the false alarm rates nor in the

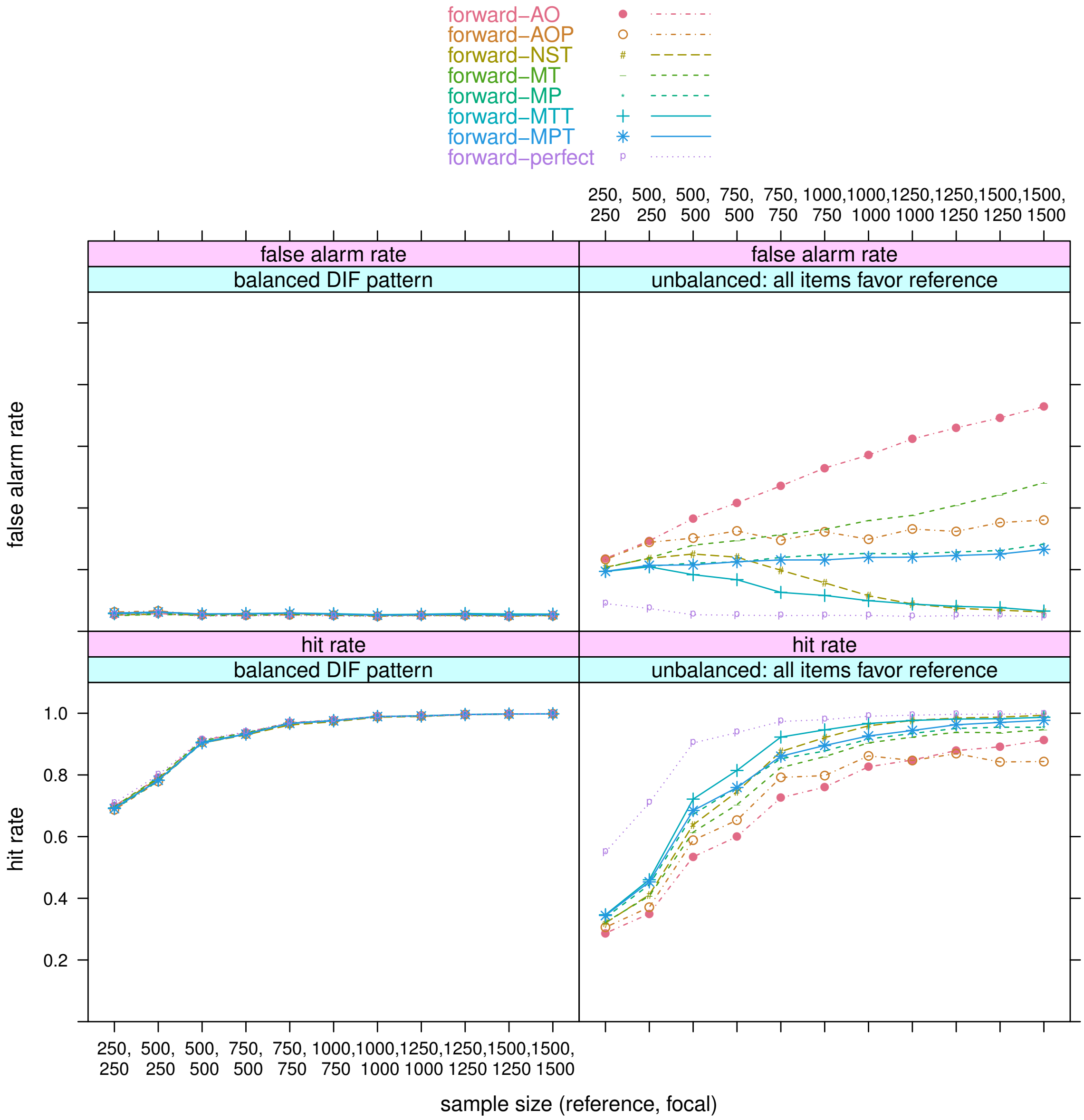

**Figure 29 –** *Balanced condition:* 45% *DIF-items with no systematic advantage for one group; unbalanced condition:* 45% *DIF-items favoring the reference group; sample size varied from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates.*

hit rates for any of the investigated methods. Again, all selections yielded test results similar to the iterative method based on the perfect anchor. Hence, all selection strategies were advisable and the *iterative forward anchor class* was robust against the anchor selection strategy employed in this case.

In contrast to this, the results of the final DIF tests varied notably with the anchor selection strategies when DIF was unbalanced. The largest false alarm rates occurred for the AO-, AOP- and also the MT-method that were again not suited to locate an anchor in the unbalanced condition. Their false alarm rates even increased with the sample size. The NST-selection was more appropriate in selecting the longer iterative anchor since it achieved a lower false alarm rate in regions of medium to large sample sizes. The MP- and the MPT-selection now yielded test results with very similar false alarm rates (with a slight advantage for the threshold method) that also clearly exceeded the significance level. The lowest false alarm rate among the empirical selection strategies occurred for the forward-MTT method. But there is still room for improvement as the perfect anchor selection still displays test results with a lower false alarm rate, especially in regions of small or medium sample sizes.

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

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

### 7.4.3 Comparison of the mean test statistic and p-value threshold selection

To explain the fact that the MPT-selection yielded better results when it was combined with the *constant four anchor class*, whereas the MTT-selection performed better combined with the *iterative forward anchor class*, the ranking order of candidate anchor items is now regarded in detail for one balanced and one unbalanced setting (again with 45% DIF-items and 1000 observations in each group). Figure 30 contains the proportions of DIF-items in the ranking order of candidate anchor items. In the regarded setting, 22 items were DIF-free and, ideally, the 22 lowest ranks (from left to the vertical line) should display low proportions of DIF-items.

In the balanced condition (Figure 30, top panel), the first items of the sequence of anchor candidates – i.e. the items up to the vertical line – displayed low proportions of DIF-items over the simulation runs for both the MPT-selection (black bars) and the MTT-selection (gray bars). In contrast to this, the items that were assigned the highest ranks – i.e. the items to the right of the vertical line – displayed large proportions of DIF-items. Thus, both anchor selection strategies yielded appropriate ranking orders that clearly separated DIF- and DIF-free items: The first candidates displayed low proportions of DIF-items, whereas the last candidates displayed large proportions of DIF-items as intended for all ranks above 22.

In the unbalanced condition (Figure 30, bottom panel), the separation of candidates with low

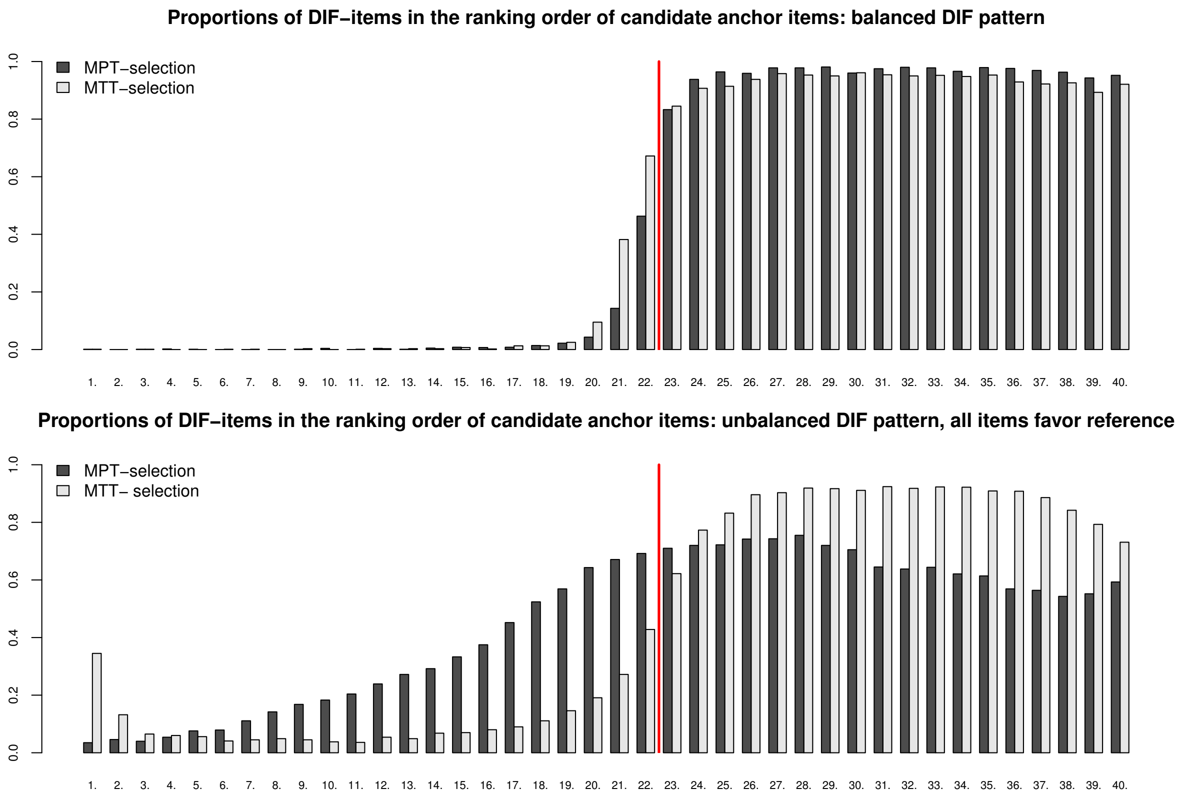

**Figure 30 –** *Top row: proportion of DIF-items in the ranking order of anchor candidates in the balanced condition:* 45% *DIF-items with no systematic advantage for one group; bottom row: proportion of DIF-items in the ranking order of anchor candidates in the unbalanced condition:* 45% *DIF-items favoring the reference group; sample size was set to* 1000 *observations in each group.*

proportions of DIF-items for the first ranks and high proportions for the last ranks was harder for both methods. Now the first anchor candidates displayed higher proportions of DIF-items. Generally, the MTT-selection (gray) yielded lower DIF-proportions for items up to the vertical line compared to the MPT-selection (black) and was, thus, better suited to locate a longer anchor. However, when an anchor of constant length four was intended, only the first four candidates were included in the anchor. The first four ranks selected by the MPT-selection displayed lower proportions of DIF-items items compared to the MTT-selection (see very left of Figure 30, bottom panel). Thus, the MPT-selection was better suited to locate four anchor items.

In summary, the first anchor candidates were more likely found from the set of DIF-free items by the MPT-selection, whereas the MTT-selection was better suited for longer anchors. The next section addresses whether one of the two best performing methods can be considered as superior.

### 7.4.4 Comparison of the best performing methods

In order to compare the results of the constant4-MPT method and of the forward-MTT method, we again restrict the analysis to 45% DIF-items (see Figure 31) that are either balanced (left column) or unbalanced (right column). The false alarm rates and the hit rates are now illustrated together with the empirical confidence intervals that are computed as the 2.5% and the 97.5% quantile from the 1000 replications.

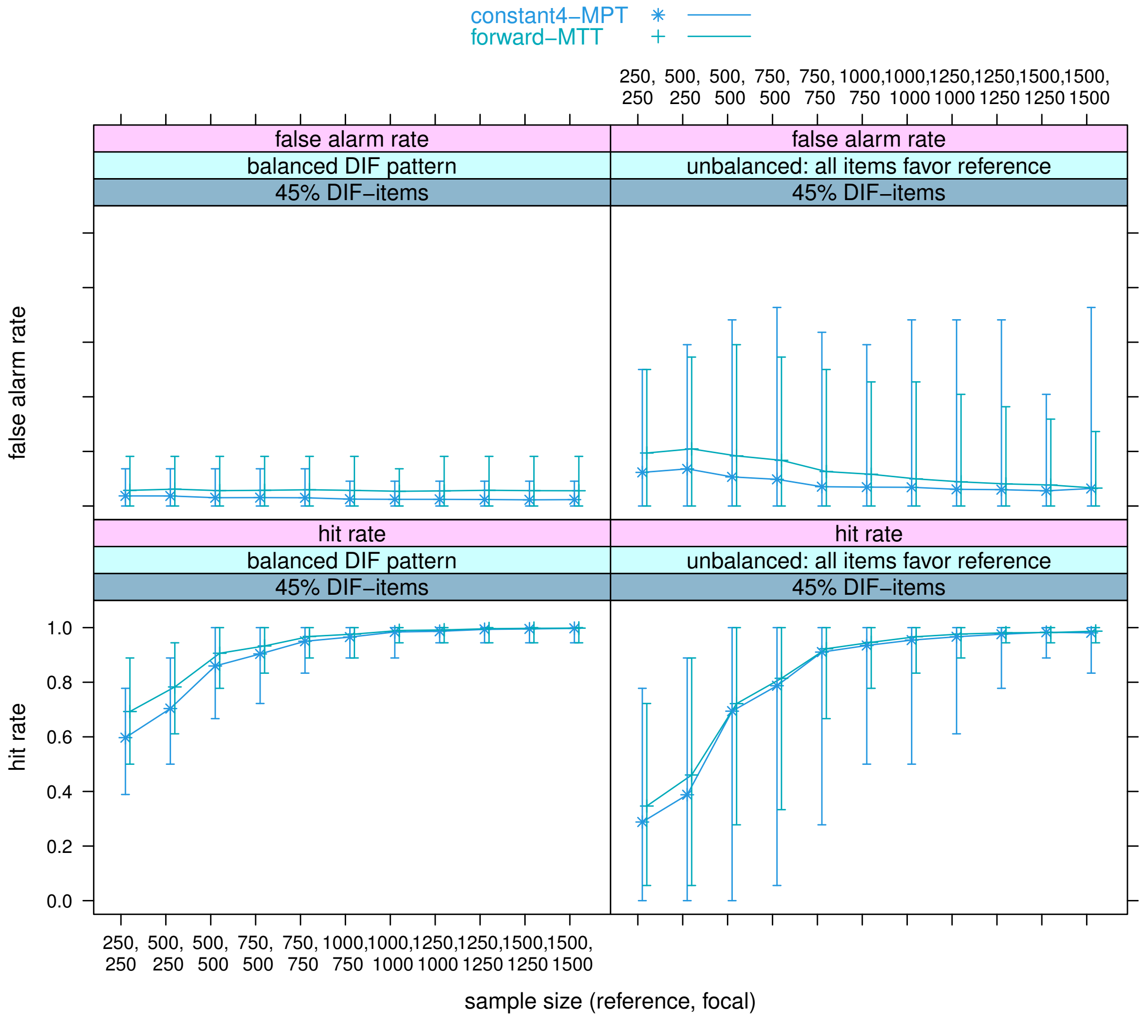

**Figure 31 –** *Balanced condition:* 45% *DIF-items with no systematic advantage for one group; unbalanced condition:* 45% *DIF-items favoring the reference group; sample size varies from* (250, 250) *up to* (1500, 1500)*; top row: false alarm rates; bottom row: hit rates.*

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

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

In summary, the comparison of the constant4-MPT and the forward-MTT method did not clearly show the superiority of one of those methods. Only when the sample size was large, the constant4-MPT method was superior regarding a lower and less fluctuating false alarm rate together with a comparably high hit rate in the balanced case and the forward-MTT method outperformed the constant4-MPT method by yielding a higher hit rate and more reliable results due to less fluctuation within the hit rate as well as within the false alarm rate in the unbalanced case. Section 7.4.3 showed that the first anchor candidates were more likely found from the set of DIF-free items by the MPT-selection, whereas the MTT-selection was better suited for longer anchors. Future research might compare both methods when a certain proportion of anchor candidates is excluded for the MTT-selection and the iterative forward anchor is only allowed to grow up to a certain proportion of currently presumed DIF-free items.

### 7.4.5 Further simulated settings

In this section, further simulated settings are summarized briefly. In addition to the settings where 45% of the items displayed DIF, also smaller proportions of DIF-items were regarded and the magnitude of DIF was also generated normally and from a sequence (see below) to avoid artificial results. All further manipulated variables influenced the results of the final DIF tests.

*Null hypothesis*

Under the null hypothesis of no DIF, the methods from the *constant four anchor class* combined with an empirical anchor selection strategy yielded false alarm rates between 0.02 and 0.04, whereas all methods from the *iterative forward class* together with the constant4-perfect method exhausted the significance level.

*Constant DIF magnitude and other DIF proportions*

Generally, the smaller the proportions of simulated DIF-items were, the weaker were the effects of the employed anchor selection strategy. Under the alternative, when the constant DIF magnitude was regarded together with different proportions of DIF-items in the balanced condition, again all *constant four anchor methods* held the 5%-level and displayed high hit rates. The only exception was the NST-selection that yielded the u-shaped patterns for the false alarm rate whose shape sharpened the larger the DIF proportion was. In the unbalanced condition with the constant DIF magnitude, the existing *constant four methods* were outperformed by the new suggestions when the DIF proportion was 30% or higher and the advantage – measured by a lower false alarm rate and a higher hit rate – of the new MP-, MPT- and MTT-selection increased with the proportion of DIF. Only with 40% of DIF-items or more, the MPT-selection had an visible advantage compared to the other new suggestions.

In accordance with the previous results from Section 7.4.2, the *iterative anchor class* was robust against the anchor selection strategy when DIF was balanced independent of the proportion of DIF-items. Similarly to the results of the *constant four anchor class*, the advantage for the forward-MTT method occurred when the unbalanced DIF proportion was 40% or higher.

*Further DIF magnitudes*

In addition to the constant DIF magnitude presented in the previous sections, we have replicated the main results of our simulation study with two other DIF magnitudes that also yielded a mean absolute difference in the DIF-item parameters of Mean($|\Delta_{\text{DIF}}|$) = 0.6. In contrast to the constant DIF magnitude, the DIF magnitude was generated normally $\pm\Delta_{\text{DIF}} \sim N(0.6, 0.01)$ and from the sequence $\{0.2, 0.4, \ldots, 1.0\}$ (abbreviated uniform) with the sign of $\Delta_{\text{DIF}}$ consistent with the intended direction of DIF.

When 45% balanced DIF-items were considered, there were no visible differences for the anchor selection strategies (except for the constant4-NST method) and the results were, thus, not displayed here. When 45% unbalanced DIF-items were considered, the newly suggested selection strategies outperformed the existing strategies in the majority of the unbalanced settings.

Figure 32 contains the false alarm rate (top row) and the hit rate (bottom row) in case of 45% unbalanced DIF-items of the selections combined with the *constant four anchor class*. The magnitudes generated constant (also depicted in Figure 28, right column), normally and from the sequence are ordered according to their variation (increasing from the left column to the right column).

One interesting point here is that the advantage of the threshold methods (the MPT- compared to the MP-selection and also the MTT- compared to the MT-selection) decreased when the variation of the DIF magnitude increased. The reason for this is that the same magnitude of DIF induces the most indistinct situation for the DIF tests since all items that have the same magnitude of DIF displayed no DIF in the tests when one of these items functions as the anchor. Anchor selection strategies must then be designed in a way such that they do not select those items displaying artificial DIF which can be caused by contamination or by random sampling fluctuation (for a more detailed discussion, see Chapter 6).

When the *iterative forward anchor class* is regarded in the setting of 45% unbalanced DIF-items (see Figure 33), again the MTT-selection – and the NST-selection for large sample sizes – reached the best results. The remaining existing methods were inadvisable. When the newly suggested methods were regarded, the superiority of the MTT-selection vanished when the variation of the DIF magnitude increased.

In settings where the DIF proportion was lower and the effects of the DIF magnitude were generated normally or from the sequence, again, the effects of the anchor selection strategies were less strongly developed. Still, large proportions of DIF may occur in practical research situations (Shih and Wang, 2009). Hence, anchor selection strategies should also perform satisfyingly in these challenging situations.

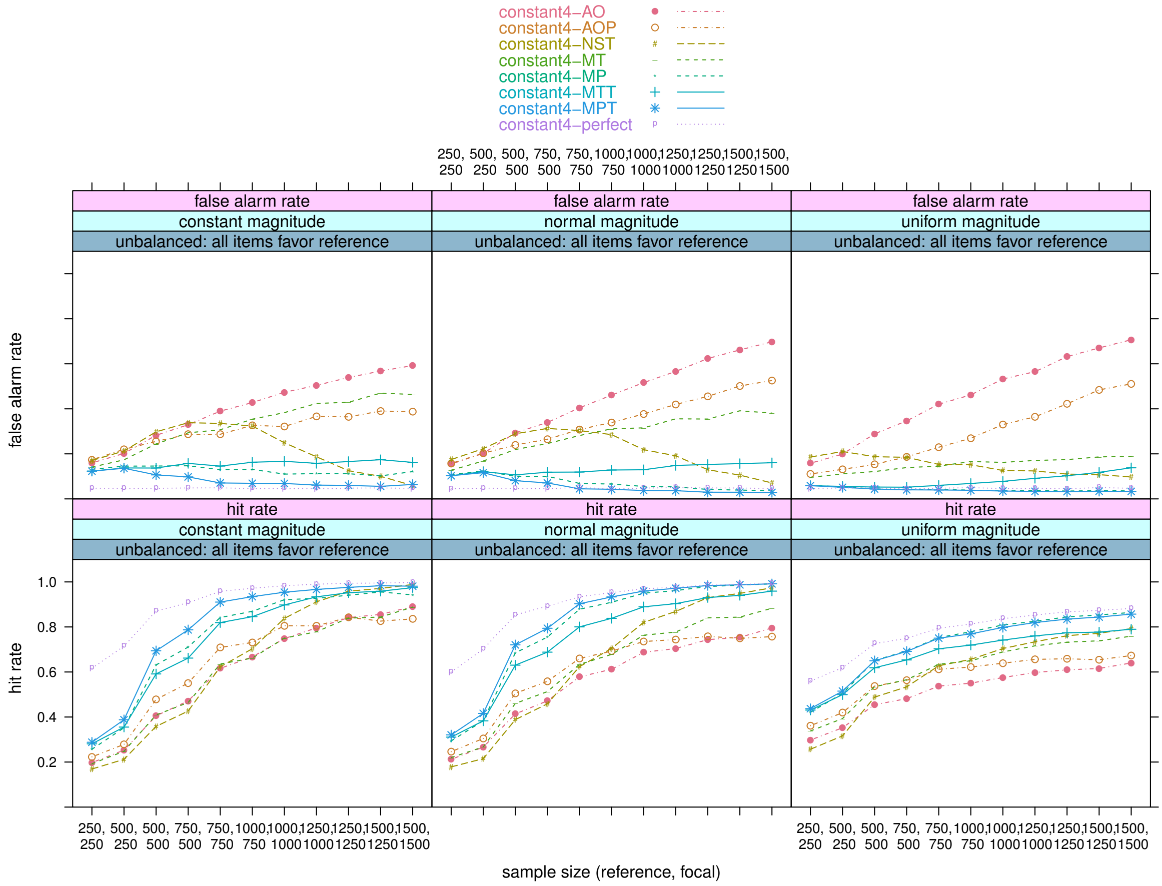

**Figure 32 –** *Unbalanced condition: 45% DIF-items favoring the reference group; sample size varied from* (250, 250) *up to* (1500, 1500)*; magnitude of DIF: constant, normal and from the sequence* $\{0.2, 0.4, \ldots, 1.0\}$*; top row: false alarm rates; bottom row: hit rates.*

In summary, a variety of simulated settings were considered in the present simulation study. The additional settings illustrate that the correct classification of DIF- and DIF-free items got even more challenging, when the proportion of DIF was large. The difference between the new anchor selections decreased when the variation of the differences between the DIF-item parameters of the reference and focal group increased. The further results substantiate that the threshold selections (MTT- and MPT-selection) outperformed the corresponding mere mean value selection strategies (MT- and MP-selection). Finally, none of the additionally considered simulation settings contradicted the recommendation to use the MPT-selection when the *constant four anchor class* is used and the MTT-selection when the *iterative forward anchor class* is employed.

## 7.5 Discussion and practical recommendations

In this chapter, we introduced three new anchor selection strategies and compared them to existing methods that do not rely on any prior knowledge of DIF-free items. Moreover, we

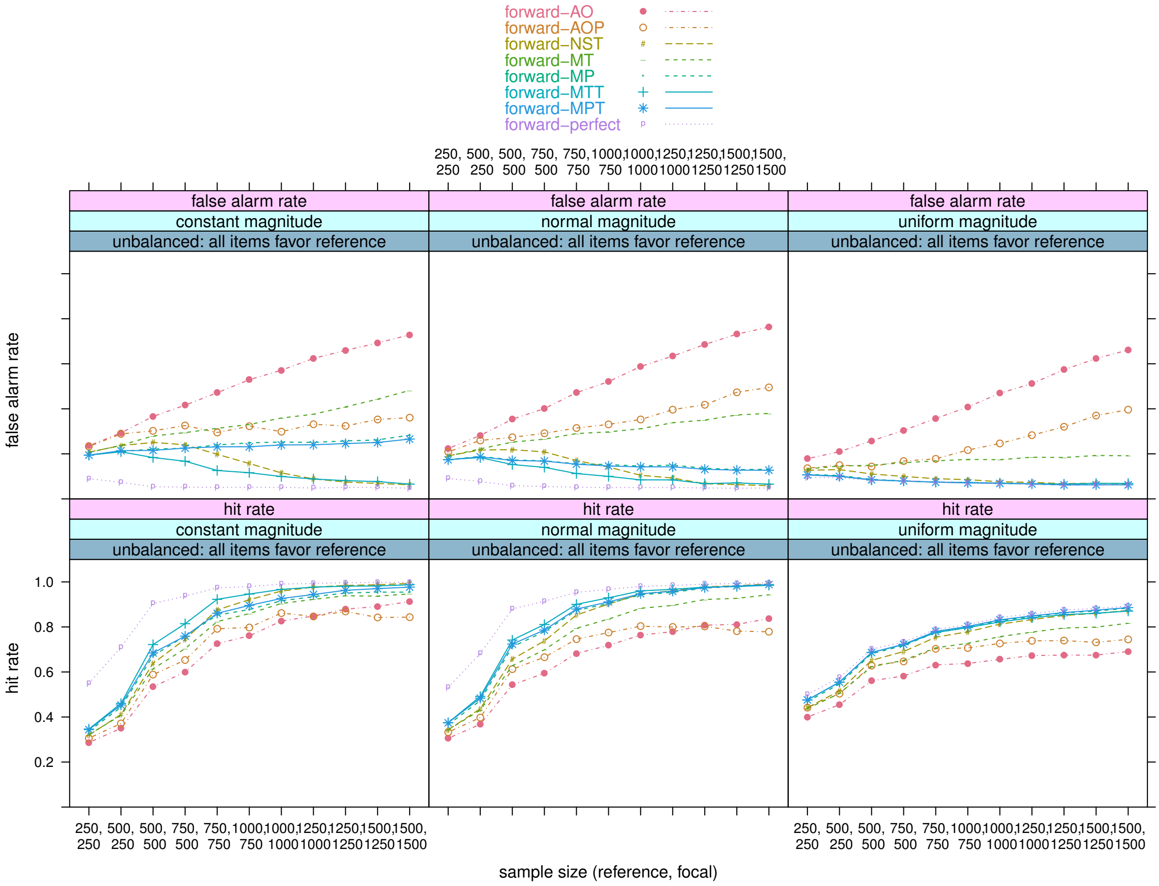

**Figure 33 –** *Unbalanced condition: 45% DIF-items favoring the reference group; sample size varied from* (250, 250) *up to* (1500, 1500)*; magnitude of DIF: constant, normal and from the sequence* {0.2, 0.4, . . . , 1.0}*; top row: false alarm rates; bottom row: hit rates.*

introduced a straightforward notation of the anchor selection strategies to facilitate the implementation and the usage of the newly suggested anchor selection strategies. An extensive simulation study was conducted to evaluate the performance of the anchor selection strategies in combination with the *constant four anchor class* and the *iterative forward anchor class*. The two anchor classes are structurally different, since the *constant four anchor class* always uses a short anchor of constant length four, whereas the *iterative forward class* determines the anchor length in an iterative way and usually yields a longer anchor.

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

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

In case of the selection of a short anchor of constant length four, the MPT-selection outperformed all other investigated empirical anchor selection strategies by yielding a low false alarm rate and simultaneously reaching a high hit rate in all regarded conditions. Thus, we recommend to use the MPT-selection if a short constant anchor length is intended.

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

Furthermore, first results indicate that neither the constant4-MPT nor the forward-MTT method was clearly superior regarding strictly smaller and less fluctuating false alarm rates and higher and less fluctuating hit rates. Only when DIF was balanced and the sample size was large, the constant4-MPT method might have a slight advantage by yielding a lower and less fluctuating false alarm rate and a comparable hit rate. On the other hand when DIF was unbalanced and the sample size was large, the forward-MTT method might have a slight advantage by yielding a higher hit rate and more reliable results due to less fluctuation within the hit rate as well as within the false alarm rate.

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

# 8 Outlook on anchor selection strategies for multiple group comparisons in DIF analysis

***Summary:*** *Explicit anchor selection strategies that locate the anchor items have been developed and investigated only for the comparison of two groups, namely the reference and the focal group. However, multiple group comparisons are highly relevant for practical testing situations even if they are controversially discussed. Thus, in this chapter, we discuss two alternative ways to address the anchor problem for multiple group comparisons. The first alternative is to select an anchor set for each paired comparison. The second alternative is a two-step approach to conduct item-wise paired multiple group comparisons using a common set of anchor items. In the first step, generalizations of anchor selection strategies to multiple group comparisons are suggested by means of two aggregation rules. In the second step, item-wise paired multiple group comparisons are conducted using the anchor set found in the first step.*

***Keywords:*** *Rasch model, differential item functioning (DIF), anchor selection, anchor class, multiple group comparisons, uniform DIF, measurement invariance*

## 8.1 Introduction

The analysis of differential item functioning (DIF) has attracted much attention in the field of psychological and educational testing. The reason for this is that if DIF is present, the test results no longer represent the ability alone and the groups of test-takers cannot be compared in an objective, fair way as highlighted in Section 3.2.

DIF occurs if test-takers from different groups display different probabilities of solving an item even if they have the same latent trait. Variables typically proposed for testing include age, gender, ethnicity and language, depending on the objective of the assessment (cf., e.g., Gelin *et al.*, 2004; Perkins *et al.*, 2006; Woods *et al.*, 2009; Pedraza *et al.*, 2009). Thus, DIF analysis is very likely intended for more than two groups (see Penfield, 2001, for more practical examples of several focal groups). Ideally, the DIF analysis would consist of item-wise paired multiple group comparisons that answer the question which items display DIF and between which groups.

However, there are several problems associated with item-wise paired multiple group comparisons: the number of falsely significant tests increases with the total number of conducted tests, the statistical power of paired comparisons may be lower compared to simultaneous tests over all groups and the computational effort may be high (Penfield, 2001). The first two arguments are inherent to the multiple testing situation. To reduce the number of falsely significant tests, adjustments for multiple testing are necessary and result in a stricter significance level that is theoretically supposed and empirically confirmed to reduce the statistical power (e.g. Penfield, 2001). Critical aspects of this decision will be addressed in Section 10.1.2. The computational burden, however, can be lowered easily by using a computationally non-intensive test statistic, such as the Wald test (see, e.g., Glas and Verhelst, 1995) for DIF-detection.

To solve the problem of the reduced power, Penfield (2001) suggested to conduct item-wise tests simultaneously over all groups. This approach has the disadvantage that the tests only answer the question which items display DIF and not yet between which groups. Since the group information may be crucial for the study of the underlying causes of DIF, the author recommends to conduct post-hoc tests consisting of item-wise paired comparisons between the groups (Penfield, 2001, p. 256). Similarly, Kim *et al.* (1995) recommended to compare the groups pairwise after the item-wise simultaneous test indicated DIF. In addition, item-wise paired multiple group comparisons may also provide useful insights when they are used as post-hoc tests for groups that were empirically determined e.g. by latent class approaches (Rost, 1990; Rost and von Davier, 1995) or by Rasch trees (Strobl *et al.*, 2013).

This highlights the importance of the item-wise paired multiple group comparisons. These paired comparisons – e.g. using the item-wise Wald test that relies on the comparison of the estimated item parameters – require an anchor, regardless of whether they are conducted for all items or only for those test items that displayed significant DIF in an item-wise simultaneous test. Note, however that statistical tests conducted post-hoc may also be affected by the effects described by Leeb and Pötscher (2005) and Berk *et al.* (2010).

For the anchor process, methods that rely on an explicit anchor selection are advisable (see Chapter 7 for a detailed discussion). Several authors have suggested anchor selection strategies to overcome the risk of anchor *contamination* i.e. the risk that at least one anchor item is a DIF-item (see, e.g., Finch, 2005; Woods, 2009; Wang *et al.*, 2012). Contamination can cause seriously increased false alarm rates (see, e.g., Wang and Yeh, 2003; Wang, 2004; Wang and Su, 2004; Finch, 2005; Stark *et al.*, 2006; Woods, 2009; Kopf *et al.*, 2013c) and, hence, items truly free of DIF may appear to have DIF. This jeopardizes the results of the DIF analysis as well as the associated investigation of the causes of DIF (Jodoin and Gierl, 2001). Thus, a suitable anchor selection is crucial for suitable DIF analysis. Nevertheless, the development of anchor selection strategies focused only on the comparison of two groups.

## 8.2 Alternative I: Selection of an anchor set for each paired comparison

If paired tests are intended for multiple groups, one strategy could be to select a (potentially) different anchor for each paired comparison. This alternative can be carried out using the strategies discussed in Chapter 7, since the paired comparisons are identical to the two-group comparisons of one reference and one focal group that were discussed in the previous chapters.

This alternative has three major disadvantages. First, the anchor set probably differs with respect to the groups that are currently tested and hence, the anchored item parameters for one specific group are likely to vary depending on the group they are compared with which is not very intuitive, but confusing. Especially when considering post-hoc tests for the Rasch trees, the representation of (potentially different) item parameters depending on the paired comparisons will be rather complex when the number of the end nodes is large, and – if it was applied for the graphical display of the item parameters in the end nodes of a Rasch tree – would make it impossible to directly compare the parameter locations.

Second, the anchor process has to be carried out for each intended group comparison. Third, only the data of two groups are considered in the anchor selection and, thus, not all available information is used. Those items that have DIF in one group comparison are likely to have DIF also in other group comparisons, e.g. when different language groups are compared, but the additional information is not considered.

## 8.3 Alternative II: Selection of a common anchor set

In our second alternative, we suggest to determine a one common anchor set for all groups by means of a two-step procedure. Then, the estimated item parameters for one specific group are the same over all comparisons and all information of the data set is used to accomplish this, but the anchor selection has to be extended for multiple groups. This is the aim of this chapter. It is, to our knowledge, the first attempt to address the generalization of anchor selection strategies to multiple group comparisons. In the following, we extend the anchor process to multiple group comparisons by suggesting two aggregation rules.

The anchor method guides the researcher through the process of selecting the anchor items. In our two-step approach, we focus on the constant anchor class where a predefined number of common anchor items is intended in the first step but the approach can be extended for other anchor classes as well. The anchor length of four items is typically used and it is claimed that this length assures sufficient power (cf. e.g., Shih and Wang 2009, Wang *et al.* 2012 and, for critical aspects, see Chapter 6). These common anchor items are included in the restriction and, thus, used to define a common metric for the estimated item parameters $\hat{\beta}_j$ with $j = 1, \ldots, k$ of a test of the length $k$ (see again Chapter 5 for a detailed description of the anchoring process). DIF in the anchored item parameters is then tested in the second step by means of paired item-wise multiple group comparisons.

In case two groups, the *reference* (ref) and the *focal* (foc) group, are compared, all anchor selection strategies that were presented in the previous Chapter 7 yield a ranking order of candidate anchor items. For example, the mean test statistic (MT) selection (Shih and Wang, 2009) defines the criterion $c_j^{\text{ref-foc}}$ for item $j$ between reference and focal group as

$$c_j^{\text{ref-foc}} = \left( \frac{1}{k-1} \sum_{\ell \in \{1,\ldots,k\} \setminus j} \left| t_j(a_\ell) \right| \right).$$

Hence, every item is assigned the mean absolute test statistic resulting from the test results using every other item once as single anchor. In the next step, the ranks of all $c_j^{\text{ref-foc}}$'s are calculated and define the order of candidate anchor items; Those items that are assigned the lowest ranks are used as anchor items (see Chapter 7).

*Step one: Selection of a common anchor*

However, if multiple groups are compared and one single anchor set is intended, this rule to rank the criteria does no longer suffice. Thus, in step one we suggest two ways to aggregate the criteria originating from every possible paired group comparison. The reason to use all-pairs comparisons (also know as Tukey's all-pairwise comparisons, see, e.g., Bretz *et al.* 2011) for the anchor process in step one is that every group comparison is regarded and, thus, every group occurs equally often in the comparisons. If one would use e.g. the many-to-one comparisons (Dunnett contrasts) for anchor selection, the reference group would be regarded more frequently if more than two groups are considered. The choice of the anchor would depend more strongly on characteristics of the reference group compared to those of the focal groups. Thus, we use all-pairs comparisons and face $\frac{G \cdot (G-1)}{2} = M$ group combinations. For each paired group combination $m = 1, \ldots, M$, the resulting criteria $c_j^m$ are calculated for every item of the test as defined by the anchor selection for the two-group case.

The generalization of anchor candidates for multiple group comparisons is carried out by aggregating the results from the paired comparisons. Therefore, we suggest two strategies that we term *aggregation* rules. The first aggregation rule is to use the average criterion value $c_j^\bullet$ for every item $j$ over all regarded comparisons

$$c_j^\bullet = \frac{1}{M} \sum_{m=1}^{M} c_j^m.$$

We term this suggestion the *mean aggregation* rule. The items corresponding to the lowest ranks of $c_j^\bullet$ are selected as anchor items. One disadvantage of this strategy might be that every group comparison obtains the same weight in the decision which item should be used as anchor. Hence, we propose the *minimax aggregation* rule that uses the minimax-strategy that minimizes the loss of the worst-case scenario in decision theory (Savage, 1951). First, every item $j$ is assigned the largest criterion value occurring in the paired comparisons

$$c_j^\bullet = \max_{m \in \{1,\ldots,M\}} c_j^m.$$

Second, the items corresponding to the lowest ranks of $c_j^\bullet$ are, again, selected as anchor items. We assume that this strategy is superior since the worst-case resulting from the paired comparisons decides the ranking order of the candidate anchor items. From decision theory, on would assume in case e.g. one item displays DIF only between some groups and not between others, the minimax rule as superior since it takes the worst-case, i.e. the situation where the strongest DIF occurs, into account, whereas the mean rule uses all test results and the average reduces the DIF effect. However, dependencies in the all-pairs comparisons complicate this argumentation and also the construction of fair comparisons of both aggregation rules, that will be addressed in future research.

*Step two: DIF tests using the final anchor*

After the anchor set is determined, the final DIF tests that now consist of item-wise paired multiple group comparisons can be carried out in the second step. Typically used paired comparisons are the all-pairs comparisons or many-to-one comparisons. All-pairs comparisons include the null hypotheses $H_0 : \beta_j^g = \beta_j^{g'}$ with $g \neq g'$ resulting in $\frac{G \cdot (G-1)}{2}$ comparisons per item, as already stated above. In total, $k \cdot \left(\frac{G \cdot (G-1)}{2}\right)$ tests result if all $k$ test items are studied for DIF. In contrast to this, many-to-one comparisons focus on the comparison of all groups to a control group which is the reference group in DIF analysis (suggested e.g. by Penfield, 2001). The null hypotheses $H_0 : \beta_j^{\text{ref}} = \beta_j^{g'}$ with ref $\neq g'$ yield $k \cdot (G - 1)$ tests in total.

| | **test length** $k$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k = 10$ | | | | $k = 20$ | | | | $k = 40$ | | | |
| **number of groups** $G$ | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| all-pairs | 10 | 30 | 60 | 100 | 20 | 60 | 120 | 200 | 40 | 120 | 240 | 400 |
| many-to-one | 10 | 20 | 30 | 40 | 20 | 40 | 60 | 80 | 40 | 80 | 120 | 160 |

**Table 12** – *Number of tests using the all-pairs or the many-to-one comparisons for* $k \in \{10, 20, 40\}$ *and* $G \in \{2, 3, 4, 5\}$.

Table 12 includes the total number of comparisons for some exemplary settings. It can be seen that the number of tests conducted increases more rapidly with the number of groups and with the number of test items for the all-pairs compared to the many-to-one comparisons. With an increasing number of tests, the problems associated with the multiple testing situation (see Section 8.1) reinforce. Thus, many-to-one comparisons are preferred in DIF testing situations. Many-to-one comparisons also represent a natural way to compare the *reference* group that is often a majority group against the present focal groups (see, e.g., Magis and De Boeck, 2011, for details). The additional information which of the focal groups differs from which other focal group might be less interesting in common DIF analysis.

Thus, we focus on many-to-one comparisons for the final DIF test in step two, but the anchor could in principle be used also for all-pairs comparisons as might be investigated in future research. Note that the reinforcement of the problems of multiple testing does not affect the anchor selection in step one since only a ranking order of candidate anchor items is defined that is intended to include all information in a symmetric way.

## 8.4 Discussion and future research questions

In this chapter we suggested two alternatives to conduct paired multiple group comparisons. We recommend to use a common anchor set as opposed to a (potentially) different anchor set for each paired comparison. Our suggestion can be implemented by means of a two-step procedure, where all-pairs comparisons are used to locate a common anchor set in a first step and many-to-one comparisons are employed in the final DIF test as a second step to answer the question which focal groups differ from the reference group with respect to their measurement properties.

This chapter was limited to the theoretical presentation of alternative ways to address multiple group comparisons. Future research has to be carried out to empirically compare the alternative ways to either select an anchor set for each paired comparison or to select a common set of anchor items.

Furthermore, the latter alternative to select a common set of anchor items for paired multiple group comparisons requires an aggregation rule. Future research is intended to compare the mean and the minimax aggregation rules regarding their performance to locate a suitable anchor set in an extensive simulation study. No research is available yet on how the anchor selection strategies perform in case of multiple groups. Moreover, an extension for the iterative forward anchor class for multiple group comparisons is needed as well, since methods from the iterative forward anchor class often showed the highest hit rates in two-group comparisons in the previous chapters.

# 9 Alternative ideas

***Summary:*** *In this chapter some alternatives to the approaches presented in the previous chapters are discussed. Therefore, the previously suggested quasi-variances are briefly reviewed. Quasi-variances are designed to summarize the information of the variance-covariance matrix and allow us to test for DIF in all k items at the same time by employing quasi-Wald tests. Firstly, the performance of these quasi-Wald tests for statistical inference on the all k items except the first anchor candidate is compared to the performance of the standard Wald tests that were used throughout this thesis. Secondly, the strategy to declare the first anchor item as DIF-free – as was done in the previous chapters of this thesis – is compared to the alternative approach to evaluate the quasi-Wald test result for the first anchor item.*

***Keywords:*** *Rasch model, anchoring, anchor methods, item response theory (IRT), differential item functioning (DIF), DIF analysis, quasi-variances, item bias*

## 9.1 Introduction

As discussed in various parts of this thesis, one linear restriction is placed on the item parameter estimates as the origin of the scale in the Rasch model can be arbitrarily chosen (Fischer, 1995). The *scale indeterminacy* is solved, firstly, by setting the restriction $\tilde{\beta}_1 = 0$. The corresponding covariance matrix $\widehat{\text{Var}}(\tilde{\beta})$ then contains zero entries in the first row and in the first column as discussed in Chapter 5.

As already stated above, other restrictions can be obtained by transformation using the equations $\hat{\beta} = A\tilde{\beta}$ and $\widehat{\text{Var}}(\hat{\beta}) = A\widehat{\text{Var}}(\tilde{\beta})A^\top$, where $A = I_k - \frac{1}{\sum_{\ell=1}^{k} a_\ell} 1_k \cdot a^\top$, $I_k$ denotes the identity matrix, $1_k$ denotes a vector of one entries and $a$ is a vector with one entries for those elements $a_\ell$ that are included in the restriction and zero entries otherwise (e.g., $a = (1, 0, 1, 0, 0, \ldots)^\top$ including item 1 and item 3). The entries of the rank deficient covariance matrix $\widehat{\text{Var}}(\hat{\beta})$ in the row and in the column of the item that is first included in the restriction were set to zero throughout this thesis. As a consequence, the first candidate anchor item obtained no DIF test and was declared DIF-free for all anchor classes that rely on an anchor selection.

An alternative way to handle the situation where the constant anchor class (that relies on an anchor selection) was regarded was implemented by Woods (2009). In her simulation study, for example, only "*all items not selected to be anchors were tested for DIF*" (Woods, 2009, p. 46). This approach excludes the possibility to detect DIF in the anchor items, does not use all information available (since $k-1$ standard errors are indeed available) and is, thus, not discussed further in this thesis.

## 9.2 Quasi-variances

To overcome the problem that $k$ items of a test are of interest, but only $k - 1$ standard errors result from the proceeding described above, the previously suggested quasi-variances (Firth, 2003; Firth and De Menezes, 2004) are briefly reviewed and employed for DIF analysis for the first time.

### 9.2.1 Basic idea of quasi-variances

In his article, Firth (2003) suggested an alternative representation of the results of a statistical model by reporting quasi-variances. The commonly used representation for e.g. generalized linear models includes summary tables where the estimated coefficients are printed together with their standard errors for the Wald tests for the null hypotheses that the respective coefficients equal zero. As an example, he refers to a situation where a categorical variable is included in the model by means of several binary dummy variables. Their effects are reported by means of their estimated coefficients, in his example $\hat{\tau}_2$ to $\hat{\tau}_5$, and the standard errors. With this information, certain other statistical hypotheses such as $\hat{\tau}_5 - \hat{\tau}_3 = 0$ cannot be tested, since the information of the complete variance-covariance matrix (denoted Var) is not reported – due to reasons of space constraints and readability. The complete variance-covariance matrix would be required to obtain the standard error as $\sqrt{\widehat{\text{Var}}(\hat{\tau}_5 - \hat{\tau}_3)} = \sqrt{\widehat{\text{Var}}(\hat{\tau})_{5,5} + \widehat{\text{Var}}(\hat{\tau})_{3,3} - 2\widehat{\text{Var}}(\hat{\tau})_{5,3}}$.

To overcome this problem, Firth (2003) proposed the usage of quasi-variances $s_j$, where $j = 1, \ldots, k$. These quasi-variances are constructed in a way such that the variances for contrasts $c = (c_1, \ldots, c_k)$ of the model coefficients $\tau_j$ are approximated by

$$\text{Var}(c^T \hat{\tau}) \approx c_1^2 s_1^2 + \ldots + c_k^2 s_k^2.$$

The quasi-variances are obtained by minimizing the sum of an error measure for simple contrasts

$$\sum_{l<j} \left[ \log \left( \frac{s_l^2 + s_j^2}{\widehat{\text{Var}}(\hat{\tau}_j - \hat{\tau}_l)} \right) \right]^2 = \sum_{l<j} \left[ \log \left( s_l^2 + s_j^2 \right) - \log \left( \widehat{\text{Var}}(\hat{\tau}_j - \hat{\tau}_l) \right) \right]^2 .$$

The minimization can be carried out using software for generalized linear models (see Firth, 2003; Firth and De Menezes, 2004, for details on the implementation and error criteria) and is provided in the R add-on package `qvcalc` (Firth, 2012). Even if the minimization is carried out including only simple contrasts, the solution is said to also hold as an error measure for more complex contrasts (Firth, 2003).

### 9.2.2 Quasi-variances in DIF detection

As opposed to the motivating example of Firth (2003) where inference should be carried out for a contrast that would require the complete variance-covariance matrix, here, quasi-variances are employed to overcome the problem that only $k - 1$ standard errors for the DIF analysis of a test are at hand, but all $k$ items with the corresponding parameters $\beta_j$ are of interest.

In this thesis, the Wald test is used for item-wise DIF detection. In the previous sections, the first anchor item was declared DIF free. The Wald test statistics for all remaining items $j$ of the test included the differences of the estimated and anchored item parameters $\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}$ (which strongly depend on the anchor items) in the numerator and the standard errors $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}})} = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{ref}})_{j,j} + \widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{foc}})_{j,j}}$ in the denominator. Therefore, the item parameters are first estimated separately in each group with the constraint $\tilde{\beta}_1 = 0$ and then anchored by transforming each group to the restriction reflecting the intended anchor method (see Section 9.2.1 or Chapter 5). The covariances for contrasts between the item parameters of the reference and the focal group are zero.

To allow for testing all $k$ items, the resulting variance-covariance matrix for each group is now used to calculate quasi-variances for all $k$ items. These quasi-variances are used to construct the *quasi-Wald test* for each item $j$, $j = 1, \ldots, k$ based on the quasi-variances $s_j^2$

$$T_j = \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{ref}})_{j,j} + \widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{foc}})_{j,j}}} \approx \frac{\hat{\beta}_j^{\mathrm{ref}} - \hat{\beta}_j^{\mathrm{foc}}}{\sqrt{s_j^{2,\mathrm{ref}} + s_j^{2,\mathrm{foc}}}}. \tag{26}$$

Note, however, that the anchor problem also occurs in the quasi-Wald test since the differences in the numerator still depend on the anchor method. To evaluate whether the estimated quasi-variances themselves are invariant to the anchor method chosen, the quasi-standard errors were calculated in a short simulation study (results not shown). The data consist of a test of length 40 with 45% balanced or unbalanced DIF-items. Quasi-standard errors were calculated for item parameters that were anchored using either the constant-single anchor (either pure or contaminated) or the constant-four anchor (either pure or contaminated with the degree of contamination equal to 50% or 100%). The quasi-standard errors of the quasi-Wald tests for two exemplary items (one simulated DIF-item and one simulated DIF-free item) did not differ in any replication, when machine precision is taken into account. Thus, in the following, it is assumed that quasi-variances are invariant against the employed anchor method. Future research may prove this result from a theoretical perspective.

### 9.2.3 Simulation study

In order to evaluate the performance of the quasi-Wald test compared to the Wald test, a short simulation study is carried out in the free R system for statistical computing (R Development Core Team, 2011) using the add-on R-package `qvcalc` by Firth (2012). The following questions are addressed: How do the quasi-Wald tests perform in comparison with the Wald tests for the $k - 1$ items excluding the first anchor candidate? Is the decision of declaring the first anchor as DIF-free appropriate or is the quasi-Wald test result superior?

*Data generating process*

Each data set, that represents one of 2000 replications from one simulation setting, corresponds to the simulated responses of two groups of subjects (the *reference* (ref) and the *focal* (foc) group) in a test with $k = 40$ items.

- *Person and item parameters*

  The person parameters are generated from a normal ability distribution with a higher mean for the reference group $\theta^{\text{ref}} \sim N(0, 1)$ than for the focal group $\theta^{\text{foc}} \sim N(-1, 1)$ similar to Wang *et al.* (2012). Values assigned to the item parameters are, again, $\beta$ = (-2.522, -1.902, . . ., 1.592) used by Wang *et al.* (2012), see Section 7.3. The responses in each group follow the Rasch model and are generated similar to the previous simulation studies (see e.g. Section 6.4).

- *DIF-items*

  The first 15, 30 or 45 percent of the items (cf. paragraph DIF proportions in the next section) are chosen to display uniform DIF by setting the difference in the item parameters of reference and focal group $\Delta_{\text{DIF}} = \beta_j^{\text{ref}} - \beta_j^{\text{foc}}$ to +.6 or −.6 consistent with the intended direction of DIF.

*Manipulated variables*

Similar to previous simulation studies for the comparison of two groups (cf. Chapter 6 and Chapter 7), the manipulated variables were the sample size, the direction of DIF, the percentage of DIF and the anchor methods.

- *Sample sizes*

  The sample sizes in reference and focal group are defined by the following pairs $(n^{\text{ref}}, n^{\text{foc}}) \in \{(250, 250), (500, 250), (500, 500), (750, 500), (750, 750), \ldots, (1500, 1500)\}$ where, again, equal and different group sizes are considered.

- *Directions and proportions of DIF*

  The direction of DIF is either *balanced* where each DIF-item favors either the reference or the focal group but no systematic advantage for one group remains because the effects cancel out or *unbalanced* where a systematic disadvantage for the focal group is generated. The proportion of DIF is set to $p \in \{15\%, 30\%, 45\%\}$. In order to save space, only the most difficult situation of 45% DIF-items is discussed in detail, but the other proportions are included in the Appendix of this chapter.

- *Anchor methods*

  The quasi-Wald test (indicated by the ending *-quasi*) and the Wald test are combined with the previously suggested *constant4-MPT* and *forward-MTT* anchor methods (see Chapter 7).

*Outcome variables*

To assess the performance of the quasi-Wald tests, two situations are regarded. Firstly, the quasi-Wald tests are compared to the Wald tests for all items except for the first anchor item.

Secondly, the declaration that the first anchor item is DIF-free (that was used throughout this thesis) is compared with the result of the quasi-Wald test for the first anchor item.

- *Classification rate*

  For a single replication the *classification rate* indicates the number of correct test decisions about whether the item is a DIF-item. The estimated classification rate for each experimental setting is computed as the mean over all 2000 replications. This outcome variable is evaluated for the first anchor candidate.

- *False alarm rate and hit rate*

  For the remaining $k - 1$ items, the false alarm rate (as a measure for the type I error) and the hit rate (that corresponds to the *power* of the statistical test) were calculated similar to Section 6.4.

- *Further outcome variables*

  In addition, the proportion where the first anchor item was a DIF-item and also characteristics of the item-wise tests for the first item (a simulated DIF-item) and the last item (a simulated DIF-free item) of the test, such as the estimated (quasi-) standard errors, were computed.

### 9.2.4 Results: Quasi-Wald tests versus Wald tests

To conclude whether the quasi-Wald tests work appropriately, we now review the test results by means of the false alarm and the hit rate of the $k - 1$ items that are not the first anchor candidate when 45% DIF are present. Figure 34 (top row) shows the false alarm rates.

In case of balanced DIF (Figure 34, top row left), the forward methods reached similar false alarm rates independent of whether the Wald test (forward-MTT) or the quasi-Wald test (forward-MTT-quasi) was carried out. These false alarm rates were slightly inflated (average 0.06). When the Wald tests were computed for the constant4-MPT method, the false alarm rates were below the significance level (average 0.03). The quasi-Wald tests for the constant4-MPT method (constant4-MPT-quasi) displayed false alarm rates that were first slightly inflated (in case of 250 observations in each group: 0.07) but then decreased with the sample size (to 0.04).

In case of unbalanced DIF (Figure 34, top row right), the forward methods reached again similar false alarm rates that were first inflated but decreased with the sample size. The lowest (but still inflated) false alarm rate occurred when the Wald tests were regarded for the constant4-MPT method. The quasi-Wald tests for the constant4-MPT method displayed inflated false alarm rates and were clearly outperformed by the Wald tests when regarding the false alarm rate.

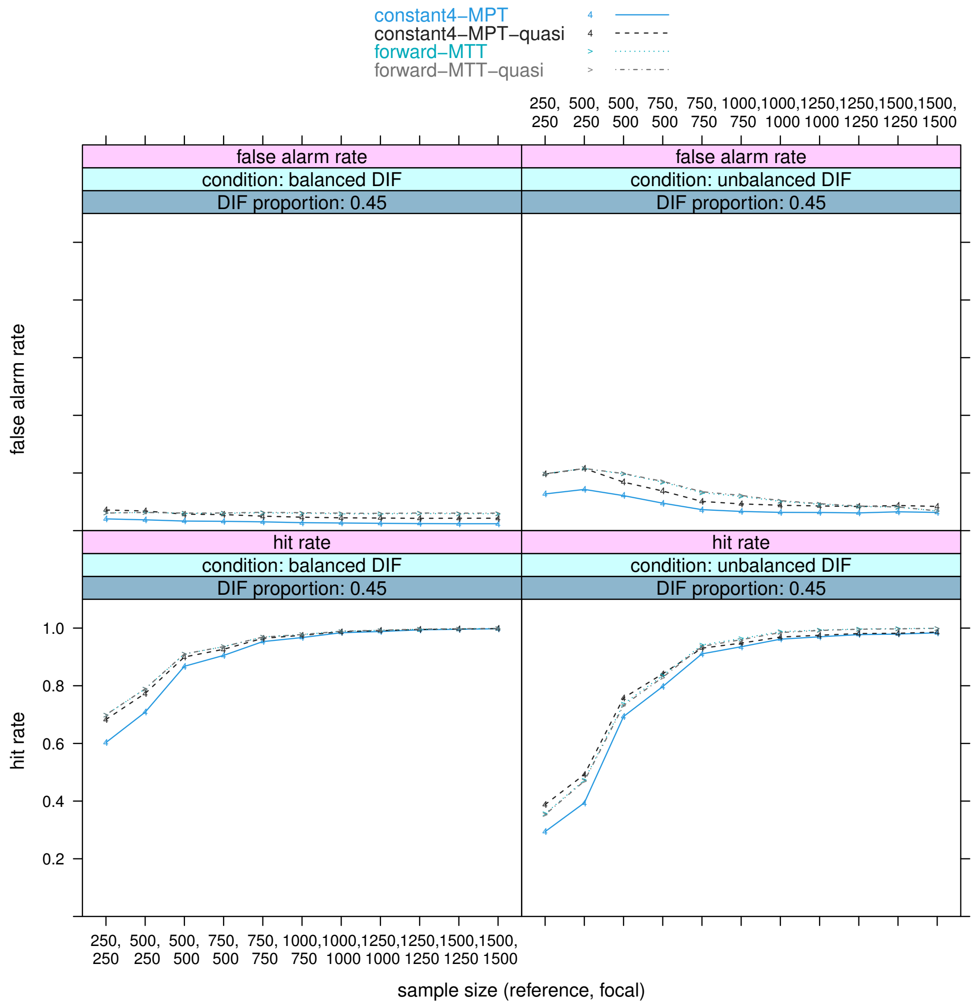

**Figure 34 –** *False alarm and hit rate of the $k-1$ DIF tests – not including the first anchor candidate – in case of* 45% *DIF.*

Figure 34 (bottom row) depicts the hit rate. Generally, the Wald test for the constant4-MPT method, that displayed the lowest false alarm rate, also showed the lowest hit rate. This result might be considered as an example where a constant anchor of four items does not assure sufficient power. It would then contradict the rule of thumb that four anchor items assure sufficient power as was stated previously (cf. e.g., Shih and Wang, 2009; Wang *et al.*, 2012). All other methods, the constant4-MPT-quasi, the forward-MTT and the forward-MTT-quasi method reached similar and high hit rates and outperformed the constant4-MPT method regarding the hit rate.

In case of lower proportions of DIF-items (see Figure 36 and Figure 37 in the Appendix), the only clear difference occurred for the constant4-MPT method that displayed a false alarm rate below 5% together with a lower hit rate and the remaining methods displayed a false alarm rate near the 5% level together with a higher power.

In summary, the Wald tests and the quasi-Wald tests reached similar results when they were combined with the forward-MTT method in any condition (see Figure 36 and Figure 37). Surprisingly, a larger difference between the Wald test and the quasi-Wald test results was visible when they were combined with the constant-MPT method. The quasi-Wald test displays a higher false alarm rate and a higher hit rate compared to the Wald test. Since the numerator $\hat{\beta}_j^{\text{ref}} - \hat{\beta}_j^{\text{foc}}$ of both tests is identical, but the quasi-Wald tests showed higher false alarm rates and hit rates, it is assumed that the test statistics are generally higher due to a smaller estimate of the quasi-variance. The estimated standard errors for the first item and the last item of the test are included in the Appendix in Figure 38 and Figure 39 and empirically confirm this hypothesis (and beyond that, the assumption that the quasi-variances are invariant against the chosen anchor items).

### 9.2.5 Results: Classification of the first anchor item

Now, the classification rate for the first anchor item is regarded. The results of the decision that the first anchor candidate is declared DIF-free are compared to the quasi-Wald test results for the first anchor candidate.

In case 45% balanced DIF-items were simulated (see Figure 35, left), the quasi-Wald tests (constant4-MPT-quasi and forward-MTT-quasi) and the declaration of the first anchor item as DIF-free (constant4-MPT, forward-MTT) method reached similar true classification rates.

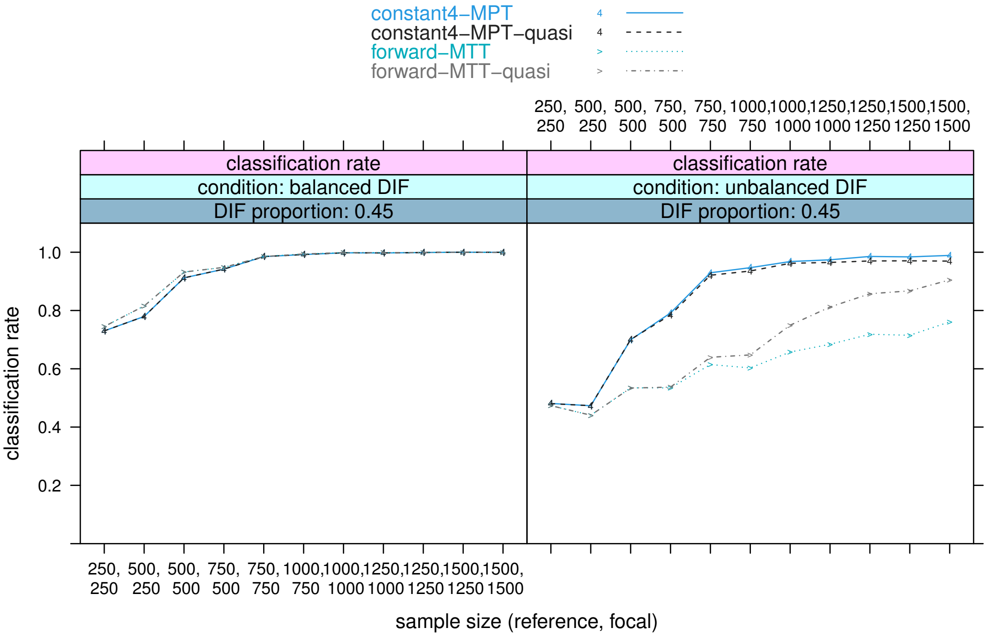

**Figure 35 –** *True classification rate of the first anchor item in case of* 45% *DIF.*

In the balanced case, the first anchor item was almost always classified as DIF-free by all methods. This was correct in approximately 74% of the replications (more precisely in 73% for the MPT- and in 74.5% for the MTT-selection) when the sample size was 250 in each group. The true classification rate increased with the sample size, since the proportion of replications where the first anchor candidate was a DIF-item decreased. The reason for this is that the anchor selection strategies perform better when the sample size is large. The forward-MTT-quasi and the forward-MTT method had a slight advantage, since the proportion of replications where the first candidate anchor item selected by the MTT-selection was a DIF-item was slightly smaller compared to the candidate selected by the MPT-selection for the balanced case. Thus, both strategies – to either declare the first anchor candidate as DIF-free or to use the quasi-Wald test – reached the same results and yielded the correct classification when the sample size was large.

In case of 45% unbalanced DIF-items (see Figure 35, right), the situation was quite different. The largest difference occurred for the methods combined with the constant4-MPT anchor method and combined with the forward-MTT method. The reason is that, in case of unbalanced DIF, the first anchor item found by the MTT-selection was more often contaminated as the first item located by the MPT-selection (see Section 7.4.3 in Chapter 7 for 1000 observations in each group). It might, thus, be reasonable to exclude a certain proportion of the first anchor candidates as discussed in Chapter 7, since the first anchor candidates displayed larger proportions of DIF-items. This would lead to a better classification result for the forward-MTT method.

When the iterative forward method was combined with the MTT-selection in the unbalanced condition, the quasi-Wald decision was superior, since it allowed to classify the first anchor item as a DIF-item and the power to detect true DIF increased with the sample size. The largest difference in the true classification rate between the declaration as DIF-free (forward-MTT) and the quasi-Wald test result (forward-MTT-quasi) was 15%. This means that the wrong decision to declare the first anchor candidate as DIF-free could be reduced by up to 15% when the quasi-Wald test is employed.

This also means that the hit rate of the forward-MTT method is underestimated in the simulation results: If the first anchor item is a simulated DIF-item, it is only possible to declare 17 out of the 18 true DIF-items as DIF-items since the first anchor is declared DIF-free. The corresponding hit rate is, thus, limited to $\frac{17}{18}$. Using the quasi-Wald test instead allows to declare up to 18 out of the 18 true DIF-items as DIF-items and the hit rate can reach the value 1. However, since the quasi-Wald test reached different results for the first anchor item in only up to 15% of the cases, the underestimation of the hit rate (regarding all items) in the previous simulation studies (with 18 out of 40 items simulated with DIF) by declaring the first anchor as DIF-free is reduced to $0.0083 = 0.15 \cdot \frac{1}{18}$ and is, thus, considered acceptable since it is less than 1% and since it would vanish if the anchor selection strategy works appropriately. Then, the first anchor candidate is expected be more often found from the set of DIF-free items. In case of the constant4-MPT method, the classification rate increased and the difference to the quasi-Wald test was marginal.

In summary, the strategy to declare the first anchor candidate found by the respective anchor selection as DIF-free yielded the same results as the quasi-Wald test in case of balanced DIF. This also holds for all conditions of balanced and unbalanced DIF when 15% or 30% of the items displayed DIF (results are shown in Figure 40 in the Appendix). In case of 45% unbalanced DIF and the constant4-MPT method, the differences were also small. However, in case the iterative forward procedure with the MTT-selection was regarded, the decision to declare the first anchor item as DIF-free was often wrong. The classification based on quasi-variances was better in regions of medium to large sample sizes. In case of 45% unbalanced DIF-items, one item of the test is often declared DIF-free while it was a simulated DIF-item. This holds for both strategies, the declaration and the quasi-Wald test for small sample sizes. In case of medium to large sample sizes, the true classification can be improved (in up to 15% of the replications) by employing quasi-Wald tests and, thus, the slight underestimation of the hit rates for the forward-MTT method could be reduced. Future research may investigate the situation when a certain proportion of the first anchor candidates of the MTT-selection (that displayed a larger proportion of DIF-items among the first anchor candidates, see again Chapter 7) is excluded.

## 9.3 Discussion

In this chapter, alternative approaches were presented. First, quasi-Wald tests were compared to Wald tests for all $k - 1$ items excluding the first anchor candidate. In case of 45% DIF, the largest differences occurred for the investigated methods, especially for the constant4-MPT method depending on whether it was combined with the Wald or the quasi-Wald test. The false alarm rates and the hit rates for the Wald test were lower compared to the respective rates of the quasi-Wald test. Therefore, none of the tests can be considered strictly superior. When the forward-MTT method was combined with both tests, no visible differences occurred. Thus, quasi-Wald tests are considered as an appropriate alternative to the Wald tests but further research is needed to explain under which conditions the quasi-Wald test and the Wald test results differ.

Second, the quasi-Wald tests were calculated again for the constant4-MPT and the forward-MTT method to assess DIF in the first anchor candidate. The test decision was compared to the declaration of the first anchor item as DIF-free. In the majority of the simulated settings, both strategies – to either declare the first anchor candidate as DIF-free or to use the quasi-Wald test – reached the same results.

The only exception occurred in case 45% unbalanced DIF-items were present. In this case, generally the constant4-MPT method had an advantage compared to the forward-MTT method, since the proportion of DIF-items among the first anchor candidates was higher. However, the declaration of the first item as DIF-free reached a classification rate similar to the quasi-Wald test when the constant4-MPT method was used. The only notable difference between the declaration as DIF-free and the quasi-Wald test result occurred in case of 45% unbalanced DIF-items when the forward-MTT method was regarded. The quasi-Wald test allowed an improvement of the correct classification rate of up to 15% and might, thus, be preferred. Nevertheless, the consequences if the quasi-Wald tests would have been used instead of the declaration strategy are considered marginal (since the improvement of the hit rate regarding the entire test is limited to $\frac{1}{18}$ in up to 15% of the replications and in total less than 1%).

Future research might investigate how the MTT-selection performs when a certain proportion of the first anchor candidates is excluded. It would also be interesting, whether the differences between the declaration as DIF-free and the quasi-Wald test result vanish as well.

## 9.4 Appendix

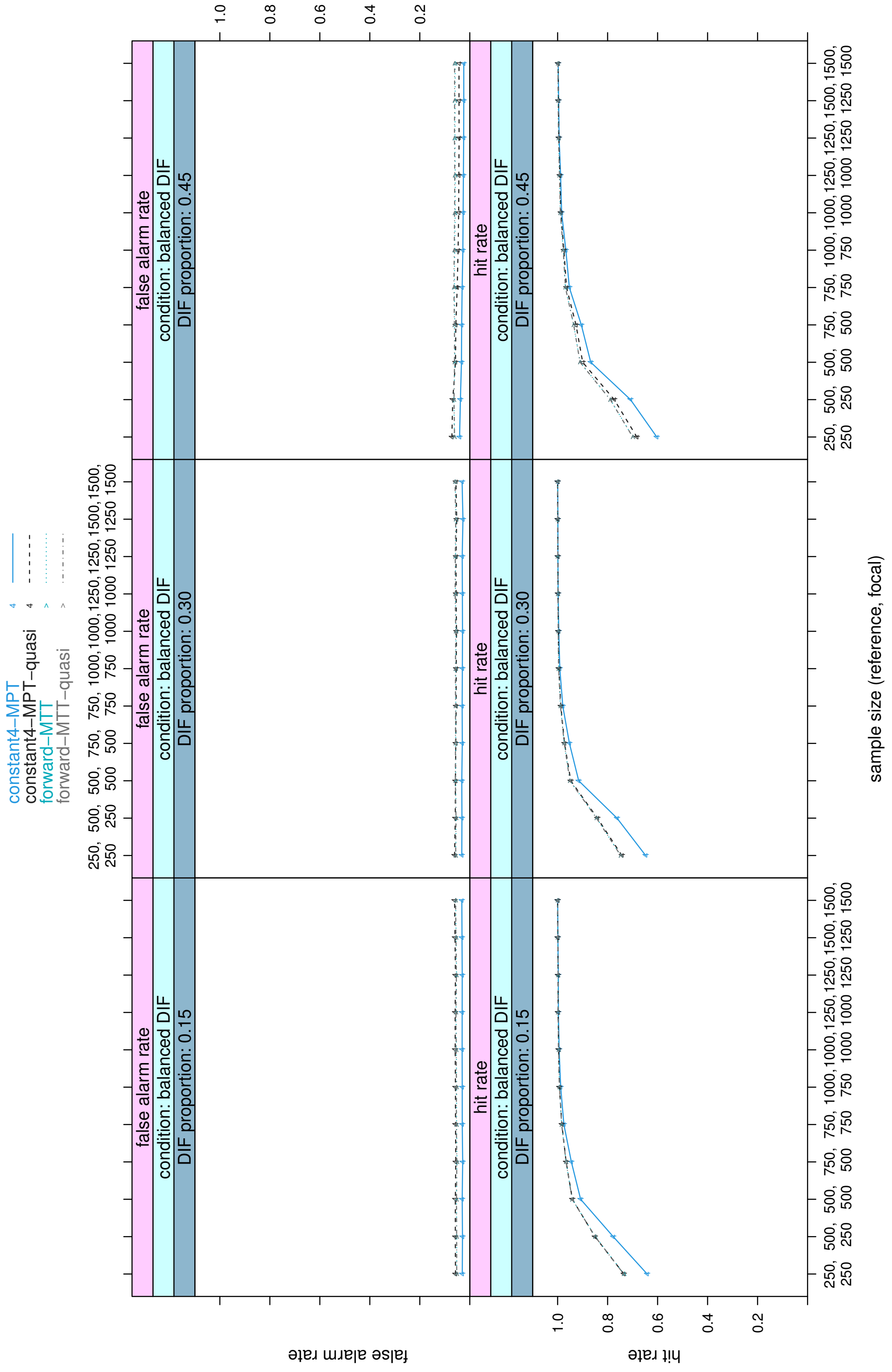

**Figure 36 –** *False alarm and hit rate of the* $k-1$ *DIF tests excluding the first anchor candidate in case of* 15%, 30% *or* 45% *balanced DIF.*

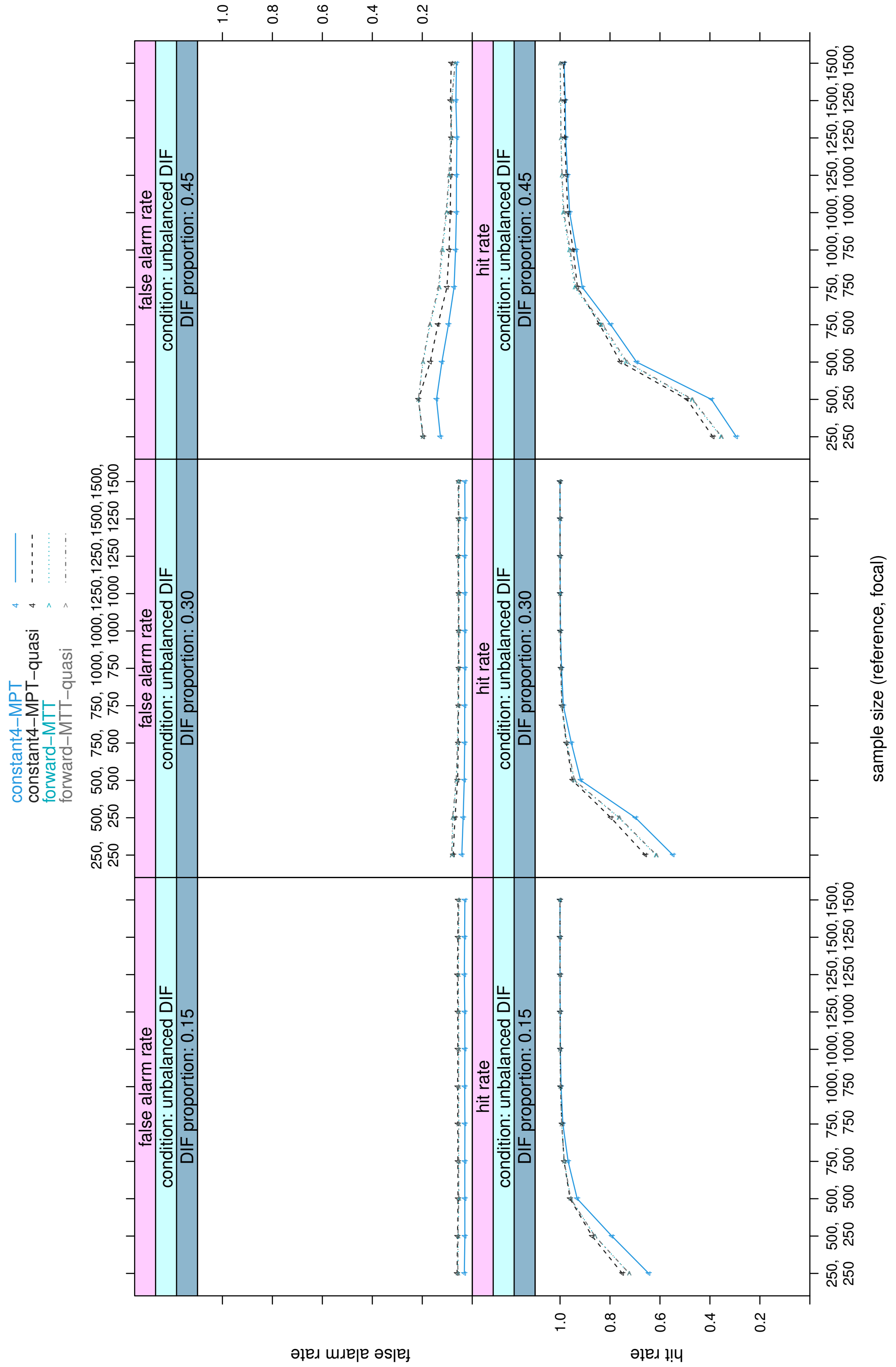

**Figure 37 –** *False alarm and hit rate of the $k-1$ DIF tests excluding the first anchor candidate in case of* 15%, 30% *or* 45% *unbalanced DIF.*

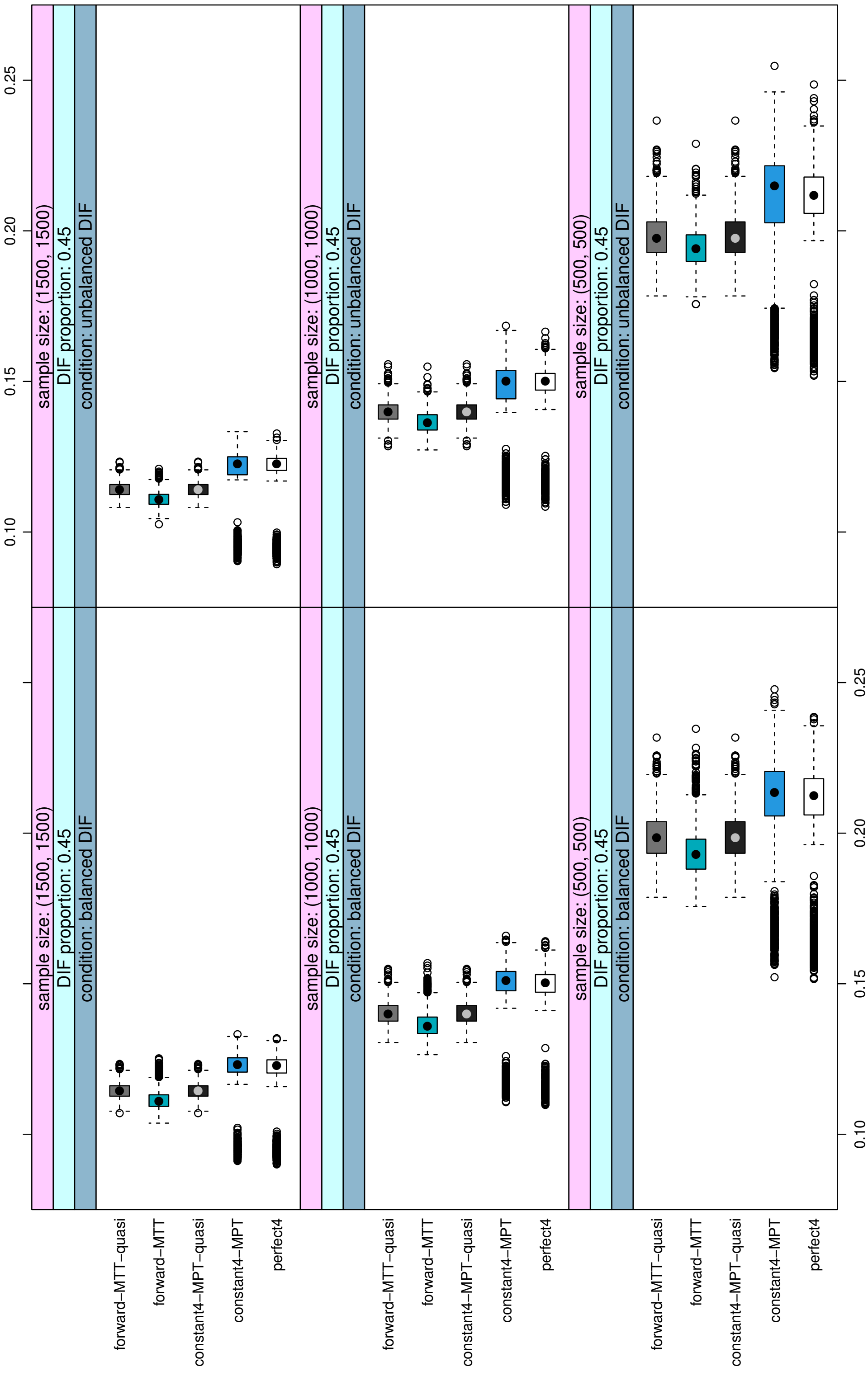

**Figure 38 –** *Estimated standard errors for the item parameter difference of the last item (DIF-free) for three different sample sizes. The replications where the last item was the first anchor candidate for either the MTT- or the MPT-selection were excluded in order to yield comparable results. The benchmark method (perfect4) consists of four randomly chosen DIF-free items.*

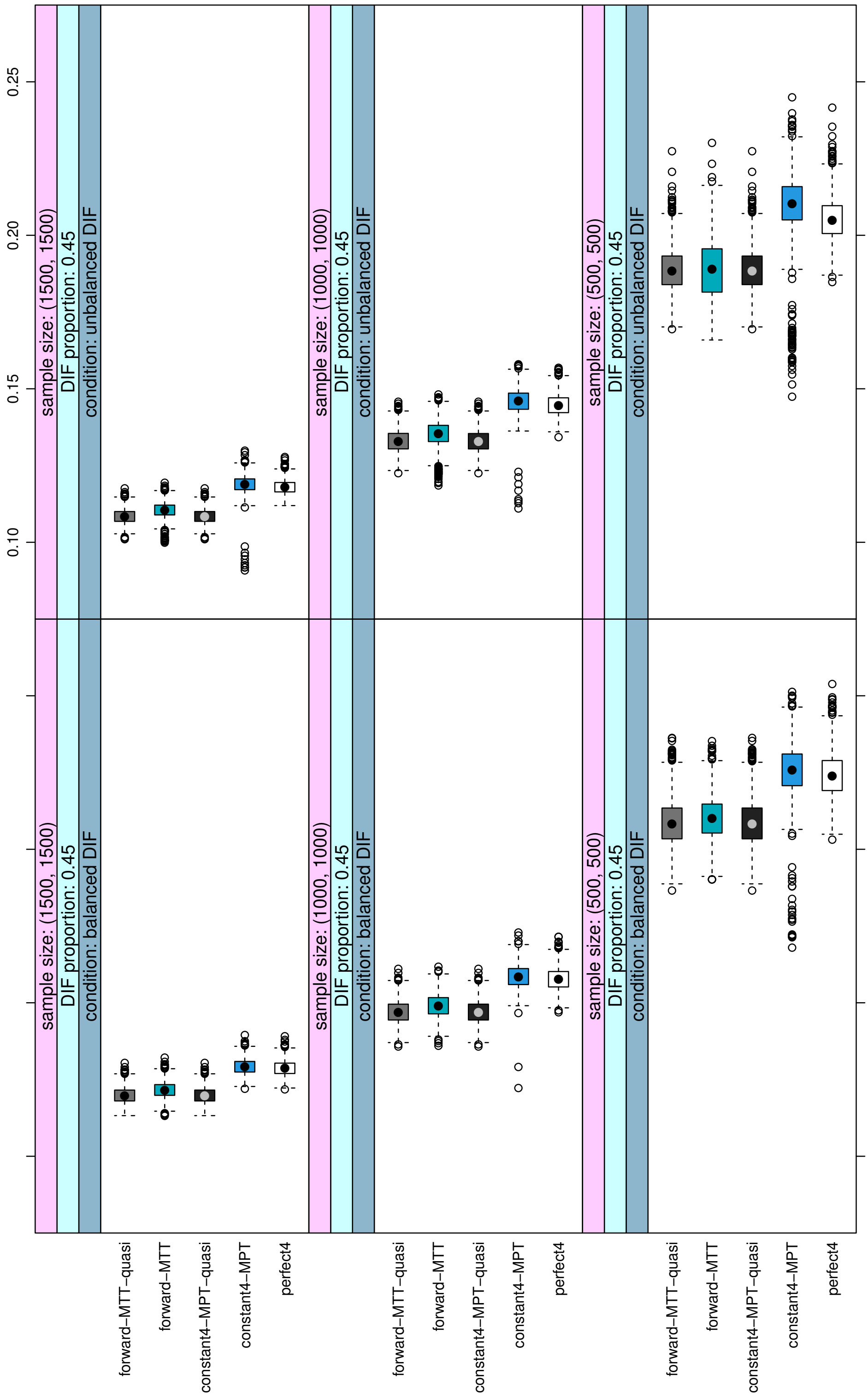

**Figure 39 –** *Estimated standard errors for the item parameter difference of the first item (DIF-item) for three different sample sizes. The replications where the first item was the first anchor candidate for either the MTT- or the MPT-selection were excluded in order to yield comparable results. The benchmark method (perfect4) consists of four randomly chosen DIF-free items.*

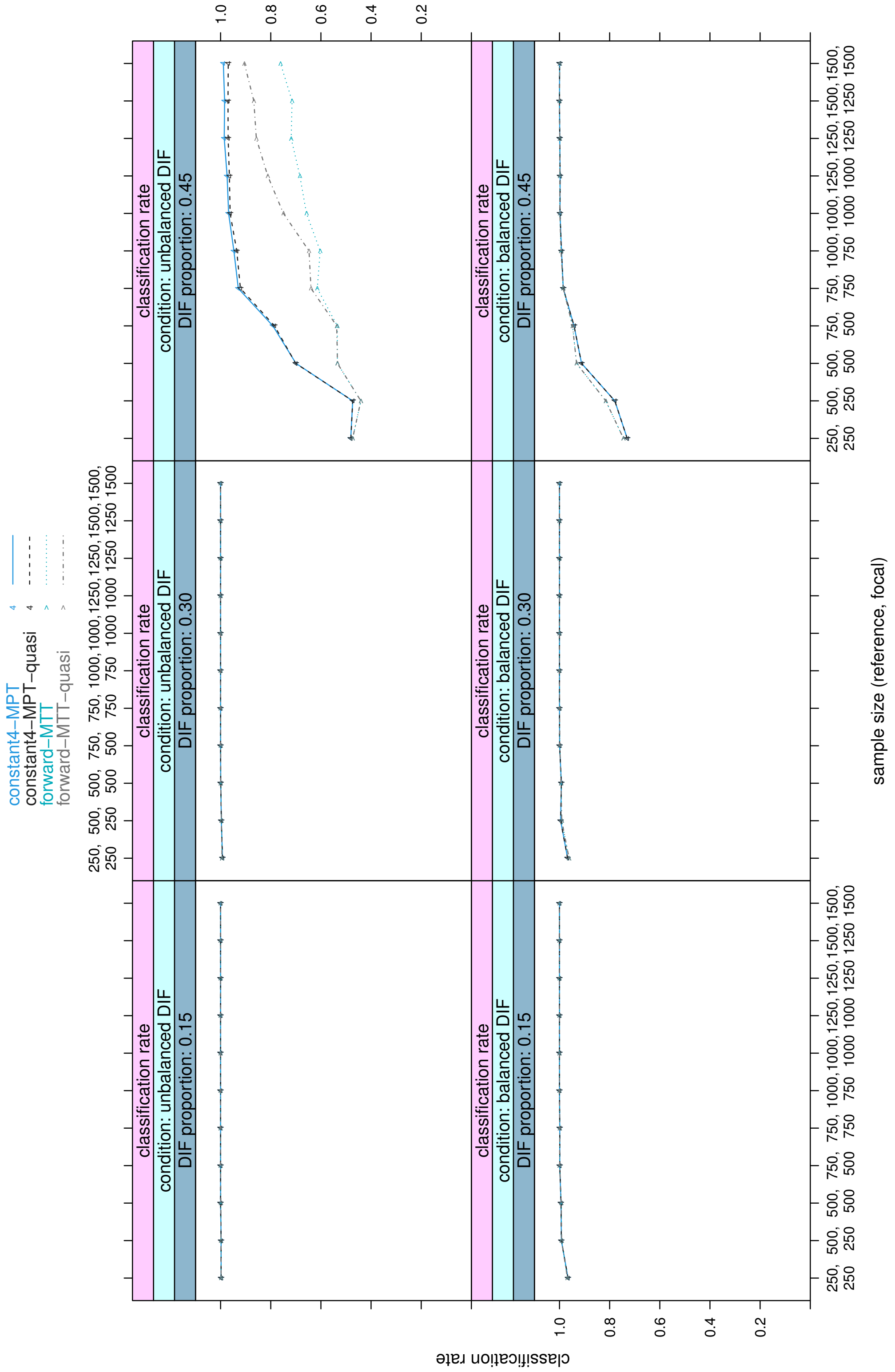

**Figure 40 –** *True classification rate of the first anchor item in case of* 15%, 30% *or* 45% *balanced or unbalanced DIF.*

# 10 Concluding remarks and outlook

To conclude, the questions addressed in this thesis are reviewed. Therefore, the contents will be summarized in the next section. Section 10.1.1 is concerned with a selection of new results provided in this thesis and highlights their impact on the development of statistical methods for the detection of DIF. The limits of this work are briefly discussed in Section 10.1.2. Finally, future research questions are proposed.

## 10.1 Concluding summary and overview

The aim of this thesis was the development of new statistical methods for the analysis of differential item functioning. One key aspect was concerned with the combination of model-based recursive partitioning (Zeileis *et al.*, 2008) and item response theory. First, the potential of model-based recursive partitioning for the social sciences was addressed in Chapter 2. The algorithmic approach is not limited to exploratory data analysis as is the case for several predecessor methods like classification and regression trees, but allows the researcher to specify a statistical model and to test whether its parameters are stable for the entire sample. This approach can be used as a test for violations of Ockham's Razor, that occur if relevant covariates have been omitted and, thus, the model is not complex enough to explain the data.

In the field of educational and psychological testing, in social and in behavioral sciences, the variables of interest, such as abilities, skills or attitudes, are often not observable. IRT models including the widely used Rasch model have been developed to measure these latent variables. The assumption that the item parameters are invariant, i.e. the item parameters are constant for all subgroups and are not affected by differential item functioning, is a crucial requirement to allow for fair comparisons between groups as was pointed out in Chapter 3.

Rasch trees that apply the model-based recursive partitioning algorithm to the Rasch model were previously suggested as a global test for uniform DIF as well as an overall model test (Strobl *et al.*, 2013). The question whether Rasch trees are also suited for the detection of non-uniform DIF was first addressed in Chapter 4. The results showed that the Rasch trees allow to find both types of DIF – uniform and also non-uniform DIF – if the sample size is large enough. Rasch trees are a flexible method for the detection of DIF, since they automatically provide the information which groups display DIF. However, Rasch trees are based on a global test and cannot yet answer the question which particular items are affected by DIF. Still, this information is crucial for test and questionnaire construction.

To allow for conclusions on the item-level, item-wise DIF tests were addressed. As a statistical test of DIF on the item-level, the Wald test, that relies on the comparison of the estimated item parameters of the Rasch model, was regarded in this thesis. To place the item parameters of different groups onto a common scale, the same restriction is imposed on the estimated item parameters. Items included in the restriction are termed *anchor items*. Strategies how the anchor items are located are termed anchor methods. Chapter 5 provided a conceptual framework for anchor methods, where they were divided in anchor classes and anchor selection strategies, and illustrated that the results of the Wald test strongly depended on the choice of the anchor items.

A new anchor class, termed the *iterative forward anchor class*, was suggested in Chapter 6. The idea behind this anchor class was to build a method that does not rely on initially biased test results and that allows for a longer anchor. The results showed that the iterative forward anchor class is an attractive alternative to the previously used anchor classes since on the one hand it allowed for a low false alarm rate and on the other hand achieved a high hit rate.

To further improve the results of the anchor methods, alternative anchor selection strategies were introduced and presented in Chapter 7. Three newly suggested anchor selection strategies were combined with the constant anchor class consisting of four anchor items and with the newly suggested iterative forward anchor class. The results showed that the newly suggested threshold methods were better suited for the detection of DIF than the previously suggested anchor methods. As a result, in the simulation study, the *mean p-value threshold* selection was best suited to locate a short anchor, consisting of four items, whereas the *mean test statistic threshold* selection was appropriate for the iterative forward method that allows for a longer anchor.

So far, the developments of the anchor methods were carried out only for the comparison of two groups. Chapter 8 first addressed the anchor problem when multiple groups are to be compared. This situation is relevant if researchers want to conduct post-hoc tests or descriptive measures after global test procedures – such as the Rasch trees – indicated any DIF. The solution suggested in this thesis was to select a common set of anchor items instead of a (potentially different) anchor set for each paired comparison. The selection strategies discussed in Chapter 7 were extended to select a common anchor set for multiple group comparisons by the development of two aggregation rules, the *minimax* aggregation rule (that relies on the worst case criterion occurring for the comparisons of two groups) and the *mean* aggregation (that is based on the mean criterion over all paired comparisons). The chapter pointed out the theoretical advantages of the common anchor set approach in combination with the minimax aggregation rule, that will be empirically investigated in future research.

Finally, some alternative ideas to those presented in this thesis were discussed in Chapter 9. The assumption that the first anchor item is DIF-free and the usage of quasi-variances in quasi-Wald tests for DIF detection were discussed.

### 10.1.1 Selection of scientific findings

In this section, the main scientific findings of this thesis are reviewed.

In Chapter 4, it was shown that Rasch trees were also suited for non-uniform DIF detection and in case of differing item difficulty parameters (that may induce uniform or non-uniform DIF) often even better suited than the extensions of the logistic regression. Rasch trees were shown to be a reasonable alternative also in cases where the logistic regression is misspecified or the group composition is unknown. The extensions of the logistic regression showed inflated false alarm rates when the groups were simulated from different ability distributions. Furthermore, many situations occurred where the post-hoc classifications of the type of DIF for the extensions of the logistic regression were far behind the detection rates. These situations should be investigated for the item-wise logistic regression as well.

A variety of proposed anchor methods was classified in Chapter 5 by means of a conceptual framework for clarification. The conceptual idea to divide the anchor class and the anchor selection was also implemented in the new iterative forward procedure and the results indicate that this idea – that the ranking order of candidate anchor items is separated from the anchor class – may also improve the anchor methods. The methods that separated both parts (the constant four class and the iterative forward class combined with the NST- and the AO-strategy) performed better compared to the methods where both parts were mixed (the constant four next candidate anchor method and the iterative backward method) in Chapter 6.

In Chapter 6, several interesting findings were discussed that explained the fact that the iterative forward method together with the NST-selection outperformed the other methods. First, it is not only the fact whether the anchor is *contaminated* (i.e. includes at least one DIF-item) but also the degree of contamination (i.e. the proportion of DIF-items in the anchor) that influences the performance of the anchor methods. The iterative forward method was often contaminated but the degree of contamination was low and the effect of the contaminated anchor items in the longer iterative forward anchor was marginal. Moreover, several anchor methods displayed an inflated false alarm rate – even if the anchor was DIF-free – that can be attributed to the fact that anchor items located by an anchor selection strategy display different characteristics compared to randomly chosen DIF-free items and may be exactly those items that again induce artificial DIF.

New anchor selection strategies were suggested in Chapter 7. The new suggestions allowed to further reduce the false alarm rates and achieved higher hit rates. The results also showed that the performance of the anchor selection may depend on the anchor length. Another finding was that the data generating process where all differences in the item parameters were simulated equal were harder for the anchor methods compared to situations where the mean differences were identical but the differences had a larger variance.

In Chapter 9, it was shown that quasi-Wald tests are suited for DIF detection. The test decision based on the quasi-Wald test for the first anchor candidate may be preferred over the decision that this item is declared DIF-free. However, if the first anchor candidate is DIF-free (which should be the case if the anchor selection works appropriately), similar results are expected.

### 10.1.2 Limits of this work

This work was limited to the analysis of differential item functioning in the Rasch model. In the Rasch model, the items are assumed to have the same discriminatory power. The items are, thus, characterized by the item difficulty parameters only. Other IRT models include further parameters for the discriminatory power (2pl model) or for guessing behavior (3pl model) or allow for more than two response categories.

Furthermore, the development of the anchor methods was evaluated using the Wald test for DIF detection. Other DIF tests such as the likelihood ratio test may provide better test results when the sample sizes are small.

In Chapter 6, it is recommended to analyze the risk of contamination together with the degree of contamination when new anchor methods are suggested. New anchor selection strategies were suggested in Chapter 7, but the risk and the degree of contamination were not discussed since the newly suggested methods yielded generally lower false alarm rates and did not display inversely u-shaped patterns for the false alarm rates.

The generalization of the anchor methods to paired multiple group comparisons in Chapter 8 was limited to the theoretical presentation of alternative ways. A simulation study is needed to assess the performance of the new suggestions. The generalizations of the anchor selections were limited to locate a predefined number of anchor items. Iterative procedures such as methods from the newly proposed iterative forward anchor class were not yet extended.

The simulation results were mainly based on the false alarm rate and the hit rate of the statistical tests. These measures reflect the fact whether or not the construction of a common scale for the item parameters was appropriate when the sample sizes are sufficiently large. However, the strategy to rely the final test decision about the classification of DIF- and DIF-free items on statistical significance tests is not indisputable. Generally, the power to detect DIF depends on the sample size. In large-scale assessments, the power to detect DIF should be reasonably high and even the correction for multiple testing may be advisable. However, if the sample size is small, DIF might be missed due to a low statistical power. This problem would even be reinforced if corrections for multiple testing are employed. Thus, the decision whether or not corrections for multiple testing are employed has to be considered carefully. Still, the results of this thesis allow to improve the construction of the common scale for the estimated item parameters and are, thus, not limited to statistical significance tests but also improve the accuracy of other measures such as descriptive statistics or effect size measures.

In general, all previously suggested procedures that were introduced in Chapter 3.3 including the methods investigated in this thesis (the Rasch trees, the methods based on logistic regression and the Wald test) are based on the null hypothesis that reflects no DIF and the alternative hypothesis that reflects the presence of DIF. This thesis follows this logic and, thus, allows to connect to the current state of research.

## 10.2 Future research

Future research questions arise from the limitations of this work. First, future research is intended that applies the model-based recursive partitioning algorithm to other IRT models including a location and a guessing parameter or allowing for more than two response categories (Strobl *et al.*, 2013).

Similar to the situation presented here for the Rasch model, additional steps to allow to answer the question which items display DIF and between which groups then also become relevant.

Item-wise DIF tests for other IRT models that include discrimination or guessing parameters or that allow for more response categories may be investigated.

It is to be shown, how the anchor methods developed in this thesis perform when other IRT models are the underlying data generating process. Furthermore, modifications of the anchor selection strategies might be necessary: González-Betanzos and Abad (2012) found that anchor items that displayed high discrimination parameters were better suited as anchor items. Future research may combine the strategies introduced in this thesis with new requirements. The ranking order of candidate anchor items may for example be modified in a way such that items with a low discriminatory power are excluded from the anchor candidates. The anchor methods may also be combined with other statistical tests for DIF.

Furthermore, it is yet to be explained how the anchor selection strategies perform in locating a suitable anchor for paired multiple group comparisons. First, the strategy to select a common set of anchor items needs to be compared with the results that occur if different anchor sets are selected for each paired comparison. The paired comparisons are similar to the two-group case discussed in Chapter 6 and in Chapter 7.

Second, a systematic evaluation of the aggregation rules presented in Chapter 8 is necessary. Third, it is important to compare different anchor selection strategies including those suggested in Chapter 7 under different conditions when more focal groups are present. Moreover, the methods from the iterative forward or backward anchor class and the quasi-Wald tests can be investigated for paired multiple group comparisons.

Addressing the problems related to the statistical significance testing with the null hypothesis of no DIF, alternative ways might be considered. First, instead of a final test decision, a DIF ranking order might be reported that no longer states whether an item displays significant DIF but reports which of the items display the lowest magnitudes of DIF and which items display the largest magnitudes of DIF.

Second, an alternative to overcome the problem that the tests presented in this thesis (see, again, Chapter 3) decide about the absence of DIF and not about the presence of DIF, is the usage of so called equivalence tests. Ongoing research considers these tests for the DIF analysis using the Mantel-Haenszel statistic (Casabianca and Lewis, 2012).

This development has the attractive property that the so called burden of proof (e.g. Walker and Nowacki, 2011) is shifted. While classical DIF tests decide about the absence of DIF, equivalence tests allow to decide about the presence of DIF. The burden of proof in the alternative hypothesis is then to show that the studied item does not display DIF which matches exactly the research hypothesis. An advantage is that DIF is then less likely missed due to small samples sizes compared to the classical DIF tests. In contrast to this, the burden of proof requires large sample sizes to underpin that the studied item does not display DIF. Thus, equivalence testing may prove helpful to set objective quality standards in test and questionnaire development.

One strategy to conduct an equivalence test that is referred to as "*[t]he simplest and most widely used approach*" (Walker and Nowacki, 2011, p. 193) is the two one-sided test (TOST) approach

(Westlake, 1981; Schuirmann, 1987). Generally, the null hypothesis states non-equivalence by means of two predefined equivalence margins ($\Lambda_1$ and $\Lambda_2$) $\mathrm{H}_0 : \mu_T - \mu_R \leq \Lambda_1$ or $\mu_T - \mu_R \geq \Lambda_2$, for example between the mean bioavailability of a test and reference product (Schuirmann, 1987) or the efficacy of a new and a current therapy (Walker and Nowacki, 2011). The alternative represents equivalence within the predefined margins $\mathrm{H}_1 : \Lambda_1 < \mu_T - \mu_R < \Lambda_2$.

Schuirmann (1987) points out that the null hypothesis of non-equivalence can be separated in two parts – $\mathrm{H}_{01} : \mu_T - \mu_R \leq \Lambda_1$ and $\mathrm{H}_{02} : \mu_T - \mu_R \geq \Lambda_2$ – and that the alternative $\mathrm{H}_1$ is empirically supported if both null hypothesis are rejected at the $\alpha$ level (or, equivalently, if the $(1-2\alpha)\cdot 100\%$ confidence interval for the difference is located between the equivalence margins).

Applied to the Wald test used in this thesis, the null hypothesis (of non-equivalence) reflects DIF in item $j$ (instead of no DIF in item $j$) and can be written as

$$\mathrm{H}_0 : \beta_j^{\mathrm{ref}} - \beta_j^{\mathrm{foc}} \leq -\Lambda_{\mathrm{DIF}} \text{ or } \beta_j^{\mathrm{ref}} - \beta_j^{\mathrm{foc}} \geq \Lambda_{\mathrm{DIF}},$$

assuming the symmetry property that DIF in either direction, in favor of the reference or of the focal group, is treated in the same way and that an equivalence margin $\Lambda_{\mathrm{DIF}}$ is given. The corresponding alternative hypothesis is

$$\mathrm{H}_1 : -\Lambda_{\mathrm{DIF}} < \beta_j^{\mathrm{ref}} - \beta_j^{\mathrm{foc}} < \Lambda_{\mathrm{DIF}}.$$

For DIF analysis, the null hypothesis can also be written in two parts, namely $\mathrm{H}_{01} : \beta_j^{\mathrm{ref}} - \beta_j^{\mathrm{foc}} \leq -\Lambda_{\mathrm{DIF}}$ and $\mathrm{H}_{02} : \beta_j^{\mathrm{ref}} - \beta_j^{\mathrm{foc}} \geq \Lambda_{\mathrm{DIF}}$.

Before the equivalence test can be carried out, research on how to choose the equivalence margin $\Lambda_{\mathrm{DIF}}$ is necessary. While Casabianca and Lewis (2012) discuss an effect size classification employed by ETS, one idea to obtain a substantially meaningful equivalence margin for the Wald test employed here could be to limit the maximum difference in the probability of solving the item between both groups to a certain value (e.g. such as to the values 5% or 15% that were discussed in the court proceedings, see Chapter 3).

To test both null hypothesis, now two item-wise one-sided DIF tests such as two one-sided Wald tests can be used. As discussed throughout this thesis, the DIF test results depend on the anchor methods what also holds for one-sided test procedures. Thus, the anchor methods developed in this thesis are not only applicable for the classical DIF tests or for descriptive and effect size measures but also for DIF analysis using equivalence tests such as the TOST approach.

## Literature

Ackerman TA (1992). “A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective.” *Journal of Educational Measurement*, **29**(1), 67–91.

Aigner DJ, Cain GG (1977). “Statistical Theories of Discrimination in Labor Markets.” *Industrial and Labor Relations Review*, **30**(2), 175–187.

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

Anderson EB (1973). “A Goodness of Fit Test for the Rasch Model.” *Psychometrika*, **38**(1), 123–140.

Andrews DWK (1993). “Tests for Parameter Instability and Structural Change with Unknown Change Point.” *Econometrica*, **61**(4), 821–856.

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

Angoff WH (1993). “Perspectives on Differential Item Functioning Methodology.” In PW Holland, H Wainer (eds.), *Differential Item Functioning*, chapter 1. Lawrence Erlbaum, Hillsdale, New Jersey.

Benjamini Y, Hochberg Y (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Berk R, Brown L, Zhao L (2010). “Statistical Inference After Model Selection.” *Journal of Quantitative Criminology*, **26**(2), 217–236.

Berk RA (2006). “An Introduction to Ensemble Methods for Data Analysis.” *Sociological Methods & Research*, **34**(3), 263–295.

Birnbaum A (1968). “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability.” In F Lord, M Novick (eds.), *Statistical Theories of Mental Test Scores*, pp. 397–479. Addison-Wesley, Reading.

Björklund A, Kjellström C (2002). “Estimating the Return to Investments in Education: How Useful Is the Standard Mincer Equation?” *Economics of Education Review*, **21**(3), 195–210.

Boeck PD, Bakker M, Zwitser R, Nivard M, Hofman A, Tuerlinckx F, Partchev I (2011). “The Estimation of Item Response Models with the lmer Function from the lme4 Package in R.” *Journal of Statistical Software*, **39**(12), 1–28.

Bolt DM, Hare RD, Vitale JE, Newman JP (2004). “A Multigroup Item Response Theory Analysis of the Psychopathy Checklist – Revised.” *Psychological Assessment*, **16**(2), 155–168.

Boulesteix AL (2006). "Maximally Selected Chi-Square Statistics and Binary Splits of Nominal Variables." *Biometrical Journal*, **48**(5), 838–848.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Bretz F, Hothorn T, Westfall P (2011). *Multiple Comparisons Using R*. Chapman & Hall/CRC Press, Boca Raton, Florida, USA.

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

Casabianca J, Lewis C (2012). "Equivalence Testing for Differential Item Functioning: Standard and Bayesian Approaches." Carnegie Mellon University Working Paper.

Chaudhuri P, Lo WD, Loh WY, Yang CC (1995). "Generalized Regression Trees." *Statistica Sinica*, **5**(2), 641–666.

Cheung GW, Rensvold RB (1999). "Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method." *Journal of Management*, **25**(1), 1–27.

Cohen A, Bolt D (2005). "A Mixture Model Analysis of Differential Item Functioning." *Journal of Educational Measurement*, **42**(3), 133–148.

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

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

Diekmann A (2007). *Empirische Sozialforschung. Grundlagen Methoden Anwendungen*. 18. edition. Rowohlt Taschenbuch Verlag, Reinbek.

Dobra A, Gehrke J (2001). "Bias Correction in Classification Tree Construction." In CE Brodley, AP Danyluk (eds.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA*, pp. 90–97. Morgan Kaufmann.

Dorans NJ (1989). "Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel-Haenszel Method." *Applied Measurement in Education*, **2**(3), 217–233.

Drasgow F (1987). "Study of the Measurement Bias of Two Standardized Psychological Tests." *Journal of Applied Psychology*, **72**(1), 19–29.

Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K (2006). "Identification of Differential Item Functioning Using Item Response Theory and the Likelihood-based Model Comparison Approach. Application to the Mini-Mental State Examination." *Medical Care*, **44**(22), 134–142.

Eggen T, Verhelst N (2006). "Loss of Information in Estimating Item Parameters in Incomplete Designs." *Psychometrika*, **71**(2), 303–322.

Ellis BB (1989). "Differential Item Functioning: Implications for Test Translations." *Journal of Applied Psychology*, **74**(6), 912 – 921.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

Finch H (2005). "The MIMIC Model As a Method for Detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio." *Applied Psychological Measurement*, **29**(4), 278–295.

Finch WH, French BF (2007). "Detection of Crossing Differential Item Functioning: A Comparison of Four Methods." *Educational and Psychological Measurement*, **67**(4), 565–582.

Firth D (2003). "Overcoming the Reference Category Problem in the Presentation of Statistical Models." *Sociological Methodology*, **33**(1), 1–18.

Firth D (2012). *qvcalc: Quasi Variances for Factor Effects in Statistical Models*. R package version 0.8-8.

Firth D, De Menezes RX (2004). "Quasi-variances." *Biometrika*, **91**(1), 65–80.

Fischer G (1981). "On the Existence and Uniqueness of Maximum-Likelihood Estimates in the Rasch Model." *Psychometrika*, **46**(1), 59–77.

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

Fischer GH (1974). *Einführung in die Theorie Psychologischer Tests. Grundlagen und Anwendungen*. Verlag Hans Huber, Bern, Stuttgart, Wien.

Fischer GH (1995). "Derivations of the Rasch Model." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 2. Springer, New York.

Flier HVD, Mellenbergh GJ, Adèr HJ, Wijn M (1984). "An Iterative Item Bias Detection Method." *Journal of Educational Measurement*, **21**(2), 131–145.

Frederickx S, Tuerlinckx F, De Boeck P, Magis D (2010). "RIM: A Random Item Mixture Model to Detect Differential Item Functioning." *Journal of Educational Measurement*, **47**(4), 432–457.

French BF, Maller SJ (2007). “Iterative Purification and Effect Size Use with Logistic Regression for Differential Item Functioning Detection.” *Educational and Psychological Measurement*, **67**(3), 373–393.

Gelin M, Carleton B, Smith M, Zumbo B (2004). “The Dimensionality and Gender Differential Item Functioning of the Mini Asthma Quality of Life Questionnaire (MiniAQLQ).” *Social Indicators Research*, **68**(1), 91–105.

Gelin MN, Zumbo BD (2003). “Differential Item Functioning Results May Change Depending on How an Item Is Scored: An Illustration with the Center For Epidemiologic Studies Depression Scale.” *Educational and Psychological Measurement*, **63**(1), 65–74.

Glas CAW (1998). “Detection of Differential Item Functioning using Lagrange Multiplier Tests.” *Statistica Sinica*, **8**(3), 647–667.

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

Gómez-Benito J, Hidalgo MD, Guilera G (2010). “Bias in Measurement Instruments. Fair Tests.” *Papeles del Psiclogo*, **31**(1), 75–84.

González-Betanzos F, Abad FJ (2012). “The Effects of Purification and the Evaluation of Differential Item Functioning with the Likelihood Ratio Test.” *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **8**(4), 134–145.

Gujarati DN (2003). *Basic Econometrics*. 4. edition. McGraw-Hill, Boston.

Gustafsson J (1980). “Testing and Obtaining Fit of Data in the Rasch Model.” *British Journal of Mathematical and Statistical Psychology*, **33**(2), 205–233.

Hambleton RK, Rogers HJ (1989). “Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods.” *Applied Measurement in Education*, **2**(4), 313–334.

Hambleton RK, Swaminathan H, Rogers HJ (1991). *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park.

Hancock G, Samuelsen K (eds.) (2007). *Advances in Latent Variable Mixture Models*. Information Age Publishing, Charlotte.

Hannöver W, Richard M, Hansen NB, Martinovich Z, Kordy H (2002). “A Classification Tree Model for Decision-Making in Clinical Practice: An Application Based on the Data of the German Multicenter Study on Eating Disorders, Project TR-EAT.” *Psychotherapy Research*, **12**(4), 445–461.

Hastie T, Tibshirani R, Friedman JH (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. 2. edition. Springer, New York.

Hidalgo-Montesinos MD, Lopez-Pina JA (2002). "Two-Stage Equating in Differential Item Functioning Detection under the Graded Response Model with the Raju Area Measures and the Lord Statistic." *Educational and Psychological Measurement*, **62**(1), 32–44.

Hochberg Y, Tamhane A (eds.) (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Holland PW, Thayer DT (1988). "Differential Item Performance and the Mantel-Haenszel Procedure." In H Wainer, HI Braun (eds.), *Test Validity*, chapter 9. Lawrence Erlbaum, Hillsdale, New Jersey.

Holland PW, Wainer H (eds.) (1993). *Differential Item Functioning*. Lawrence Erlbaum, Hillsdale, New Jersey.

Hothorn T, Hornik K, van de Wiel M, Zeileis A (2006a). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263.

Hothorn T, Hornik K, Zeileis A (2006b). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Hothorn T, Lausen B (2003). "On the Exact Distribution of Maximally Selected Rank Statistics." *Computational Statistics & Data Analysis*, **43**(2), 121–137.

Jodoin MG, Gierl MJ (2001). "Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection." *Applied Measurement in Education*, **14**(4), 329–349.

Kelderman H, MacReady G (1990). "The Use of Loglinear Models for Assessing Differential Item Functioning across Manifest and Latent Examinee Groups." *Journal of Educational Measurement*, **27**(4), 307–327.

Kim J, Oshima TC (2013). "Effect of Multiple Testing Adjustment in Differential Item Functioning Detection." *Educational and Psychological Measurement*, **73**(3), 458–470.

Kim SH, Cohen AS (1995). "A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning." *Applied Measurement in Education*, **8**(4), 291–312.

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

Kim SH, Cohen AS, Kim HO (1994). "An Investigation of Lord's Procedure for the Detection of Differential Item Functioning." *Applied Psychological Measurement*, **18**(3), 217–228.

Kim SH, Cohen AS, Park TH (1995). "Detection of Differential Item Functioning in Multiple Groups." *Journal of Educational Measurement*, **32**(3), 261–276.

Kitsantas P, Moore T, Sly D (2007). "Using Classification Trees to Profile Adolescent Smoking Behaviors." *Addictive Behaviors*, **32**(1), 9–23.

Kopf J, Augustin T, Strobl C (2013a). "The Potential of Model-Based Recursive Partitioning in the Social Sciences: Revisiting Ockam's Razor." In J McArdle, G Ritschard (eds.), *Contemporary Issues in Exploratory Data Mining*, chapter 3. Routeledge. To appear in 2013.

Kopf J, Strobl C (2013). "Detecting Non-uniform DIF with Rasch Trees." To be submitted.

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

Kopf J, Zeileis A, Strobl C (2013c). "Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches." *Technical Report 150*, Department of Statistics, LMU Munich.

Koziol J (1991). "On Maximally Selected Chi-Square Statistics." *Biometrics*, **47**(4), 1557–1561.

Kuckulenz A, Zwick T (2005). "Heterogene Einkommenseffekte Betrieblicher Weiterbildung." *Die Betriebswirtschaft*, **65**(3), 258–275.

Leeb H, Pötscher BM (2005). "Model Selection and Inference: Facts and Fiction." *Econometric Theory*, **21**(1), 21–59.

Leisch F (2004). "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R." *Journal of Statistical Software*, **11**(8), 1–18.

Lesnoff M, Lancelot R (2012). *aod: Analysis of Overdispersed Data*. R package version 1.3.

Li HH, Stout W (1996). "A New Procedure for Detection of Crossing DIF." *Psychometrika*, **61**(4), 647–677.

Li KC, Lue HH, Chen CH (2000). "Interactive Tree-Structured Regression via Principal Hessian Directions." *Journal of the American Statistical Association*, **95**(450), 547–560.

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

Linn RL, Levine MV, Hastings CN, Wardrop JL (1981). "Item Bias in a Test of Reading Comprehension." *Applied Psychological Measurement*, **5**(2), 159–173.

Liou M (1994). "More on the Computation of Higher-Order Derivatives on the Elementary Symmetric Functions in the Rasch Model." *Applied Psychological Measurement*, **18**(1), 53–62.

Loh WY (2002). "Regression Trees with Unbiased Variable Selection and Interaction Detection." *Statistica Sinica*, **12**(2), 361–386.

Lopez Rivas GE, Stark S, Chernyshenko OS (2009). "The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test." *Applied Psychological Measurement*, **33**(4), 251–265.

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

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

Magis D, Raîche G, Béland S, Gérard P (2011). "A Generalized Logistic Regression Procedure to Detect Differential Item Functioning Among Multiple Groups." *International Journal of Testing*, **11**(4), 365–386.

Maij-de Meij A, Kelderman H, Van der Flier H (2008). "Fitting a Mixture Item Response Theory Model to Personality Questionnaire Data: Characterizing Latent Classes and Investigating Possibilities for Improving Prediction." *Applied Psychological Measurement*, **32**(8), 611–631.

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), 1–20.

Mair P, Hatzinger R, Maier MJ (2012). *eRm: Extended Rasch Modeling.* R package version 0.15-0.

Maller SJ (2001). "Differential Item Functioning in the Wisc-III: Item Parameters for Boys and Girls in the National Standardization Sample." *Educational and Psychological Measurement*, **61**(5), 793–817.

Mantel N, Haenszel W (1959). "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease." *Journal of the National Cancer Institute*, **22**(4), 719–748.

Marcus R, Peritz E, Gabriel K (1976). "Closed Testing Procedures with Special Reference to Ordered Analysis of Variance." *Biometrika*, **63**(3), 655–660.

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

McLean LD, Ragsdale RG (1983). "The Rasch Model for Achievement Tests: Inappropriate in the Past, Inappropriate Today, Inappropriate Tomorrow." *Canadian Journal of Education / Revue canadienne de l'éducation*, **8**(1), 71–76.

Meade AW, Lautenschlager GJ (2004). "A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance." *Organizational Research Methods*, **7**(4), 361–388.

Mellenbergh GJ (1982). "Contingency Table Models for Assessing Item Bias." *Journal of Educational Statistics*, **7**(2), 105–118.

Merkle EC, Zeileis A (2012). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*. Forthcoming.

Miller R, Siegmund D (1982). "Maximally Selected Chi Square Statistics." *Biometrics*, **38**(4), 1011–1016.

Millsap RE, Everson HT (1993). "Methodology Review: Statistical Approaches for Assessing Measurement Bias." *Applied Psychological Measurement*, **17**(4), 297–334.

Mincer JA (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research, New York.

Mislevy R, Verhelst N (1990). "Modeling Item Responses when Different Subjects Employ Different Solution Strategies." *Psychometrika*, **55**(2), 195–215.

Molenaar IW (1995a). "Estimation of Item Parameters." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 3. Springer, New York.

Molenaar IW (1995b). "Some Background for Item Response Theory and the Rasch Model." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 1. Springer, New York.

Morgan JN, Sonquist JA (1963). "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association*, **58**(302), 415–434.

Narayanan P, Swaminathan H (1996). "Identification of Items That Show Nonuniform DIF." *Applied Psychological Measurement*, **20**(3), 257–274.

OECD (2009). *PISA Data Analysis Manual: SPSS®*. 2. edition. OECD Publishing, Paris.

Osterlind SJ, Everson HT (2009). *Differential Item Functioning*. 2. edition. SAGE Publications, Thousand Oaks.

Pae TI (2004). "DIF for Examinees with Different Academic Backgrounds." *Language Testing*, **21**(1), 53–73.

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

Park DG, Lautenschlager GJ (1990). "Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification." *Applied Psychological Measurement*, **14**(2), 163–173.

Pedraza O, Graff-Radford N, Smith G, Ivnik R, Willis F, Petersen R, Lucas J (2009). "Differential Item Functioning of the Boston Naming Test in Cognitively Normal African American and Caucasian Older Adults." *Journal of the International Neuropsychological Society*, **15**(5), 758–768.

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

Perkins A, Stump T, Monahan P, McHorney C (2006). "Assessment of Differential Item Functioning for Demographic Comparisons in the MOS SF-36 Health Survey." *Quality of Life Research*, **15**, 331–348.

Phelps ES (1972). "The Statistical Theory of Racism and Sexism." *The American Economic Review*, **62**(4), 659–661.

Quinlan JR (1993). *C4.5: Programms for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco.

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

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

Raykov T, Marcoulides GA, Lee CL, Chang C (2013). "Studying Differential Item Functioning via Latent Variable Modeling: A Note on a Multiple-Testing Procedure." *Educational and Psychological Measurement*. Online first.

Richardson LF (1958). "Mathematics of War and Foreign Politics." In JR Newman (ed.), *The World of Mathematics*. Simon and Schuster, New York.

Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G, Lanfranchi G (2003). "Pattern Recognition in Gene Expression Profiling Using DNA Array: A Comparison Study of Different Statistical Methods Applied to Cancer Classification." *Human Molecular Genetics*, **12**(8), 823–836.

Rost J (1990). "Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis." *Applied Psychological Measurement*, **14**(3), 271–282.

Rost J, von Davier M (1995). "Mixture Distribution Rasch Models." In GH Fischer, IW Molenaar (eds.), *Rasch Models – Foundations, Recent Developments, and Applications*, chapter 14. Springer, New York.

Savage LJ (1951). "The Theory of Statistical Decision." *Journal of the American Statistical Association*, **46**(253), 55–67.

Scheuneman JD, Bleistein CA (1989). "A Consumer's Guide to Statistics for Identifying Differential Item Functioning." *Applied Measurement in Education*, **2**(3), 255–275.

Schnell R, Hill PB, Esser E (2008). *Methoden der Empirischen Sozialforschung.* 8. edition. Oldenburg Verlag, München, Wien.

Schuirmann DJ (1987). "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics*, **15**(6).

Setodji CM, Reise SP, Morales LS, Fongwa MN, Hays RD (2011). "Differential Item Functioning by Survey Language Among Older Hispanics Enrolled in Medicare Managed Care: A New Method for Anchor Item Selection." *Medical Care*, **49**(5), 461–468.

Shealy R, Stout W (1993a). "A Model-Based Standardization Approach That Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF." *Psychometrika*, **58**(2), 159–194.

Shealy RT, Stout WF (1993b). "An Item Response Theory Model for Test Bias and Differential Test Functioning." In PW Holland, H Wainer (eds.), *Differential Item Functioning*, chapter 10. Lawrence Erlbaum, Hillsdale, New Jersey.

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

Shih YS (2004). "A Note on Split Selection Bias in Classification Trees." *Computational Statistics & Data Analysis*, **45**(3), 457–466.

Smit J, Kelderman H, Van der Flier H (2000). "The Mixed Birnbaum Model: Estimation using Collateral Information." *Methods of Psychological Research Online*, **5**, 1–13.

Stark S, Chernyshenko OS, Drasgow F (2006). "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology*, **91**(6), 1292–1306.

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250.

Strobl C (2008). *Statistical Issues in Machine Learning. Towards Reliable Split Selection and Variable Importance Measures*. Ph.D. thesis, Department of Statistics, LMU Munich.

Strobl C (2013). "Data Mining." In T Little (ed.), *The Oxford Handbook on Quantitative Methods*, chapter 29. Oxford University Press USA.

Strobl C, Boulesteix AL, Augustin T (2007). “Unbiased Split Selection for Classification Trees Based on the Gini Index.” *Computational Statistics & Data Analysis*, **52**(1), 483–501.

Strobl C, Kopf J, Zeileis A (2010). “Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben.” In S Trepte, M Verbeet (eds.), *Wissenswelten des 21. Jahrhunderts – Erkenntnisse aus dem Studentenpisa-Test des SPIEGEL*, pp. 255–272. VS Verlag, Wiesbaden.

Strobl C, Kopf J, Zeileis A (2013). “Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model.” *Psychometrika*. Accepted for publication.

Strobl C, Malley J, Tutz G (2009). “An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests.” *Psychological Methods*, **14**(4), 323–348.

Strobl C, Wickelmaier F, Zeileis A (2011). “Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning.” *Journal of Educational and Behavioral Statistics*, **36**(2), 135–153.

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

Thissen D (2001). *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. Unpublished manuscript, University of North Carolina, Chapel Hill.

Thissen D, Steinberg L, Kuang D (2002). “Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons.” *Journal of Educational and Behavioral Statistics*, **27**(1), 77–83.

Thissen D, Steinberg L, Wainer H (1988). “Use of Item Response Theory in the Study of Group Differences in Trace Lines.” In H Wainer, HI Braun (eds.), *Test Validity*, chapter 10. Lawrence Erlbaum, Hillsdale, New Jersey.

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

Tutz G, Schauberger G (2012). “A Penalty Approach to Differential Item Functioning in Rasch Models.” *Technical Report 134*, Department of Statistics, LMU Munich.

Ünlü A (2011). “A Note on the Connection Between Knowledge Structures and Latent Class Models.” *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **7**(2), 63–67.

Vermunt J (2010). “Latent Class Models.” In P Peterson, E Baker, B McGaw (eds.), *International Encyclopedia of Education*, volume 7 of *International Encyclopedia of Education*.

Wagner GG, Frick JR, Schupp J (2007). "The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements." *Schmollers Jahrbuch*, **127**(1), 139–169.

Walker E, Nowacki AS (2011). "Understanding Equivalence and Noninferiority Testing." *Journal of General Internal Medicine*, **26**(2), 192–196.

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

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

Wang WC, Su YH (2004). "Effects of Average Signed Area Between Two Item Characteristic Curves and Test Purification Procedures on the DIF Detection via the Mantel-Haenszel Method." *Applied Measurement in Education*, **17**(2), 113–144.

Wang WC, Yeh YL (2003). "Effects of Anchor Item Methods on Differential Item Functioning Detection with the Likelihood Ratio Test." *Applied Psychological Measurement*, **27**(6), 479–498.

Wang Y, Witten IH (1997). "Induction of Model Trees for Predicting Continuous Classes." In *Proceedings of the European Conference on Machine Learning*. University of Economics, Faculty of Informatics and Statistic, Prague.

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

Westlake WJ response to Kirkwood TBL (1981). "Bioequivalence Testing – A Need to Rethink." *Biometrics*, **37**(3), 589–594.

Wicherts JM, Dolan CV, Hessen DJ (2005). "Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance." *Journal of Personality and Social Psychology*, **89**(5), 696–716.

Williams VSL, Jones LV, Tukey JW (1999). "Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement." *Journal of Educational and Behavioral Statistics*, **24**(1), 42–69.

Woods C, Oltmanns T, Turkheimer E (2009). "Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality." *Journal of Psychopathology and Behavioral Assessment*, **31**, 320–330.

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

Woods CM, Cai L, Wang M (2013). "The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT." *Educational and Psychological Measurement*, **73**(3), 532–547.

Yee TW (2010a). "The VGAM Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34.

Yee TW (2010b). *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.8-1.

Yee TW, Hastie TJ (2003). "Reduced-rank Vector Generalized Linear Models." *Statistical Modelling*, **3**(1), 15–41.

Yee TW, Wild CJ (1996). "Vector Generalized Additive Models." *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(3), 481–493.

Zeileis A (2005). "A Unified Approach to Structural Change Tests Based on ML Scores, *F* Statistics, and OLS Residuals." *Econometric Reviews*, **24**(4), 445–466.

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**(4), 488–508.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514.

Zeileis A, Hothorn T, Hornik K (2010). *Party with the Mob: Model-Based Recursive Partitioning in R*. R package version 0.9-9999.

Zeileis A, Strobl C, Wickelmaier F, Kopf J (2011). *psychotree: Recursive Partitioning Based on Psychometric Models*. R package version 0.12-1.

Zhang H, Yu CY, Singer B, Xiong M (2001). "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data." *Proceedings of the National Academy of Sciences of the United States of America*, **98**(12), 6730–6735.

Zumbo BD (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

# Eidesstattliche Versicherung

**(Gemäß Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)**

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 14. August 2013 Julia Kopf