FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Mgr. Adéla Hladká

Statistical models for detection of differential item functioning

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis: RNDr. Patrícia Martinková, PhD.
Consultant of the doctoral thesis: doc. Mgr. Michal Kulich, PhD.
Study programme: Probability and Statistics, Econometrics and Financial Mathematics

Prague 2021

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to express my gratitude to my supervisor Patrícia Martinková who was profusely helpful and offered invaluable assistance, support, and guidance. I would also like to thank Michal Kulich, Marek Brabec, and Marek Omelka for their valuable advices and recommendations.

My special thanks belongs to my husband Miroslav and to all my family members for their constant support in my studies.

This thesis was partly supported by the Charles University (projects SVV 260580 and PRIMUS/17/HUM/11), by the Czech Science Foundation under the grant No. 21-03658S, and by the long-term strategic development financing of the Institute of Computer Science of the Czech Academy of Sciences (RVO: 67985807).

Title: Statistical models for detection of differential item functioning

Author: Mgr. Adéla Hladká

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Patrícia Martinková, PhD., Institute of Computer Science of the Czech Academy of Sciences

Abstract: This thesis focuses on topic of Differential Item Functioning (DIF), a phenomenon that can arise in various contexts of educational, psychological, or health-related multi-item measurements. We discuss several statistical methods and models to detect DIF among dichotomous, ordinal, and nominal items.

In the first part, generalized logistic regression models for DIF detection among dichotomous items are introduced, which account for possibility of guessing and/or inattention. Techniques for estimation of item parameters are presented, including a newly proposed algorithm based on a parametric link function. Two simulation studies are presented. The first compares the generalized logistic regression models to other widely used DIF detection methods. The second illustrates differences between the techniques to estimate item parameters. Implementation of the models into the **R** software and its **difNLR** package is illustrated.

In the second part, generalized logistic regression models for DIF detection among polytomous items are discussed. Cumulative logit, adjacent category logit, and nominal models are introduced together with the maximum likelihood method to estimate item parameters and with examples of implementation in the **difNLR** package.

The third part deals with a nonparametric comparison of regression curves for DIF detection based on kernel smoothing. We discuss several settings and we newly propose an estimate of an optimal weight function for a test statistic to identify DIF. Nonparametric approaches are compared to the logistic regression method in a simulation study.

In the fourth and last part, further topics of DIF detection are discussed, including item purification, multiple comparison corrections, and DIF effect sizes. Different approaches are compared in a complex simulation study on three of the most used DIF detection methods.

Keywords: differential item functioning, generalized logistic regression, nonparametric methods, differential distractor functioning

Contents

Introduction

This thesis deals with the topic of Differential Item Functioning (DIF). DIF is a well-known phenomenon that can arise in various contexts including educational measurement, admission tests, mental health inventories, and other types of behavioral assessments (Osterlind & Everson, 2009; Penfield & Camilli, 2006). An item is said to function differently (or, in short, to be a DIF item) when two respondents with the same underlying latent trait but from various social groups or with distinct characteristics have different probability of endorsing or correctly answering an item in multi-item measurement. Latent trait can be knowledge, ability, or health-related outcome such as depression or quality of life, and it is often estimated by total test score. The group membership can be characterized for example by gender, race, or socio-economic status.

Consider dichotomous response Y_i to a particular item i on the test, where $Y_i = 1$ means correct answer or endorsement and $Y_i = 0$ indicates incorrect answer or opposition. Let θ be an underlying latent trait intended to be measured by the test. Further, suppose that we are concerned with comparing the conditional probability of Y_i for two different respondent groups, described by variable G, where $G = 0$ stands for a *reference group* (usually majority) and $G = 1$ is a *focal group* (often minority or potentially disadvantaged group).

In case that the conditional probability of the response Y_i is dependent only on the underlying latent trait θ and independent on the grouping variable G, i.e.,

$$\mathsf{P}(Y_i = 1|\theta, G = 0) = \mathsf{P}(Y_i = 1|\theta, G = 1), \quad \forall \theta \in \Theta,$$

where Θ is the continuum of the latent trait θ, we conclude there is no DIF in item i (Figure 1A).

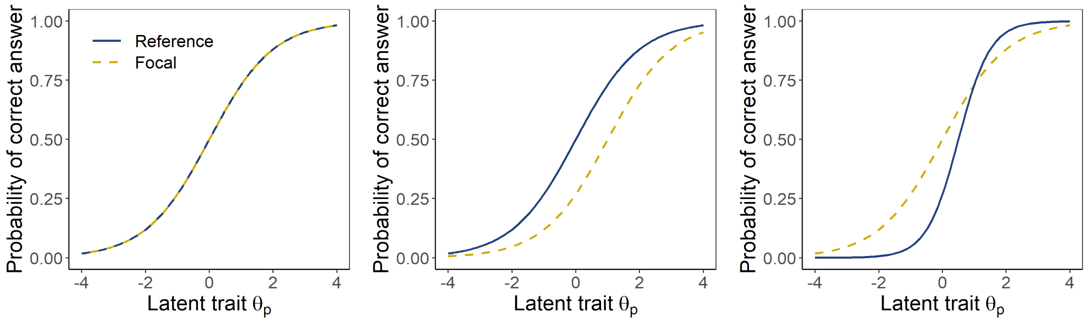
Generally, two types of DIF are distinguished in the literature (see for example Penfield & Camilli, 2006). *Uniform DIF* captures a situation in which the item of interest provides a constant relative advantage for the same group, regardless of the level of the latent trait $\theta \in \Theta$. In other words, for all levels of the latent ability θ, the ratio of odds of answering given item correctly for the focal group to the odds of answering this item correctly for the reference group is a constant:

$$\frac{\mathsf{P}(Y_i = 1|\theta, G = 1)/\,\mathsf{P}(Y_i = 0|\theta, G = 1)}{\mathsf{P}(Y_i = 1|\theta, G = 0)/\,\mathsf{P}(Y_i = 0|\theta, G = 0)} = \zeta_i \in \mathbb{R}, \quad \forall \theta \in \Theta,$$

see also Figure 1B. *Nonuniform DIF* then describes a situation when the conditional dependency between the item response and a group membership changes across the continuum of θ. This may even result in a situation when respondents form one group are advantaged by the item for some of the levels of the latent trait (for instance $\theta \in \Theta_1$) while they are disadvantaged for other levels ($\theta \in \Theta \backslash \Theta_1$), i.e.,

$$\mathsf{P}(Y_i = 1|\theta, G = 0) \geq \mathsf{P}(Y_i = 1|\theta, G = 1), \quad \forall \theta \in \Theta_1,$$
$$\text{and}$$
$$\mathsf{P}(Y_i = 1|\theta, G = 0) < \mathsf{P}(Y_i = 1|\theta, G = 1), \quad \forall \theta \in \Theta \backslash \Theta_1,$$

which is also called *crossing nonuniform DIF* (Figure 1C).

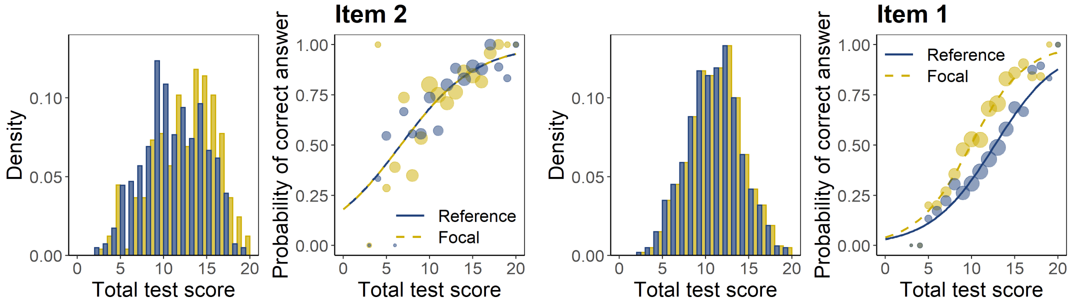


(A) No DIF. (B) Uniform DIF. (C) Nonuniform DIF.

Figure 1: Definition of DIF.

While DIF typically refers to the differences in probabilities of correctly answering or endorsing an item with respect to group membership, the covariates of interest may, however, be more complex including combination of categorical and continuous variables with a hierarchical or more complex structure.

DIF analysis provides useful guidance in detecting potentially biased items which can be a possible threat to the fairness and validity in the measurement. DIF analysis should therefore be a routine part of validation process of educational, psychological, or health-related multi-item tests. Sometimes, fairness is incorrectly examined by comparing total test scores or item scores separately. Martinková et al. (2017) provided powerful simulated as well as real-data examples showing that between-group differences in total score do not necessarily result into DIF while DIF may be present even when the distribution of total scores is identical in the two groups. Using the Homeostasis Concept Inventory (HCI) dataset (McFarland et al., 2017), there is a significant difference in the total test scores between the two groups of respondents (Figure 2A, left), however there is no DIF item (Figure 2A, right). On the other hand, considering the Graduation Management Admission Test (GMAT) dataset which was simulated with the exact match of the distribution of the total test scores (Figure 2B, left), two items were identified as functioning differently (Figure 2B, right) and being potentially unfair.

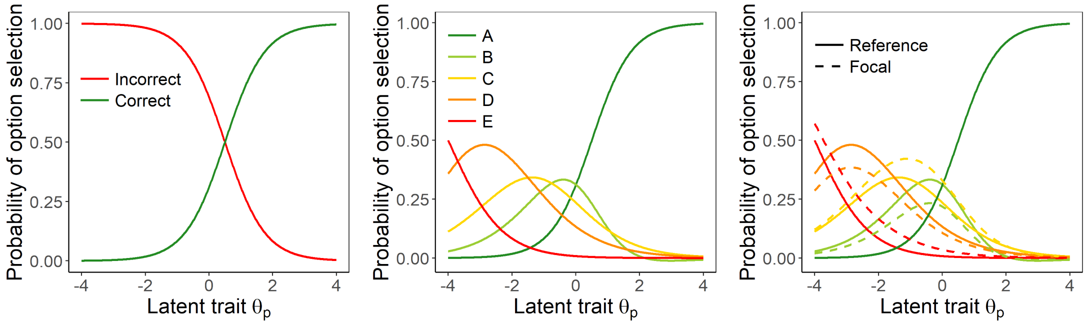


(A) HCI dataset. (B) GMAT dataset.

Figure 2: DIF vs. between-group differences in distribution of total scores.

Besides uncovering potential unfairness, DIF may also point to misconceptions held by groups. Moreover, when generalizing the concept of DIF to longitudinal setting, the DIF in change can provide proofs of the instructional or treatment sensitivity, even in cases when the differences in gains are not detected in overall score (Martinková, Hladká, & Potužníková, 2020).

The concept of DIF can be easily extended to account for ordinal responses. Moreover, it can be generalized to nominal responses in a way that it captures differences not only in the probabilities of correct answers between two groups (Figure 3A) but also in other answer options (also called *distractors*, Figure 3B). In such a case, the group differences are called Differential Distractor Functioning (DDF). Formally, DDF then refers to a situation when two respondents from different social groups, but with the same level of the underlying latent trait, have different probabilities of selection of answer options in the given test item. The DDF may even refer to situation when there is no difference in probability of correct answer (or in endorsing the item), therefore no DIF in dichotomized item present, but there is a difference in the probability of selecting some specific distractors (Figure 3C).



(A) Binary evaluation of an item.

(B) Options on a multiple-choice item.

(C) DDF but no DIF.

Figure 3: Illustration of DDF.

State of the art of DIF detection

Many statistical methods were developed to identify DIF items, using either Item Response Theory (IRT) models or score-based techniques (here also referenced as non-IRT) while both branches are still being studied intensively (Berger & Tutz, 2016; Cho, Suh, & Lee, 2016; Penfield, Gattamorta, & Childs, 2009). In this part we introduce a number of often used DIF detection approaches with emphasis on those used in the simulation studies presented in this thesis (see Sections 1.5, 1.6, 3.2, and 4.2). Recent and more detailed reviews of DIF detection methods can be found for example in Magis, Béland, Tuerlinckx, and De Boeck (2010), Osterlind and Everson (2009), or Penfield and Camilli (2006).

IRT methods for DIF detection

IRT covers class of nonlinear mixed effect models in which latent trait (ability) is often estimated as respondents' random effect and the item parameters are usually treated as fixed effects. Furthermore, respondents' abilities as well as parameters of all items are estimated simultaneously and not by separate models for each item of the test (Bock & Moustaki, 2006).

The underlying nonlinear mixed effect model framework (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003) and the concept of latent variable may be a bit more complex to understand and harder to implement without specialized software, especially when higher number of item parameters is considered. Moreover, the IRT models are generally known to be computationally demanding and a large sample size is often required (Kim & Oshima, 2013).

The most widely used unidimensional models for dichotomous responses are (generalized) logistic models with one to four item parameters. We further denote them as 1-4 Parameter Logistic (PL) IRT models. The 4PL IRT model (Barton & Lord, 1981) is given by the equation

$$\mathsf{P}(Y_{pi} = 1|\theta_p) = c_i + (d_i - c_i)\frac{e^{a_i(\theta_p - b_i)}}{1 + e^{a_i(\theta_p - b_i)}}, \tag{1}$$

where $\theta_p \in \mathbb{R}$ is an ability of person p and parameters $a_i, b_i \in \mathbb{R}$, and $c_i, d_i \in [0, 1]$ represent discrimination, difficulty, probability of guessing, and a parameter related to the probability of inattention in item i. Rasch and 1-3PL models are special cases of the 4PL model (1), where some of the parameters are fixed to a certain value. For example, 2PL IRT model is a 4PL IRT model with $c_i = 0$ and $d_i = 1$, while Rasch model additionally constraints $a_i = 1$. For estimation of item parameters and respondents' abilities, the marginal maximum likelihood method (Bock & Aitkin, 1981) is typically used.

Although potential applications of IRT models are much broader, they are also widely used in the DIF detection. To account for group-based differences in the item responses and to test for DIF in the IRT framework, two IRT models are often fitted, each for one group, and then the estimated parameters are re-scaled (Candell & Drasgow, 1988; Lautenschiager & Park, 1988).

The mostly used DIF detection procedures within the IRT framework include for example Lord's test (Lord, 1980), likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), or Raju's test based on area between Item Characteristic Curve (ICC)s (Raju, 1988, 1990).

Lord's test. Lord's test (Lord, 1980) is one of the most popular methods for testing DIF within the IRT models. Lord's approach uses test statistic of Wald's type with the null hypothesis assuming equal item parameters in both groups. Related test statistic W_i for item i and 4PL IRT model (1) has the following form:

$$W_i = (\gamma_{i0} - \gamma_{i1})' \left(\boldsymbol{\Sigma}_{i0} + \boldsymbol{\Sigma}_{i1}\right)^{-1} (\gamma_{i0} - \gamma_{i1}), \tag{2}$$

where $\gamma_{i0} = (a_{i0}, b_{i0}, c_{i0}, d_{i0})$ are parameters for the reference group for item i, $\gamma_{i1} = (a_{i1}, b_{i1}, c_{i1}, d_{i1})$ parameters for the focal group, and $\boldsymbol{\Sigma}_{i0}$ and $\boldsymbol{\Sigma}_{i1}$ the corresponding covariance matrices (Lord, 1980, p. 223). The test statistic W_i has an asymptotic chi-square distribution with 4 degrees of freedom when considering 4PL IRT model (1).

Raju's area. Raju's test estimates area between the ICCs for the two groups to detect DIF (Raju, 1988, 1990). This method is applicable for models in which the asymptotes of the underlying IRT models have the same value for both groups. This holds for example for the Rasch and the 1-2PL IRT models, however, it is also applicable for the 3PL and 4PL models in which the asymptotes are fixed to the same value or are estimated simultaneously for both groups (see, for example, Magis et al., 2010).

Consider the 4PL IRT model (1) for both the groups with the same values of the asymptotes (i.e., $c_{i0} = c_{i1} = c_i$ and $d_{i0} = d_{i1} = d_i$), where index 0 refers to the reference group and index 1 to the focal group. Then the Signed Area (SA) and Unsigned Area (UA) between the two ICCs for the item i are given by

$$\begin{aligned} \mathrm{SA}_i &= (d_i - c_i)(b_{i1} - b_{i0}), \\ \mathrm{UA}_i &= (d_i - c_i)\left|\frac{2(a_{i1} - a_{i0})}{Da_{i1}a_{i0}} \log\left(1 + \exp\left(\frac{Da_{i1}a_{i0}(b_{i1} - b_{i0})}{a_{i1} - a_{i0}}\right)\right) - (b_{i1} - b_{i0})\right|, \end{aligned} \tag{3}$$

where $D = 1.7$ is a scaling constant, see Raju (1988) for details. While the SA is easy to compute, it is a function of parameters b_{i0}, b_{i1}, c_i, and d_i only, therefore, it can be misleading when parameters a_{i0} and a_{i1} differ (see, for example, Osterlind & Everson, 2009). Thus, the SA is more suitable when uniform DIF is expected, while the UA may be more appropriate in case when the ICCs cross.

It can be shown that the estimated SA for item i, i.e., area based on estimated item parameters, divided by its standard deviation asymptotically follows standard normal distribution, while the test statistic based on estimated UA for item i has asymptotically half-normal distribution (Raju, 1990).

Non-IRT methods for DIF detection

Non-IRT approaches have been studied and used in the DIF detection for decades. These are typically straightforward methods which are easy to explain to audiences and easy to apply in standard statistical software, or even by hand. Non-IRT techniques include for example Angoff's delta plot (Angoff & Ford, 1973, see also Magis & Facon, 2012), well known Mantel-Haenszel test (Mantel & Haenszel, 1959, see also P. W. Holland, 1985 and P. W. Holland & Thayer, 1988), the Simultaneous Item Bias Test (SIBTEST) method (Shealy & Stout, 1993), or the logistic regression model for DIF detection (Swaminathan & Rogers, 1990).

Angoff's delta plot. Angoff's delta plot (Angoff & Ford, 1973) compares nonlinear transformations of the empirical probabilities (also called the *delta scores*) per item in the two groups. The delta scores are plotted for each item for the two groups in a scatter-plot called *diagonal plot* or *delta plot*. An item is under suspicion of DIF if the delta point considerably departs from the main axis of ellipsoid created by delta scores. The detection threshold is either fixed or based on a bivariate normal approximation (Magis & Facon, 2012).

The Mantel-Haenszel test. The Mantel-Haenszel test (Mantel & Haenszel, 1959) is one of the most popular and used methods in the DIF framework (P. W. Holland & Thayer, 1988) despite the fact that it is only able to detect a uniform DIF (see, e.g., Swaminathan & Rogers, 1990). It tests an association between the item responses and the group membership variable conditionally on the level of the total test score. In more detail, assuming I dichotomously scored items of the test, for each level of the total test score $k = 0, \dots, I$ for given item i, a 2 x 2 contingency table is produced (Table 1).

Table 1: Contingency table for item i and for the total test score of k.

	$Y_i = 1$	$Y_i = 0$
Reference group (0)	n_{i01k}	n_{i00k}
Focal group (1)	n_{i11k}	n_{i10k}

These contingency tables summarize responses on item i by the respondents with the total test score equal to k. Terms n_{i01k} and n_{i11k} correspond to a number of respondents from the reference and the focal group who answered correctly, while terms n_{i00k} and n_{i10k} indicate the numbers of those from the reference and the focal group who answered incorrectly.

The Mantel-Haenszel test statistic then combines all levels of the total score $k = 0, \dots, I$ for given item i and takes the following form:

$$\mathrm{MH}_i = \frac{\left[\left|\sum_{k=0}^{I}\left(n_{i01k} - \frac{(n_{i01k}+n_{i00k})(n_{i01k}+n_{i11k})}{n_{ik}}\right)\right| - 0.5\right]^2}{\sum_{k=0}^{I} \frac{(n_{i01k}+n_{i00k})(n_{i11k}+n_{i10k})(n_{i01k}+n_{i11k})(n_{i00k}+n_{i10k})}{n_{ik}^2(n_{ik}-1)}}. \tag{4}$$

where $n_{ik} = n_{i01k}+n_{i00k}+n_{i11k}+n_{i10k}$. Under the null hypothesis of no conditional association between the item responses and group membership given the level of the total score, i.e., no DIF, the MH_i statistic (4) has an asymptotic chi-square distribution with one degree of freedom for the large sample sizes (see, e.g., Agresti, 2010, p. 232, and P. W. Holland & Thayer, 1988).

SIBTEST. In the case that the target ability distributions are the same for both the groups, Shealy and Stout (1993) proposed a test statistic to identify DIF in item i in the following form

$$\Omega_i = \frac{\hat{\omega}_i}{\hat{\sigma}(\hat{\omega}_i)}, \tag{5}$$

where $\hat{\omega}_i$ is given by

$$\hat{\omega}_i = \sum_{k=0}^{I} p_k \left(\bar{Y}_{i0k} - \bar{Y}_{i1k}\right). \tag{6}$$

The term p_k is a proportion of respondents from the focal group who gained total test score equal to $k = 0, \dots, I$, while $\bar{Y}_{i0k}$ and $\bar{Y}_{i1k}$ are the mean item responses of the respondents with the total test score k from the reference and the focal group, respectively. The term $\hat{\sigma}(\hat{\omega}_i)$ is the estimated standard error of the estimate $\hat{\omega}_i$ (for equations see Shealy & Stout, 1993). Under the null hypothesis $\omega_i = 0$, i.e., no DIF in item i, the test statistic Ω_i has an asymptotic standard normal distribution for large sample sizes.

However, the assumption of equal ability distributions across groups is an unrealistic condition for most applications. Thus, terms $\bar{Y}_{i0k}$ and $\bar{Y}_{i1k}$ in the (6) are replaced by their regression-based estimates $\bar{Y}^*_{i0k}$ and $\bar{Y}^*_{i1k}$ (Shealy & Stout, 1993).

Logistic regression method. Finally, the natural way how to model probability of the correct answer and how to test for group-based differences in item responses is a method based on logistic regression (Swaminathan & Rogers, 1990). This method fits a logistic model for the probability of answering the tested item correctly with the observed ability and group membership variable as regressors. It also includes their mutual interaction:

$$\mathsf{P}(Y_{pi} = 1|X_p, G_p) = \frac{e^{\beta_{i0}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}{1 + e^{\beta_{i0}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}, \tag{7}$$

where X_p is an observed ability of person p and G_p is a variable describing respondents group membership ($G_p = 0$ for the reference group and $G_p = 1$ for focal group, usually seen as the disadvantaged one). Item parameters β_{i0}, β_{i1}, β_{i2}, and β_{i3} are intercept, slope, group membership effect, and interaction between the observed ability and group variable respectively.

The logistic regression model (7) can be used to detect a uniform DIF by testing the effect of group membership, i.e., $\beta_{i2} = 0$ vs. $\beta_{i2} \neq 0$ while constraining $\beta_{i3} = 0$. A nonuniform DIF effect can be detected by testing the effect of the interaction of the observed ability X_p and the group membership variable G_p, i.e., $\beta_{i3} = 0$ vs. $\beta_{i3} \neq 0$. Finally, it is possible to detect any DIF by testing for nonzero effects of both parameters β_{i2} and β_{i3} connected with group membership simultaneously. Item parameters $\beta_{i0}, \beta_{i1}, \beta_{i2}$, and β_{i3} are typically estimated using iteratively re-weighted least squares. Evaluation of significance of their effects and thus DIF detection can be performed using classical statistical test procedures such as likelihood ratio test or Wald's test.

Structure of the thesis and the main results

The main topic of this thesis is detection of DIF and DDF among binary, ordinal, and nominal items. We first deal with extensions of the logistic regression method (Swaminathan & Rogers, 1990) for binary items with possible guessing and/or inattention in Chapter 1. We focus on specification of the newly proposed statistical models, their interpretation, estimation, and implementation within the free statistical software **R** (R Core Team, 2020). We propose innovations to estimation methods and compare existing and newly proposed methods in simulation studies. In Chapter 2, the same concepts are extended for polytomous items. Chapter 3 proposes a new detection method based on nonparametric approach. Finally, Chapter 4 is devoted to further topics in DIF and DDF detection, namely the *item purification*, *corrections for multiple comparisons*, and *DIF effect sizes*. The contribution of each chapter is in more detail presented below.

Chapter 1 describes generalized logistic regression (also called nonlinear) models for DIF detection among binary items. This class of models allows for possibility of guessing and/or inattention when responding and also for detection of between-group differences in these item characteristics. Section 1.1 comprises model specification and discusses different parameterizations of the model together with their interpretation and mutual relationship. Section 1.2 reviews estimation techniques including the nonlinear least squares, the maximum likelihood method, and the Expectation-Maximization (EM) algorithm. It also offers a newly proposed method based on the parametric link function as will be summarized in Hladká, Brabec, and Martinková (2021). Section 1.4 describes the implementation of the generalized logistic models among dichotomous items into the **R** software and its package **difNLR** (Hladká & Martinková, 2020). Finally, Sections 1.5 and 1.6 present two simulation studies: The first simulation study evaluates the properties of the DIF detection method based on the nonlinear models (Drabinová & Martinková, 2017) and the second one compares various estimation techniques with the focus on group-specific models discussed in Section 1.2.

Chapter 2 reviews the group-specific generalized logistic regression models for DIF and DDF detection among polytomous items. Section 2.1 offers two models for DIF detection for ordinal items, namely the cumulative logit model and the adjacent category logit model. Section 2.2 describes nominal model for DDF detection. Besides the model specification, Chapter 2 comprises techniques to estimate item parameters and also presents implementation into the **difNLR** **R** package using simulated data examples (Hladká & Martinková, 2020).

Chapter 3 describes nonparametric comparison of ICCs for the DIF detection as will be summarized in Hladká and Martinková (2021). This new approach builds on work by Srihera and Stute (2010) which focuses on general comparison of regression functions. Section 3.1 deals with two challenges: kernel smoothing estimation of characteristic curves and fine-tuning of a test statistic allowing their comparison and thus DIF detection. While asymptotic normality was proven by Srihera and Stute (2010) for the test statistic using optimal weights maximizing power of the test proposed by these authors, the optimal weights are available only in case of known item characteristic functions. In this thesis, an estimate of optimal weights is proposed and resulting test statistic is evaluated in the context of DIF detection problem. Section 3.2 offers a simulation study evaluating properties of the proposed DIF detection procedure.

Chapter 4 discusses further issues in the DIF detection. Typically, DIF detection is performed item by item, which may cause two issues. First, potentially unfair items are included in calculation of the matching criterion which may be thus biased. Second, the type I error rate may be increased due to multiple testing. Section 4.1.1 discusses so called *item purification* dealing with the first issue and Section 4.1.2 studies correction methods for *multiple comparison corrections* in DIF detection which can tackle the second issue. Section 4.2 summarizes complex simulation study by Hladká, Martinková, and Magis (2021) which jointly evaluates properties of both approaches and moreover offers and studies their mixtures in various settings of DIF detection.

Publications by the author related to the thesis

The thesis comprises, summarizes, and partially extends the following papers:

Drabinová, A., & Martinková, P. (2017) Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498–517, doi: 10.1111/jedm12158

The paper introduced nonlinear regression for DIF detection. The model proposed there is a restricted version of the four parameter nonlinear regression model discussed in this thesis. It allows for possibility of guessing the correct answer without necessary knowledge, possibly varying between the two groups. The paper also included simulation study evaluating the properties of the proposed model and comparing it to other commonly used methods for DIF detection, which is presented in this thesis in its original form. It also offered practical illustrations on real data from admission test in Biology to medical faculty in the Czech Republic. An early version of this paper received Travel Award and was presented at International Meeting of the Psychometric Society (IMPS, 2016, Asheville, NC, USA).

Hladká, A., & Martinková, P. (2020) difNLR: Generalized logistic regression model for DIF and DDF detection. *The R Journal, 12*(1), 300–323. doi: 10.32614/RJ-2020-014

The second paper described `R` package **`difNLR`** which offers implementation of generalized logistic models for DIF and DDF detection discussed in this thesis including four parameter nonlinear regression model, which allows also for possibility of inattention when answering. The paper offered practical guide to fit models in `R`, from data generation to visualisation of the results. It also included real data example which illustrated complex use of the proposed models.

Hladká, A., Brabec, M., & Martinková, P. (2021) Estimation in generalized logistic regression model. In preparation for submission.

This paper will built on Section 1.2 and will summarize several approaches to estimate parameters in the four parameter nonlinear regression model including the newly proposed two-step estimation procedure based on a generalized logistic model using parametric link function. Estimation methods will be compared in a simulation study.

Hladká, A., & Martinková, P. (2021) Nonparametric comparison of regression curves for DIF detection. In preparation for submission.

The fourth paper will build on Chapter 3 offering nonparametric approach based on comparison of regression curves for DIF detection. Several choices of weight functions will be discussed and an estimate of optimal weights together with the use of the wild bootstrap will be proposed.

Hladká, A., Martinková, P., & Magis, D. (2021) Issues and practice in detection of differential item functioning: Applying item purification, correction for multiple comparisons, or combination of both? In preparation for re-submission.

The last core paper builds on Chapter 4 and offers extensive simulation study to examine various approaches to deal with the maintaining the level of type I error and to improve DIF detection under diverse conditions. Specifically, performance of item purification and corrections for multiple comparison are jointly evaluated and their combinations is also studied.

Topics covered in this thesis further relate to the following works:

Martinková, P., & Drabinová, A. (2018) ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503–515, doi: 10.32614/RJ-2018-074

This paper describes an **`R`** package and an online application **`ShinyItemAnalysis`** which offers methods for complex psychometric analysis of the educational and psychological tests. Software provides wide range of psychometric methods for testing reliability, validity and detailed analysis of functioning of single test items. It also makes available models for DIF detection included and described in this thesis.

Martinková, P., Hladká, A., & Potužníková, E. (2020) Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction, 66*, 101286,
doi: 10.1016/j.learninstruc.2019101286

The paper includes detailed psychometric analysis of test of learning competences and its items with the focus on differences in gains between the basic school track and the selective academic track. The concept of DIF is extended to longitudinal setting and DIF analysis is combined with the propensity score matching technique. DIF analysis uses generalized logistic regression model described in this thesis.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017) Checking equity: Why DIF analysis should be a routine part of developing conceptual assessments. *CBE–Life Sciences Education 16*(2), rm2, doi: 10.1187/cbe.16-10-0307

This methodological paper presents number of DIF detection approaches. It also provides a simulated as well as a real data example showing importance of DIF analysis and its added value over traditional methods for detection of between-group differences in development and validation of educational tests.

Martinková, P., & Hladká, A. (2021) *Computational aspects of psychometric methods in education, psychology, and health: With examples in R*. CRC Press. (In preparation)

This book will cover key psychometric topics, from Classical Test Theory (CTT) to IRT including DIF detection, with focus on computational aspects of statistical methods. Book will also include examples of implementation in the statistical software **`R`** on real data from educational, psychological, and health-related measurements.

Notation

The mathematical and special symbols together with abbreviations used throughout the thesis are summarized in Glossaries on page 140. Note that index i is related to an item of the interest while index p to the person or respondent. Within the thesis, *italic* is used to emphasize new or important concepts.

1. Generalized logistic regression models for binary items

In this chapter we review the *generalized logistic regression model* (also called the *nonlinear model*) for DIF detection among dichotomous items as was proposed and published in papers by Drabinová and Martinková (2017) and Hladká and Martinková (2020).

In Section 1.1 we introduce a class of nonlinear models which account for possibility of guessing and/or inattention when responding. Then, in Section 1.2, we focus on estimation procedures to estimate item parameters in the nonlinear models. We review the method of nonlinear least squares and the maximum likelihood method to estimate item parameters, both discussed in Drabinová and Martinková (2017) and Hladká and Martinková (2020). Newly, we deal with these methods in more detail including asymptotic properties of the item parameter estimates and estimates of their asymptotic variance. We further present two new estimation approaches for the nonlinear models for DIF detection – the EM algorithm and a newly proposed algorithm based on parametric link function.

Section 1.3 is devoted to DIF detection when using the nonlinear models. Further, in Section 1.4 we show implementation of the methods in the freely available statistical software **R** (R Core Team, 2020) and its package **difNLR** (Hladká & Martinková, 2020). Finally, in Sections 1.5 and 1.6, we offer two simulation studies. The first one explores properties of the proposed DIF detection procedure based on the nonlinear model as was published in Drabinová and Martinková (2017). The second simulation study newly compares estimation techniques discussed in Section 1.2.

1.1 Model specification

The nonlinear regression models for DIF detection among binary items are extensions of the logistic regression method (7) proposed by Swaminathan and Rogers (1990). These extensions may account for the possibility that an item can be correctly answered without possessing the necessary knowledge, i.e., a lower asymptote of the ICC can be larger than zero. Similarly, these models can take into consideration the possibility that an item is incorrectly answered by a person with high ability due to, for example, inattention or lack of time, i.e., an upper asymptote of the ICC can be lower than one.

The probability of a correct answer on item i by person p is then given by

$$\mathsf{P}(Y_{pi} = 1 | X_p, G_p) = c_{iG_p} + (d_{iG_p} - c_{iG_p}) \frac{e^{a_{iG_p}(X_p - b_{iG_p})}}{1 + e^{a_{iG_p}(X_p - b_{iG_p})}}, \tag{1.1}$$

where X_p is a *matching criterion*, a variable describing knowledge or ability of person p, and G_p stands for respondent's membership to a social group ($G_p = 0$ for the reference group and $G_p = 1$ for the focal group). Parameters $a_{iG_p}, b_{iG_p} \in \mathbb{R}$, and $c_{iG_p}, d_{iG_p} \in [0, 1]$ represent discrimination, difficulty, probability of guessing, and a parameter related to the probability of inattention in item i, while they can differ for the reference and the focal group. Thus, there are eight parameters for each item in total.

The parametrization of the terms a_{iG_p}, b_{iG_p}, c_{iG_p}, and d_{iG_p} in model (1.1) can take the form of sum of baseline parameters and a difference in these parameters between the two groups, that is, for example, $a_{iG_p} = a_i + a_{i\mathrm{DIF}} G_p$. In other words, $a_{i0} = a_i$ for the reference group and $a_{i1} = a_i + a_{i\mathrm{DIF}}$ for the focal group. The second possibility is to define the terms a_{iG_p}, b_{iG_p}, c_{iG_p}, and d_{iG_p} as the sums of two mutually exclusive group-based parameters, e.g., $a_{iG_p} = a_{i0}\,(1 - G_p) + a_{i1} G_p$. Then parameters a_i and a_{i0} are the same and they refer to the item properties of the reference group, while parameter $a_{i\mathrm{DIF}}$ gives a difference in related parameters between the focal and the reference group, i.e., $a_{i\mathrm{DIF}} = a_{i1} - a_{i0}$. In what follows, we stick with the first option which describes the group-based differences in the parameters and we set $\boldsymbol{\gamma}_i = \{a_i, a_{i\mathrm{DIF}}, b_i, b_{i\mathrm{DIF}}, c_i, c_{i\mathrm{DIF}}, d_i, d_{i\mathrm{DIF}}\}$ to be a set of parameters of the model (1.1) for item i.

1.1.1 Interpretation of the parameters

Guessing parameter c_i (sometimes also called *pseudo-guessing parameter*) can be interpreted as a probability of guessing correct answer of the item i without necessary knowledge or ability when respondent comes from the reference group. In other words, it is the probability of correct answer when the ability level X_p goes to $-\infty$, i.e.,

$$c_i = \lim_{x \to -\infty} \mathsf{P}(Y_{pi} = 1 | X_p = x, G_p = 0),$$

and describes thus a left asymptote of the ICC (in case that $a_i > 0$ the lower one, see Figure 1.1). While the parameter c_i may take values from the whole interval $[0, 1]$, in multiple-choice answers it would typically be around 1 divided by a number of choices in given item, i.e., for 4 choices it would be around 0.25, but it may also depend on attractiveness of the distractors.

Inattention parameter d_i is related to the probability of answering the item i incorrectly by a respondent from the reference group even in case when they have the full knowledge of the construct being measured, i.e., when the ability level goes to ∞:

$$d_i = \lim_{x \to \infty} \mathsf{P}(Y_{pi} = 1 | X_p = x, G_p = 0),$$

and gives thus a right asymptote of the ICC (in case that $a_i > 0$ the upper one, see Figure 1.1). Term $1 - d_i$ then can be interpreted as the probability of inattention for the respondent from the reference group. Unlike the case of parameter c_i, it is not entirely obvious what the typical values of parameter d_i should be. The parameter d_i is rather related to the situations when respondent did not have enough time, was inattentive, or did not want to admit condition, such as crying in health-related or attitude measurements, because they consider it socially or culturally unacceptable.

Difficulty parameter b_i corresponds to a value of the matching criterion X_p which is necessary to answer item i correctly with a probability of $\frac{d_i + c_i}{2}$, that is a midpoint between the two asymptotes of the ICC for the reference group (see

Figure 1.1):

$$\begin{aligned}\mathsf{P}(Y_{pi}=1|X_p=b_i,G_p=0)&=c_i+(d_i-c_i)\frac{e^{a_i(b_i-b_i)}}{1+e^{a_i(b_i-b_i)}}\\&=c_i+(d_i-c_i)\frac{e^0}{1+e^0}\\&=\frac{d_i+c_i}{2}.\end{aligned}$$

For example, in case of the logistic regression model (7), i.e., setting $c_i=0$ and $d_i=1$ in (1.1), parameter b_i corresponds to the ability level at which respondents from the reference group have the probability of 0.5 to answer given item correctly.

Finally, the *discrimination* parameter a_i is related to the slope of the ICC (1.1) in inflection point b_i (see Figure 1.1). This can be seen when looking at the first derivative with respect to a variable related to the matching criterion:

$$\begin{aligned}\mathsf{P}'(Y_{pi}=1|X_p,G_p=0)&=\frac{\partial\,\mathsf{P}(Y_{pi}=1|X_p,G_p)}{\partial X_p}\\&=(d_i-c_i)\frac{e^{a_i(X_p-b_i)}}{\left(1+e^{a_i(X_p-b_i)}\right)^2},\end{aligned}$$

and for the inflection point $X_p=b_i$ we get

$$\mathsf{P}'(Y_{pi}=1|X_p=b_i,G_p=0)=(d_i-c_i)\frac{a_i}{4}.$$

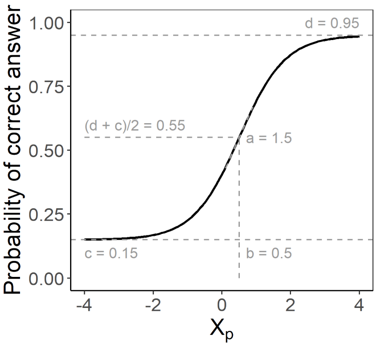


Figure 1.1: Interpretation of the parameters of the nonlinear model.

As mentioned above, parameters $a_{i\mathrm{DIF}}, b_{i\mathrm{DIF}}, c_{i\mathrm{DIF}}$, and $d_{i\mathrm{DIF}}$ describe differences between the focal and the reference group in the discrimination, difficulty, guessing, and inattention, respectively. Parameters $c_{i\mathrm{DIF}}$ and $d_{i\mathrm{DIF}}$ are related to the probabilities of guessing and inattention and need to be bounded by 0 and 1, thus, it is necessary to set $c_{i\mathrm{DIF}}\in[-c_i,1-c_i]$ and $d_{i\mathrm{DIF}}\in[-d_i,1-d_i]$ while $c_i\in[0,1]$ and $d_i\in[0,1]$.

1.1.2 Alternative parametrization

Current parametrization is convenient as it describes psychometric properties of the items. Moreover, it is similar to the one used in IRT models and makes thus parameter estimates directly comparable. We further reference it as the IRT parametrization. However, for computational purposes and to allow for multiple regressors, it is sometimes more suitable to use the classical intercept-slope parametrization as for the logistic regression method (7). The nonlinear model (1.1) then takes the following form:

$$\mathsf{P}(Y_{pi} = 1|X_p, G_p) = c_{iG_p} + (d_{iG_p} - c_{iG_p})\frac{e^{\beta_{i0}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}{1 + e^{\beta_{i0}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}. \qquad (1.2)$$

Parameters c_{iG_p} and d_{iG_p} can be defined analogously as for the IRT parametrization described above and their interpretation remains the same. Interpretation of parameters β_{i0}, β_{i1}, β_{i2}, and β_{i3} is the same as for the logistic regression model (7): Intercept β_{i0} is related to the baseline probability of answering item i correctly, that is a probability when $X_p = 0$; parameter β_{i1} is related to the change in the odds ratio of the correct answer in item i when X_p increases by one point; β_{i2} indicates a difference between the reference and the focal group in the intercept, and β_{i3} indicates a difference between the two groups in the slope.

Mutual relationship between the parameters of (1.1) and the parameters of (1.2) is then given as follows:

$$\beta_{i0} = -a_i b_i,\ \beta_{i1} = a_i,\ \beta_{i2} = -a_i b_{i\mathrm{DIF}} - a_{i\mathrm{DIF}} b_i - a_{i\mathrm{DIF}} b_{i\mathrm{DIF}},\ \beta_{i3} = a_{i\mathrm{DIF}}.$$

Standard errors of the estimates can be obtained using a delta method which allows to obtain an approximate distribution for a function of an asymptotically normal estimator (see, for example, Doob, 1935) whatever parametrization is used. In what follows we use the IRT parametrization.

1.1.3 Matching criterion

We use a term the *matching criterion* for the observed ability X_p used in analysis of DIF. Using the *total test score* X_p as the matching criterion may lead to contradictions for the nonlinear model (1.1), especially when respondents gained zero points or maximum score of the test. In such cases, the probability of correct answer on each item should be 0 and 1 respectively, but the model will predict otherwise:

$$\begin{aligned}\mathsf{P}(Y_{pi} = 1|X_p = 0, G_p = 0) &= c_i + (d_i - c_i)\frac{1}{1 + e^{a_i b_i}} > c_i \geq 0,\\ \mathsf{P}(Y_{pi} = 1|X_p = I, G_p = 0) &= c_i + (d_i - c_i)\frac{e^{a_i(I-b_i)}}{1 + e^{a_i(I-b_i)}} < d_i \leq 1.\end{aligned}$$

Similarly for the focal group we get $\mathsf{P}(Y_{pi} = 1|X_p = 0, G_p = 1) > c_i + c_{i\mathrm{DIF}} \geq 0$ and $\mathsf{P}(Y_{pi} = 1|X_p = I, G_p = 1) < d_i + d_{i\mathrm{DIF}} \leq 1$. This is also an issue for the logistic regression model (7).

In what follows, we further often use the *standardized total test score*, also called *Z-score*, as an estimate of the respondents' ability, that is

$$X_p = \frac{\sum_{i=1}^{I} Y_{pi} - \bar{Y}}{\sqrt{\frac{1}{p-1}\sum_{i=1}^{I}\left(Y_{pi} - \bar{Y}\right)^2}},$$

where $\bar{Y} = \frac{1}{p}\sum_{p=1}^{n}\sum_{i=1}^{I} Y_{pi}$. However, even though not directly apparent, even this choice of the matching criterion is not fully appropriate and the problem persists for the lowest and the highest values of the Z-score.

One possibility how to deal with this issue, is to apply a continuous transformation of the total test score to the real numbers $\mathbb{R}$. For example, the average item score is calculated for each respondent and then a logit function is applied:

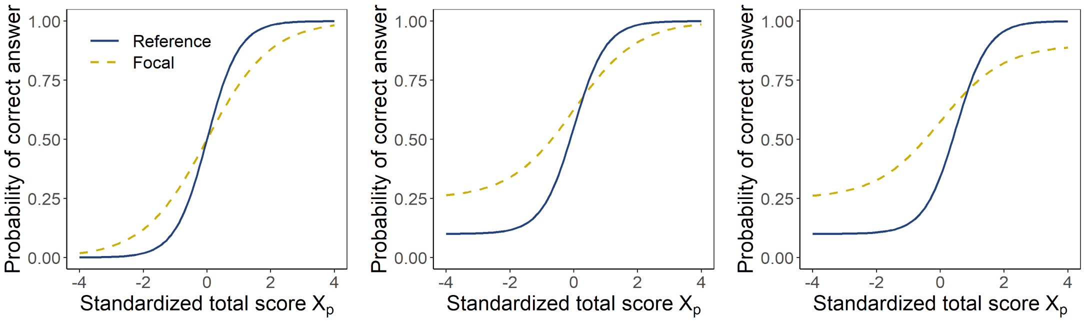
$$X_p = \log\left(\frac{\frac{1}{I}\sum_{i=1}^{I} Y_{pi}}{1 - \frac{1}{I}\sum_{i=1}^{I} Y_{pi}}\right).$$

Thus $X_p = -\infty$ if and only if no item was correctly answered and $X_p = \infty$ if and only if all items were correctly answered.

Another possibility is to use a jackknife estimate of the total test score or its standardization. The matching criterion is then calculated based on all items except the one which is currently examined. In other words, the matching criterion for the given item i is computed as follows:

$$X_{pi} = \sum_{j=1,\ j\neq i}^{I} Y_{pj}.$$

The class of models determined by equation (1.1) contains a wide range of scenarios for DIF detection, for more details see Section 1.3. For instance, with $c_i = c_{i\mathrm{DIF}} = d_{i\mathrm{DIF}} = 0$ and $d_{iG_p} = 1$ one can obtain the classical logistic regression model (7) for the detection of uniform and non-uniform DIF (Swaminathan & Rogers, 1990, see Figure 1.2A). Assuming $d_i = 1$ and $d_{i\mathrm{DIF}} = 0$, we get a nonlinear model for the DIF detection allowing for differential guessing in the groups (Drabinová & Martinková, 2017, see Figure 1.2B). However, the nonlinear model (1.1) can also be used to detect differences between the two groups in any of the four parameters a_i, b_i, c_i, and d_i as shown in Figure 1.2C.



(A) 2PL model.

(B) 3PL model.

(C) 4PL model.

Figure 1.2: Examples of the nonlinear regression models.

In contrast to the 4PL IRT model (1), the nonlinear model (1.1) assumes that the underlying latent trait is estimated by the test-score based matching criterion, typically standardized total test score or its suitable transformation, and thus the described method belongs to a class of non-IRT approaches. As such, the nonlinear model (1.1) can be seen as a proxy to the 4PL IRT model (1) for DIF detection. While estimation of the asymptote parameters is notoriously challenging in the IRT models, it was shown that the generalized logistic models require a smaller sample size to be fitted while they keep pleasant properties in terms of power and rejection rates (Drabinová & Martinková, 2017, see also Section 1.5).

1.2 Estimation of parameters

In this part we provide a review of approaches to estimate item parameters $\boldsymbol{\gamma}_i = \{a_i, a_{i\mathrm{DIF}}, b_i, b_{i\mathrm{DIF}}, c_i, c_{i\mathrm{DIF}}, d_i, d_{i\mathrm{DIF}}\}$ in the model (1.1) including the nonlinear least squares and the maximum likelihood method, as described in Drabinová and Martinková (2017) and Hladká and Martinková (2020). However, the methods are presented here in more detail, notably we also study asymptotic properties of the parameter estimates. We further propose using the EM algorithm to estimate item parameters and we newly propose an algorithm based on parametric link function as will be summarized in Hladká, Brabec, and Martinková (2021). The properties of the estimation techniques are compared in a simulation study described in Section 1.6.

1.2.1 Nonlinear least squares

The first method described here is the *nonlinear least squares* (see, for example, Ritz & Streibig, 2008; Seber & Wild, 1989; van der Vaart, 2000), that is, a minimization of the Residual Sum of Squares (RSS) of item i with respect to item parameters $\boldsymbol{\gamma}_i$. By setting

$$\pi_{pi} = \mathsf{P}(Y_{pi} = 1 | X_p, G_p) \tag{1.3}$$

in the model (1.1), we get the RSS of the item i in the following form:

$$\mathrm{RSS}_i(\boldsymbol{\gamma}_i) = \sum_{p=1}^{n} \left(Y_{pi} - \pi_{pi}\right)^2, \tag{1.4}$$

where n denotes the number of respondents, Y_{pi} is the response of respondent p to the item i, X_p is their matching criterion, and G_p is their group membership variable. The nonlinear least squares estimator is then given by

$$\hat{\boldsymbol{\gamma}}_i = \left\{\widehat{a}_i, \widehat{a}_{i\mathrm{DIF}}, \widehat{b}_i, \widehat{b}_{i\mathrm{DIF}}, \widehat{c}_i, \widehat{c}_{i\mathrm{DIF}}, \widehat{d}_i, \widehat{d}_{i\mathrm{DIF}}\right\} = \arg\min_{\boldsymbol{\gamma}_i} \mathrm{RSS}_i(\boldsymbol{\gamma}_i)$$

and thus can be then seen as an *M-estimator.* As the *criterion function* $\mathrm{RSS}_i(\boldsymbol{\gamma}_i)$ (1.4) is continuously differentiable with respect to parameters $\boldsymbol{\gamma}_i$, the minimizer can be obtained when gradient is zero, i.e., $\nabla_{\boldsymbol{\gamma}_i}\mathrm{RSS}_i(\boldsymbol{\gamma}_i) = 0$. So the minimization process involves a calculation of the first partial derivatives with respect to parameters $\boldsymbol{\gamma}_i$ and finding a solution of relevant nonlinear *estimating equations*:

$$\nabla_{\boldsymbol{\gamma}_i}\mathrm{RSS}_i(\boldsymbol{\gamma}_i) = \sum_{p=1}^{n} \frac{\partial \left(Y_{pi} - \pi_{pi}\right)^2}{\partial \boldsymbol{\gamma}_i} = \sum_{p=1}^{n} \boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i) = 0,$$

where

$$\begin{aligned}\boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i) =&\ (\psi_{i1}(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i), \ldots, \psi_{i8}(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i))\\ =&\ \left(-2(Y_{pi} - \pi_{pi})\frac{\partial \pi_{pi}}{\partial \gamma_{i1}}, \ldots, -2(Y_{pi} - \pi_{pi})\frac{\partial \pi_{pi}}{\partial \gamma_{i8}}\right),\end{aligned}$$

see also van der Vaart (2000, Example 5.27).

Let's first substitute

$$\phi_{pi} = \frac{e^{(a_i + a_{i\mathrm{DIF}} G_p)(X_p - b_i - b_{i\mathrm{DIF}} G_p)}}{1 + e^{(a_i + a_{i\mathrm{DIF}} G_p)(X_p - b_i - b_{i\mathrm{DIF}} G_p)}}.$$

The term ϕ_{pi} describes a logistic regression curve, that is a probability of correct answer when respondents were not guessing and were not inattentive. The partial derivatives of the $\mathrm{RSS}_i(\boldsymbol{\gamma}_i)$ with respect to the k-th parameter, where $k = 1, \ldots, 8$, are then

$$\frac{\partial \mathrm{RSS}_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}} = \sum_{p=1}^{n} \psi_{ik}(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i) = -2\sum_{p=1}^{n} (Y_{pi} - \pi_{pi}) \frac{\partial \pi_{pi}}{\partial \gamma_{ik}}, \tag{1.5}$$

where

$$\frac{\partial \pi_{pi}}{\partial a_i} = (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p)\frac{\partial \phi_{pi}}{\partial a_i}, \tag{1.6}$$

$$\frac{\partial \pi_{pi}}{\partial a_{i\mathrm{DIF}}} = (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p)\frac{\partial \phi_{pi}}{\partial a_{i\mathrm{DIF}}}, \tag{1.7}$$

$$\frac{\partial \pi_{pi}}{\partial b_i} = (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p)\frac{\partial \phi_{pi}}{\partial b_i}, \tag{1.8}$$

$$\frac{\partial \pi_{pi}}{\partial b_{i\mathrm{DIF}}} = (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p)\frac{\partial \phi_{pi}}{\partial b_{i\mathrm{DIF}}}, \tag{1.9}$$

$$\frac{\partial \pi_{pi}}{\partial c_i} = 1 - \phi_{pi}, \tag{1.10}$$

$$\frac{\partial \pi_{pi}}{\partial c_{i\mathrm{DIF}}} = (1 - \phi_{pi})\, G_p, \tag{1.11}$$

$$\frac{\partial \pi_{pi}}{\partial d_i} = \phi_{pi}, \tag{1.12}$$

$$\frac{\partial \pi_{pi}}{\partial d_{i\mathrm{DIF}}} = \phi_{pi} G_p, \tag{1.13}$$

with

$$\begin{aligned}\frac{\partial \phi_{pi}}{\partial a_i} &= \phi_{pi}\,(1 - \phi_{pi})\,(X_p - b_i - b_{i\mathrm{DIF}} G_p)\,,\\ \frac{\partial \phi_{pi}}{\partial a_{i\mathrm{DIF}}} &= \phi_{pi}\,(1 - \phi_{pi})\,(X_p - b_i - b_{i\mathrm{DIF}} G_p)\, G_p,\\ \frac{\partial \phi_{pi}}{\partial b_i} &= -\,\phi_{pi}\,(1 - \phi_{pi})\,(a_i + a_{i\mathrm{DIF}} G_p)\,,\\ \frac{\partial \phi_{pi}}{\partial b_{i\mathrm{DIF}}} &= -\,\phi_{pi}\,(1 - \phi_{pi})\,(a_i + a_{i\mathrm{DIF}} G_p)\, G_p.\end{aligned} \tag{1.14}$$

In our case, the nonlinear least squares minimization problem includes a system of the eight nonlinear estimating equations with the eight unknown parameters: $\frac{\partial \mathrm{RSS}_i}{\partial \gamma_{ik}} = 0$, $k = 1, \ldots, 8$, given by (1.5).

Asymptotic properties. Asymptotic properties, such as consistency and asymptotic distribution, of the nonlinear least squares estimator can be shown under the classical set of regularity conditions (see, for example, van der Vaart, 2000, Theorems 5.41 and 5.42). We reformulate these conditions for our situation:

[R0] A vector of true parameters $\boldsymbol{\gamma}_{iX}$ satisfies

$$\mathsf{E}\left(\boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_{iX})\right) = \mathsf{E}\left(-2(Y_{pi} - \pi_{pi})\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_{iX}}\right) = \mathbf{0}.$$

[R1] The true parameter $\boldsymbol{\gamma}_{iX}$ is an interior point of the parameter space.

[R2] The function $\boldsymbol{\psi}_i(y, x, g; \boldsymbol{\gamma}_i)$ is twice continuously differentiable with respect to $\boldsymbol{\gamma}_i$ for every (y, x, g).

[R3] For each $\boldsymbol{\gamma}_i^*$ in a neighborhood of $\boldsymbol{\gamma}_{iX}$ there exists an integrable function $\ddot{\boldsymbol{\psi}}(y, x, g)$ such that

$$\left|\frac{\partial^2 \psi_{ik}(y, x, g; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij} \partial \gamma_{il}}\right| \leq \ddot{\boldsymbol{\psi}}(y, x, g),$$

for each $k, j, l = 1, \ldots, 8$.

[R4] The matrix

$$\begin{aligned}\mathbb{\Gamma}_i(\boldsymbol{\gamma}_i) = {} & \mathsf{E}\left(\frac{\partial \boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i^\top}\right) \\ = {} & 2\,\mathsf{E}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\right)^\top - (Y_{pi} - \pi_{pi})\frac{\partial^2 \pi_{pi}}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right) \\ = {} & 2\,\mathsf{E}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\right)^\top\right)\end{aligned}$$

is finite and regular in a neighbourhood of $\boldsymbol{\gamma}_{iX}$.

[R5] The variance matrix

$$\begin{aligned}\mathbb{\Sigma}_i(\boldsymbol{\gamma}_i) = {} & \mathsf{E}\left(\boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i)\boldsymbol{\psi}_i^\top(Y_{pi}, X_p, G_p; \boldsymbol{\gamma}_i)\right) \\ = {} & 4\,\mathsf{E}\left((Y_{pi} - \pi_{pi})^2\,\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\right)^\top\right) \\ = {} & 4\,\mathsf{E}\left(\pi_{pi}\,(1 - \pi_{pi})\,\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi_{pi}}{\partial \boldsymbol{\gamma}_i}\right)^\top\right)\end{aligned}$$

is finite for $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_{iX}$.

Specifically, the Theorem 5.42 (van der Vaart, 2000, p. 68) implies that under the conditions [R0]–[R5], the probability that the estimating equations $\frac{\partial \mathrm{RSS}_i}{\partial \gamma_{ik}} = 0$, $k = 1, \ldots, 8$ have at least one root tending to 1, as $n \to \infty$, and there exists a sequence $\hat{\boldsymbol{\gamma}}_i$ (depending on n) such that $\hat{\boldsymbol{\gamma}}_i \xrightarrow[n\to\infty]{P} \boldsymbol{\gamma}_{iX}$, and, moreover, the sequence

$\hat{\boldsymbol{\gamma}}_i$ can be chosen to be a local maximum for each n. Theorem 5.41 (van der Vaart, 2000, p. 68) shows that every consistent estimator $\hat{\boldsymbol{\gamma}}_i$ has asymptotically normal distribution, that is:

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_{iX}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbb{\Gamma}_i^{-1}(\boldsymbol{\gamma}_{iX})\boldsymbol{\Sigma}_i(\boldsymbol{\gamma}_{iX})\mathbb{\Gamma}_i^{-1}(\boldsymbol{\gamma}_{iX})\right),$$

It is easy to see that the conditions [R0] and [R2] hold. To satisfy the condition [R1], we need to bound parameters c_i, $c_{i\text{DIF}}$, d_i, and $d_{i\text{DIF}}$ to the open intervals, i.e., $c_i, d_i \in (0,1)$, $c_{i\text{DIF}} \in (-c_i, 1-c_i)$, and $d_{i\text{DIF}} \in (-d_i, 1-d_i)$. In case that parameters of the asymptotes are on the boundary of parameter space, i.e., $c_i = 0$, $d_i = 1$, and $c_{i\text{DIF}} = d_{i\text{DIF}} = 0$, the logistic regression model (7) may be used instead. Model (1.1) with some of the parameters fixed, e.g., $d_i = 1$ and $d_{i\text{DIF}} = 0$, may be also considered analogously. Note that the asymptotic properties derived here will hold also for such submodels, however, we are limited in that it is not possible to test whether the full model or its submodel fit the data better.

Regarding the condition [R3], in our case, the polynomial of x of the fourth degree can be taken as an integrable dominating function (see Appendix A.1).

Considering that X_p is the standardized total score, we can assume that the range of X_p is bounded. Moreover, $G_p \in \{0,1\}$ and thus partial derivatives $\frac{\partial \pi_{pi}}{\partial \gamma_i}$, $k = 1, \ldots, 8$, are all bounded. Thus matrices $\mathbb{\Gamma}_i(\boldsymbol{\gamma}_i)$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\gamma}_i)$ are both finite and the condition [R5] holds. Finally, in case that rows/columns of the matrix $\mathbb{\Gamma}_i(\boldsymbol{\gamma}_i)$ are linearly independent, the matrix has a full rank and therefore it is regular, satisfying condition [R4]. Singularity of the matrix may occur, for instance, when $G_p = 0, \forall p$ (or $G_p = 1, \forall p$).

We have shown that all assumptions [R1]–[R5] hold under the mild additional conditions, and thus $\hat{\boldsymbol{\gamma}}_i$ has pleasant properties such as consistency and asymptotic normality.

Estimate of asymptotic variance. The natural estimate of the asymptotic variance of $\hat{\boldsymbol{\gamma}}_i$ is a *sandwich estimator* given by

$$\frac{1}{n}\widehat{\mathbb{\Gamma}}_{in}^{-1}(\hat{\boldsymbol{\gamma}}_i)\widehat{\boldsymbol{\Sigma}}_{in}(\hat{\boldsymbol{\gamma}}_i)\widehat{\mathbb{\Gamma}}_{in}^{-1}(\hat{\boldsymbol{\gamma}}_i), \tag{1.15}$$

where

$$\widehat{\mathbb{\Gamma}}_{in}(\hat{\boldsymbol{\gamma}}_i) = \frac{1}{n}\sum_{p=1}^{n}\left(\frac{\partial \boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \hat{\boldsymbol{\gamma}}_i)}{\partial \boldsymbol{\gamma}_i^\top}\right),$$

$$\widehat{\boldsymbol{\Sigma}}_{in}(\boldsymbol{\gamma}_i) = \frac{1}{n}\sum_{p=1}^{n}\left(\boldsymbol{\psi}_i(Y_{pi}, X_p, G_p; \hat{\boldsymbol{\gamma}}_i)\boldsymbol{\psi}_i^\top(Y_{pi}, X_p, G_p; \hat{\boldsymbol{\gamma}}_i)\right).$$

Components of the matrix $\nabla^2 \text{RSS}_i(\boldsymbol{\gamma}_i) = \widehat{\mathbb{\Gamma}}_{in}(\boldsymbol{\gamma}_i)$ are given by:

$$\begin{aligned}
\frac{\partial^2 \text{RSS}_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}\partial \gamma_{ij}} &= -2\sum_{p=1}^{n}\left\{(Y_{pi} - \pi_{pi})\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik}\partial \gamma_{ij}} - \frac{\partial \pi_{pi}}{\partial \gamma_{ik}}\frac{\partial \pi_{pi}}{\partial \gamma_{ij}}\right\},\\
\frac{\partial^2 \text{RSS}_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}^2} &= -2\sum_{p=1}^{n}\left\{(Y_{pi} - \pi_{pi})\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik}^2} - \left(\frac{\partial \pi_{pi}}{\partial \gamma_{ik}}\right)^2\right\},
\end{aligned}$$

for $k = 1, \ldots, 8$, $j = 1, \ldots, 8$, and $k \neq j$, where

$$\begin{aligned}
\frac{\partial^2 \pi_{pi}}{\partial a_i^2} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial a_{i\mathrm{DIF}}} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_i \partial a_{i\mathrm{DIF}}}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial b_i} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_i \partial b_i}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial b_{i\mathrm{DIF}}} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_i \partial b_{i\mathrm{DIF}}}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial c_i} &= -\frac{\partial^2 \phi_{pi}}{\partial a_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial c_{i\mathrm{DIF}}} &= -\frac{\partial^2 \phi_{pi}}{\partial a_i^2} G_p, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial d_i} &= \frac{\partial^2 \phi_{pi}}{\partial a_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_i \partial d_{i\mathrm{DIF}}} &= \frac{\partial^2 \phi_{pi}}{\partial a_i^2} G_p, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}}^2} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial b_i} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}} \partial b_i}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial b_{i\mathrm{DIF}}} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}} \partial b_{i\mathrm{DIF}}}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial c_i} &= -\frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial c_{i\mathrm{DIF}}} &= -\frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2} G_p, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial d_i} &= \frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial a_{i\mathrm{DIF}} \partial d_{i\mathrm{DIF}}} &= \frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2} G_p, \qquad (1.16) \\
\frac{\partial^2 \pi_{pi}}{\partial b_i^2} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial b_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial b_i \partial b_{i\mathrm{DIF}}} &= (d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p) \frac{\partial^2 \phi_{pi}}{\partial b_i \partial b_{i\mathrm{DIF}}}, \\
\frac{\partial^2 \pi_{pi}}{\partial b_i \partial c_i} &= -\frac{\partial^2 \phi_{pi}}{\partial b_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial b_i \partial c_{i\mathrm{DIF}}} &= -\frac{\partial^2 \phi_{pi}}{\partial b_i^2} G_p, \\
\frac{\partial^2 \pi_{pi}}{\partial b_i \partial d_i} &= \frac{\partial^2 \phi_{pi}}{\partial b_i^2}, \\
\frac{\partial^2 \pi_{pi}}{\partial b_i \partial d_{i\mathrm{DIF}}} &= \frac{\partial^2 \phi_{pi}}{\partial b_i^2} G_p,
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \pi_{pi}}{\partial b_{i\mathrm{DIF}}^2} &= (d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)\frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2},\\
\frac{\partial^2 \pi_{pi}}{\partial b_{i\mathrm{DIF}}\partial c_i} &= -\frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2},\\
\frac{\partial^2 \pi_{pi}}{\partial b_{i\mathrm{DIF}}\partial c_{i\mathrm{DIF}}} &= -\frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2}G_p,\\
\frac{\partial^2 \pi_{pi}}{\partial b_{i\mathrm{DIF}}\partial d_i} &= \frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2},\\
\frac{\partial^2 \pi_{pi}}{\partial b_{i\mathrm{DIF}}\partial d_{i\mathrm{DIF}}} &= \frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2}G_p,\\
\frac{\partial^2 \pi_{pi}}{\partial c_i^2} &= \frac{\partial^2 \pi_{pi}}{\partial c_i \partial c_{i\mathrm{DIF}}} = \frac{\partial^2 \pi_{pi}}{\partial c_i \partial d_i} = \frac{\partial^2 \pi_{pi}}{\partial c_i \partial d_{i\mathrm{DIF}}} = 0,\\
\frac{\partial^2 \pi_{pi}}{\partial c_{i\mathrm{DIF}}^2} &= \frac{\partial^2 \pi_{pi}}{\partial c_{i\mathrm{DIF}} \partial d_i} = \frac{\partial^2 \pi_{pi}}{\partial c_{i\mathrm{DIF}} \partial d_{i\mathrm{DIF}}} = 0,\\
\frac{\partial^2 \pi_{pi}}{\partial d_i^2} &= \frac{\partial^2 \pi_{pi}}{\partial d_i \partial d_{i\mathrm{DIF}}} = \frac{\partial^2 \pi_{pi}}{\partial d_{i\mathrm{DIF}}^2} = 0,
\end{aligned}
$$

with

$$
\begin{aligned}
\frac{\partial^2 \phi_{pi}}{\partial a_i^2} &= \phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(X_{pi}-b_i-b_{i\mathrm{DIF}}G_p\right)^2,\\
\frac{\partial^2 \phi_{pi}}{\partial a_{i\mathrm{DIF}}^2} &= \frac{\partial^2 \phi_{pi}}{\partial a_i \partial a_{i\mathrm{DIF}}} = \phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(X_{pi}-b_i-b_{i\mathrm{DIF}}G_p\right)^2 G_p,\\
\frac{\partial^2 \phi_{pi}}{\partial a_i \partial b_i} &= -\phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(a_i+a_{i\mathrm{DIF}}G_p\right)\left(X_{pi}-b_i-b_{i\mathrm{DIF}}G_p\right),\\
\frac{\partial^2 \phi_{pi}}{\partial a_i \partial b_{i\mathrm{DIF}}} &= -\phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(a_i+a_{i\mathrm{DIF}}G_p\right)\left(X_{pi}-b_i-b_{i\mathrm{DIF}}G_p\right)G_p,\\
\frac{\partial^2 \phi_{pi}}{\partial b_i^2} &= -\phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(a_i+a_{i\mathrm{DIF}}G_p\right)^2,\\
\frac{\partial^2 \phi_{pi}}{\partial b_i \partial b_{i\mathrm{DIF}}} &= \frac{\partial^2 \phi_{pi}}{\partial b_{i\mathrm{DIF}}^2} = -\phi_{pi}\left(1-\phi_{pi}\right)\left(1-2\phi_{pi}\right)\left(a_i+a_{i\mathrm{DIF}}G_p\right)^2 G_p.
\end{aligned}
$$

In practice, all parameters are estimated simultaneously using suitable numerical approaches, such as ”nl2sol” algorithm from the Port library (Dennis, Gay, & Welsch, 1981; Gay, n.d.) which accounts for bounds of the estimated parameters, i.e., keeps asymptotes into the interval of $(0, 1)$. The nonlinear least squares estimation, described in this section, is implemented in the **difNLR** package (see Section 1.4.1).

It should be noted, that the nonlinear least squares method as presented here does not account for heteroscedasticity of binary data. In case of binary data, the Pearson's residuals might be more appropriate to use. This choice takes the original squares of residuals and divides them by the variance $\pi_{pi}(1-\pi_{pi})$. The RSS of item i (1.4) would take the following form:

$$
\mathrm{RSS}_i(\boldsymbol{\gamma}_i) = \sum_{p=1}^{n} \frac{\left(Y_{pi}-\pi_{pi}\right)^2}{\pi_{pi}\left(1-\pi_{pi}\right)}.
$$

1.2.2 Maximum likelihood

The second option to estimate item parameters in the model (1.1) is the *maximum likelihood* method. Using the notation from the previous section, the likelihood function for the item i has the following form:

$$L_i(\boldsymbol{\gamma}_i) = \prod_{p=1}^{n} \pi_{pi}^{Y_{pi}} (1 - \pi_{pi})^{1-Y_{pi}} .$$

The log-likelihood function for the item i is then given by

$$l_i(\boldsymbol{\gamma}_i) = \sum_{p=1}^{n} \{Y_{pi} \log(\pi_{pi}) + (1 - Y_{pi}) \log(1 - \pi_{pi})\} , \tag{1.17}$$

that is a log-likelihood for binary data. The parameter estimates are obtained by a maximization of the log-likelihood function (1.17) and we can therefore proceed similarly as for the logistic regression model, however, the nonlinear model (1.1) is no longer generalized linear model with the canonical link function. Maximization involves calculation of the score statistics, that is the first partial derivatives of the log-likelihood function with respect to the parameters of vector $\boldsymbol{\gamma}_i$. The general form of the partial derivatives by the parameter $k = 1, \ldots, 8$ is then

$$\frac{\partial l_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}} = \sum_{i=1}^{n} \frac{\partial \pi_{pi}}{\partial \gamma_{ik}} \frac{Y_{pi} - \pi_{pi}}{\pi_{pi}(1 - \pi_{pi})}, \tag{1.18}$$

with $\frac{\partial \pi_{pi}}{\partial \gamma_{ik}}$ given by (1.6)–(1.13).

To find the critical points of the log-likelihood function (1.17), the first partial derivatives (1.18) are set to zero and these likelihood equations are to be solved. However, the solution $\hat{\boldsymbol{\gamma}}_i$ of a system of the nonlinear equations cannot be derived algebraically and it has to be numerically estimated using a suitable iterative process. For example, modification of the quasi-Newton method allowing for box constraints (Byrd, Lu, Nocedal, & Zhu, 1995) can be used in which each parameter may be given a lower and/or upper bound.

Elements of the observed information matrix

$$\mathbb{I}_{in}(\boldsymbol{\gamma}_i | \boldsymbol{X}, \boldsymbol{G}) = \frac{1}{n} \sum_{p=1}^{n} \mathbb{I}_i(\boldsymbol{\gamma}_i | X_p, G_p) = -\frac{1}{n} \frac{\partial^2 l_i(\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top},$$

that is the matrix of the second partial derivatives of the log-likelihood function (1.17), are given by

$$\begin{aligned}
\frac{\partial^2 l_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}^2} &= \sum_{p=1}^{n} \frac{Y_{pi} - \pi_{pi}}{\pi_{pi}(1 - \pi_{pi})} \left[\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik}^2} - \left(\frac{\partial \pi_{pi}}{\partial \gamma_{ik}} \right)^2 \frac{Y_{pi} - \pi_{pi}}{\pi_{pi}(1 - \pi_{pi})} \right], \\
\frac{\partial^2 l_i(\boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \gamma_{ij}} &= \sum_{p=1}^{n} \frac{Y_{pi} - \pi_{pi}}{\pi_{pi}(1 - \pi_{pi})} \left[\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik} \partial \gamma_{ij}} - \frac{\partial \pi_{pi}}{\partial \gamma_{ik}} \frac{\partial \pi_{pi}}{\partial \gamma_{ij}} \frac{Y_{pi} - \pi_{pi}}{\pi_{pi}(1 - \pi_{pi})} \right],
\end{aligned}$$

for $k, j = 1, \ldots, 8$, $k \neq j$, and the second partial derivatives $\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik}^2}$ and $\frac{\partial^2 \pi_{pi}}{\partial \gamma_{ik} \partial \gamma_{ij}}$ given by the terms (1.16).

The Fisher information matrix then has the form of

$$\mathbb{I}_i(\boldsymbol{\gamma}_i) = \mathsf{E}\,\mathbb{I}_i(\boldsymbol{\gamma}_i|X_p, G_p) = \mathsf{E}\left(\frac{1}{\pi_{pi}\left(1-\pi_{pi}\right)}\frac{\partial \pi_{pi}}{\partial \gamma_{ik}}\frac{\partial \pi_{pi}}{\partial \gamma_{ij}}\right)_{k,j=1}^{8}, \tag{1.19}$$

where the first partial derivatives $\frac{\partial \pi_{pi}}{\partial \gamma_{ik}}$, $k = 1, \ldots, 8$, are given by (1.6)–(1.13), which is in fact a quadratic form and thus it is positive definite.

Asymptotic properties. Asymptotic properties of the maximum likelihood estimator can be shown under the set of the following regularity conditions:

[R0*] The support set $S = \{y \in \mathbb{R} : f(y|x, g, \boldsymbol{\gamma}_i) > 0\}$ does not depend on the parameter $\boldsymbol{\gamma}_i$.

[R1*] The true parameter $\boldsymbol{\gamma}_{iX}$ is an interior point of the parameter space.

[R2*] The density $f(y|x, g, \boldsymbol{\gamma}_i) = y \log\left(\pi(x, g; \boldsymbol{\gamma}_i)\right) + (1-y)\log\left(1 - \pi(x, g; \boldsymbol{\gamma}_i)\right)$ is twice continuously differentiable with respect to $\boldsymbol{\gamma}_i$ for each (y, x, g).

[R3*] The Fisher information matrix $\mathbb{I}_i(\boldsymbol{\gamma}_i)$ is finite, regular, and positive definite in a neighborhood of γ_{iX}.

[R4*] The order of differentiation and integration with respect to $\boldsymbol{\gamma}_i$ can be interchanged for terms $f(y|x, g, \boldsymbol{\gamma}_i)$ and $\frac{\partial f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}$.

It is easy to see that the conditions [R0*] and [R2*] hold. In our case, the regularity condition for the maximum likelihood estimator [R1*] is the same as condition [R1] for the nonlinear least squares and we thus need to bound parameters of asymptotes to open intervals as was discussed in the previous Section 1.2.1.

Regarding the condition [R3*], we have already shown that the Fisher information matrix (1.19) is positive definite. Again, considering X_p to be the standardized total score, we can assume that its range is bounded. Together with the fact that $G_p \in \{0, 1\}$, it is easy to see that partial derivatives $\frac{\partial \pi_{pi}}{\partial \gamma_{ik}}$, $k = 1, \ldots, 8$, are all bounded and thus the Fisher information matrix is finite. Similarly as in Section 1.2.1, in case that rows/columns of the Fisher information matrix $\mathbb{I}_i(\boldsymbol{\gamma}_i)$ are linearly independent, the matrix has a full rank and therefore it is regular, satisfying the condition [R3*]. Singularity of the matrix may occur in similar cases as for the matrix $\mathbb{\Gamma}_i(\boldsymbol{\gamma}_i)$ described in Section 1.2.1.

Finally, regarding the condition [R4*], the order of differentiation and integration can be interchanged by dominated convergence theorem, as far as both $\frac{\partial f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}$ and $\frac{\partial^2 f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}$ are dominated by an integrable function. In our case a polynomial of x of the fourth degree can be taken as an integrable dominating function (see Appendix A.2).

From Hogg, McKean, and Craig (2018) it follows that when the regularity conditions [R0*]–[R4*] hold, then there exists n_0 and a sequence $\widehat{\boldsymbol{\gamma}}_{in}(n > n_0)$ of solutions to the likelihood equations such that

$$\widehat{\boldsymbol{\gamma}}_{in} \xrightarrow[n\to\infty]{P} \boldsymbol{\gamma}_{iX},$$

where $\boldsymbol{\gamma}_{iX}$ is a vector of true parameters. As the log-likelihood function (1.17) is not strictly concave, described approach does not guarantee finding a unique solution of the corresponding likelihood equations. In other words, there may be multiple solutions, each of them being a local maximum. However, there is one solution among them, which provides a consistent sequence of estimators, while other solutions may not even be close to γ_{iX} and may not converge to it. Therefore, in practice, the crucial part of estimating procedures is to find suitable starting values, preferably easily calculated but consistent estimate of parameters, see also Section 1.6.1. Further, for such consistent sequence of solutions to the likelihood equations it can be shown that

$$\sqrt{n}\,(\hat{\boldsymbol{\gamma}}_{in} - \boldsymbol{\gamma}_{iX}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbb{I}_i^{-1}(\boldsymbol{\gamma}_{iX})).$$

Estimate of asymptotic variance. Estimate of the asymptotic variance of the item parameters $\hat{\boldsymbol{\gamma}}_i$ is an inverse of the observed information matrix, i.e., an inverse of the Hessian matrix:

$$\mathbb{I}_{in}^{-1}(\hat{\boldsymbol{\gamma}}_i|\boldsymbol{X}, \boldsymbol{G}) = \left(-\frac{1}{n}\frac{\partial^2 l_i(\hat{\boldsymbol{\gamma}}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right)^{-1}. \tag{1.20}$$

The maximum likelihood estimation, covered in this section, is implemented in the **difNLR** package which is described in Section 1.4.1.

1.2.3 EM algorithm

The third approach to estimate parameters in the model (1.1) is to use the *EM algorithm*. The EM algorithm is another approach to get approximation of the maximum likelihood estimates of item parameters, as was described in Section 1.2.2.

The problem can be reformulated in the context of the latent variables (Dinse, 2011). In our setting, we consider four mutually exclusive latent variables (Z_{pi1}, Z_{pi2}, Z_{pi3}, Z_{pi4}), where variable $Z_{pij} = 1$ indicates that respondent p belongs to the category $j = 1, \ldots, 4$ for an item i, while $Z_{pij} = 0$ indicates he/she does not belong to this category.

The categories 1 and 2 denote whether a respondent who correctly answered item i (i.e., $Y_{pi} = 1$) guessed correct answer while their knowledge or ability was insufficient ($Z_{pi1} = 1$) or had a sufficient knowledge or ability to do so and did not guessed ($Z_{pi2} = 1$). The categories 3 and 4, on the other hand, point to whether the respondent who did not answer correctly item i (i.e., $Y_{pi} = 0$) did not have sufficient knowledge or ability ($Z_{pi3} = 1$) or incorrectly answered due to another reason such as inattention ($Z_{pi4} = 1$). The observed indicator Y_{pi} and its complement $1 - Y_{pi}$ can be then rewritten as $Y_{pi} = Z_{pi1} + Z_{pi2}$ and $1 - Y_{pi} = Z_{pi3} + Z_{pi4}$ (see Figure 1.3).

Let $c_i + c_{i\mathrm{DIF}} G_p$ be the probability that respondent p from the group G_p guessed item i correctly without necessary knowledge (category 1) and let $d_i + d_{i\mathrm{DIF}} G_p$ be the probability that respondent p from the group G_p was not inattentive in that item (categories 1–3). Then $d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p$ gives the probability that respondent p from the group G_p did not guess and was not inattentive (categories 2 and 3). For these two categories, ϕ_{pi} and $1 - \phi_{pi}$ are the probabilities to answer given item correctly (category 2) and incorrectly (category 3), respectively, depending on the regressors X_p and G_p. Finally, the probability of respondent p

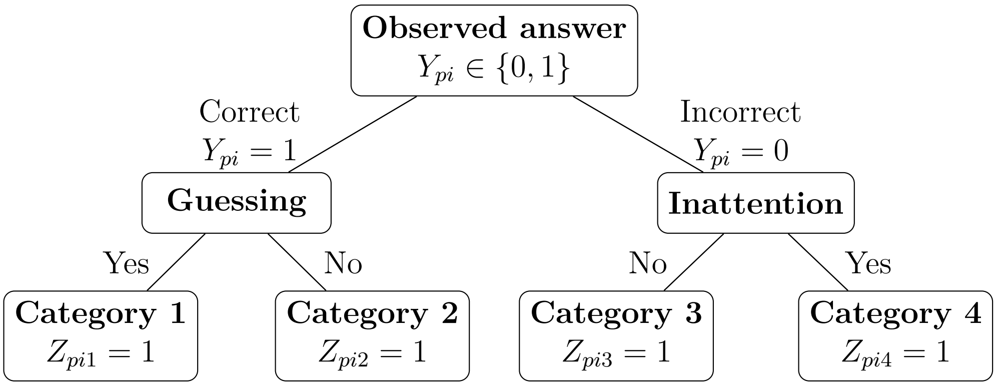


Figure 1.3: Latent variables for the EM algorithm.

from the group G_p to be inattentive in item i is $1 - (d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p) - c_i - c_{i\mathrm{DIF}}G_p = 1 - d_i - d_{i\mathrm{DIF}}G_p$ (category 4). In summary, the expected values of the latent variables are then given by terms

$$\begin{aligned}
\mathsf{E}(Z_{pi1}|X_p, G_p) &= c_i + c_{\mathrm{iDIF}}G_p,\\
\mathsf{E}(Z_{pi2}|X_p, G_p) &= (d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)\phi_{pi},\\
\mathsf{E}(Z_{pi3}|X_p, G_p) &= (d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)(1 - \phi_{pi}),\\
\mathsf{E}(Z_{pi4}|X_p, G_p) &= 1 - d_i - d_{i\mathrm{DIF}}G_p.
\end{aligned} \tag{1.21}$$

Note that $\mathsf{E}(Z_{pij}|X_p, G_p) = \mathsf{P}(Z_{pij} = 1|X_p, G_p)$, $j = 1, \ldots, 4$, as the latent variables Z_{pij} are dichotomous. The probability of correct answer can be then expressed as

$$\begin{aligned}
\mathsf{P}(Y_{pi} = 1|X_p, G_p) &= \mathsf{P}(Z_{pi1} + Z_{pi2} = 1|X_p, G_p)\\
&= \mathsf{P}(Z_{pi1} = 1|X_p, G_p) + \mathsf{P}(Z_{pi2} = 1|X_p, G_p)\\
&= c_i + c_{i\mathrm{DIF}}G_p + (d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)\phi_{pi},
\end{aligned}$$

which under the logistic model

$$\phi_{pi} = \frac{e^{(a_i + a_{i\mathrm{DIF}}G_p)(X_p - b_i - b_{i\mathrm{DIF}}G_p)}}{1 + e^{(a_i + a_{i\mathrm{DIF}}G_p)(X_p - b_i - b_{i\mathrm{DIF}}G_p)}}$$

results into the model (1.1). In other words, $Z_{pi} = (Z_{pi1}, Z_{pi2}, Z_{pi3}, Z_{pi4})$ has a multinomial distribution with one trial and corresponding probabilities given by (1.21).

The log-likelihood function for the item i takes the following form:

$$\begin{aligned}
l_i^{\mathrm{EM}}(\boldsymbol{\gamma}_i) =& \sum_{p=1}^{n} \{Z_{pi1} \log\left(\mathsf{P}(Z_{pi1} = 1|X_p, G_p)\right) + Z_{pi2} \log\left(\mathsf{P}(Z_{pi2} = 1|X_p, G_p)\right)\\
&\quad + Z_{pi3} \log\left(\mathsf{P}(Z_{pi3} = 1|X_p, G_p)\right) + Z_{pi4} \log\left(\mathsf{P}(Z_{pi4} = 1|X_p, G_p)\right)\}\\
=& \sum_{p=1}^{n} \{Z_{pi1} \log\left(c_i + c_{i\mathrm{DIF}}G_p\right) + Z_{pi2} \log\left((d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)\phi_{pi}\right)\\
&\quad + Z_{pi3} \log\left((d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p)(1 - \phi_{pi})\right)\\
&\quad + Z_{pi4} \log\left(1 - d_i - d_{i\mathrm{DIF}}G_p\right)\}\\
=& \sum_{p=1}^{n} \{Z_{pi2} \log\left(\phi_{pi}\right) + Z_{pi3} \log\left(1 - \phi_{pi}\right)\}
\end{aligned} \tag{1.22}$$

$$
\begin{aligned}
&+\sum_{p=1}^{n}\left\{Z_{pi1}\log\left(c_i + c_{i\mathrm{DIF}}G_p\right) + Z_{pi4}\log\left(1 - d_i - d_{i\mathrm{DIF}}G_p\right)\right. \\
&\qquad \left. +\ \left(Z_{pi2}+Z_{pi3}\right)\log\left(d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p\right)\right\} \qquad (1.23) \\
&= l_{i1}^{\mathrm{EM}} + l_{i2}^{\mathrm{EM}},
\end{aligned}
$$

where l_{i1}^{EM} and l_{i2}^{EM} are given by terms (1.22) and (1.23), respectively.

The log-likelihood function l_{i1}^{EM} (1.22) includes only parameters a_i, $a_{i\mathrm{DIF}}$, b_i, and $b_{i\mathrm{DIF}}$, while the log-likelihood function l_{i2}^{EM} (1.23) incorporates only parameters related to the asymptotes of the ICCs and does not include regressor X_p. The first log-likelihood function (1.22) actually has a form of the log-likelihood function for the logistic regression. However, in contrast to logistic regression, here it does not necessary hold that $Z_{pi2} + Z_{pi3} = 1$ as the correct answer could be guessed or respondent could be inattentive which would result in $Z_{pi2} + Z_{pi3} = 0$. The second log-likelihood function (1.23) takes a form of the log-likelihood for multinomial data with one trial and with the group-based probabilities $c_i + c_{i\mathrm{DIF}}G_p$, $d_i + d_{i\mathrm{DIF}}G_p - c_i - c_{i\mathrm{DIF}}G_p$, and $1 - d_i - d_{i\mathrm{DIF}}G_p$.

E-step. At the E-step, conditionally on the item responses Y_{pi} and the current parameter estimate $\hat{\boldsymbol{\gamma}}_i = (\widehat{a}_i, \widehat{a}_{i\mathrm{DIF}}, \widehat{b}_i, \widehat{b}_{i\mathrm{DIF}}, \widehat{c}_i, \widehat{c}_{i\mathrm{DIF}}, \widehat{d}_i, \widehat{d}_{i\mathrm{DIF}})$, the estimates of the latent variables are calculated as their expected values using the Bayes's theorem and (1.21):

$$
\begin{aligned}
\widehat{Z}_{pi1} &= \mathsf{E}(Z_{pi1}|Y_{pi}, X_p, G_p, \hat{\boldsymbol{\gamma}}_i) \\
&= \mathsf{P}(Z_{pi1} = 1|Y_{pi}, X_p, G_p, \hat{\boldsymbol{\gamma}}_i) \\
&= \frac{\mathsf{P}(Z_{pi1} = 1 \ \&\ Y_{pi} = y|X_p, G_p, \hat{\boldsymbol{\gamma}}_i)}{\mathsf{P}(Y_{pi} = y|X_p, G_p, \hat{\boldsymbol{\gamma}}_i)} \\
&= \begin{cases} \frac{\widehat{c}_i + \widehat{c}_{i\mathrm{DIF}}G_p}{\widehat{c}_i + \widehat{c}_{i\mathrm{DIF}}G_p + (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, & y = 1 \\ \frac{0}{1 - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p - (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, & y = 0 \end{cases} \\
&= \frac{Y_{pi}\left(\widehat{c}_i + \widehat{c}_{i\mathrm{DIF}}G_p\right)}{\widehat{c}_i + \widehat{c}_{i\mathrm{DIF}}G_p + (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, \\
\widehat{Z}_{pi2} &= Y_{pi} - \widehat{Z}_{pi1},
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{Z}_{pi4} &= \mathsf{E}(Z_{pi4}|Y_{pi}, X_p, G_p, \hat{\boldsymbol{\gamma}}_i) \\
&= \mathsf{P}(Z_{pi4} = 1|Y_{pi}, X_p, G_p, \hat{\boldsymbol{\gamma}}_i) \\
&= \frac{\mathsf{P}(Z_{pi4} = 1 \ \&\ Y_{pi} = y|X_p, G_p, \hat{\boldsymbol{\gamma}}_i)}{\mathsf{P}(Y_{pi} = y|X_p, G_p, \hat{\boldsymbol{\gamma}}_i)} \\
&= \begin{cases} \frac{0}{\widehat{c}_i + \widehat{c}_{i\mathrm{DIF}}G_p + (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, & y = 1 \\ \frac{1 - d_i - d_{i\mathrm{DIF}}G_p}{1 - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p - (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, & y = 0 \end{cases} \\
&= \frac{\left(1 - Y_{pi}\right)\left(1 - \widehat{d}_i - \widehat{d}_{i\mathrm{DIF}}G_p\right)}{1 - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p - (\widehat{d}_i + \widehat{d}_{i\mathrm{DIF}}G_p - \widehat{c}_i - \widehat{c}_{i\mathrm{DIF}}G_p)\widehat{\phi}_{pi}}, \\
\widehat{Z}_{pi3} &= 1 - Y_{pi} - \widehat{Z}_{pi4}.
\end{aligned}
$$

M-step. At the M-step, conditionally on the current estimates of the latent variables $\widehat{Z}_{pi2}$ and $\widehat{Z}_{pi3}$, the estimates of parameters $\boldsymbol{\gamma}_{i1} = \{\gamma_{ik}\}_{k=1}^{4} = \{a_i, a_{i\mathrm{DIF}}, b_i, b_{i\mathrm{DIF}}\}$ maximize the log-likelihood function l_{i1}^{EM} (1.22). As noted, (1.22) has the form as the log-likelihood for the logistic regression, therefore, we proceed analogously. This again involves a calculation of its partial derivatives with respect to the item parameters:

$$\frac{\partial l_{i1}^{\mathrm{EM}}}{\partial \gamma_{i1k}} = \sum_{p=1}^{n} \frac{\partial \phi_{pi}}{\partial \gamma_{i1k}} \left(\frac{\widehat{Z}_{pi2}}{\phi_{pi}} - \frac{\widehat{Z}_{pi3}}{1-\phi_{pi}} \right),$$

for $k = 1, \ldots, 4$, where partial derivatives of $\frac{\partial \phi_{pi}}{\partial \gamma_{i1k}}$ are given by (1.14). However, the corresponding likelihood equations $\frac{\partial l_{i1}^{\mathrm{EM}}}{\partial \gamma_{ik}} \overset{!}{=} 0$ have no closed form and estimates need to be calculated using appropriate numerical methods. Considering the fact that the log-likelihood function l_{i1}^{EM} (1.22) takes a form of the log-likelihood for the logistic regression model, iteratively re-weighted least squares may be used in practice.

The estimates $\widehat{c}_i$, $\widehat{c}_{i\mathrm{DIF}}$, $\widehat{d}_i$, and $\widehat{d}_{i\mathrm{DIF}}$ are given by a maximization of the log-likelihood function l_{i2}^{EM} (1.23) conditionally on current estimates of the latent variables $\widehat{Z}_{pi1}$, $\widehat{Z}_{pi2}$, $\widehat{Z}_{pi3}$, and $\widehat{Z}_{pi4}$. As mentioned, (1.23) has the form of the log-likelihood for the multinomial data, thus, we proceed analogously. This again involves a calculation of its partial derivatives with respect to the item parameters:

$$\begin{aligned}
\frac{\partial l_{i2}^{\mathrm{EM}}}{\partial c_i} &= \sum_{p=1}^{n} \left(\frac{\widehat{Z}_{pi1}}{c_i + c_{i\mathrm{DIF}} G_p} - \frac{\widehat{Z}_{pi2} + \widehat{Z}_{pi3}}{d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p} \right), \\
\frac{\partial l_{i2}^{\mathrm{EM}}}{\partial c_{i\mathrm{DIF}}} &= \sum_{p=1}^{n} \left(\frac{\widehat{Z}_{pi1}}{c_i + c_{i\mathrm{DIF}} G_p} - \frac{\widehat{Z}_{pi2} + \widehat{Z}_{pi3}}{d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p} \right) G_p, \\
\frac{\partial l_{i2}^{\mathrm{EM}}}{\partial d_i} &= \sum_{p=1}^{n} \left(\frac{\widehat{Z}_{pi2} + \widehat{Z}_{pi3}}{d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p} - \frac{\widehat{Z}_{pi4}}{1 - d_i - d_{i\mathrm{DIF}} G_p} \right), \\
\frac{\partial l_{i2}^{\mathrm{EM}}}{\partial d_{i\mathrm{DIF}}} &= \sum_{p=1}^{n} \left(\frac{\widehat{Z}_{pi2} + \widehat{Z}_{pi3}}{d_i + d_{i\mathrm{DIF}} G_p - c_i - c_{i\mathrm{DIF}} G_p} - \frac{\widehat{Z}_{pi4}}{1 - d_i - d_{i\mathrm{DIF}} G_p} \right) G_p.
\end{aligned}$$

As in the previous case, the corresponding equations $\frac{\partial l_{i2}^{\mathrm{EM}}}{\partial \gamma_{ik}} \overset{!}{=} 0$, $k = 5, \ldots, 8$, do not have a closed form and suitable numerical techniques need to be applied, such as the quasi-Newton method allowing for the box constraints (Byrd et al., 1995) to account for upper and lower bounds of the asymptote parameters c_i, $c_{i\mathrm{DIF}}$, d_i, and $d_{i\mathrm{DIF}}$.

Estimate of asymptotic variance. As was noted, the EM algorithm is designed to gain maximum likelihood estimates of the item parameters. Therefore, the asymptotic properties of the estimates are analogous to those derived in Section 1.2.2, and also the estimate of the asymptotic covariance matrix is the same, i.e., the inverse of the Hessian matrix (1.20).

In summary, introducing the latent variables Z_{pi1}, Z_{pi2}, Z_{pi3}, and Z_{pi4} leads to simplification of the estimation as we work with the likelihoods of the logistic regression and multinomial regression models. The proposed EM algorithm is therefore easy to implement in the `R` software and can take advantage of its existing functions, see Section 1.4.3. The empirical evidence of the convergence of the parameter estimates based on the EM algorithm is given in Section 1.6.

1.2.4 Parametric link function

In our setting, the nonlinear regression model (1.1) can be viewed as a generalized linear model with a known parametric link function

$$\begin{aligned} g(\mu_{pi}; c_i, c_{i\mathrm{DIF}}, d_i, d_{i\mathrm{DIF}}) &= \log\left(\frac{\frac{\mu_{pi}-c_i-c_{i\mathrm{DIF}}G_p}{d_i+d_{i\mathrm{DIF}}G_p-c_i-c_{i\mathrm{DIF}}G_p}}{1-\frac{\mu_{pi}-c_i-c_{i\mathrm{DIF}}G_p}{d_i+d_{i\mathrm{DIF}}G_p-c_i-c_{i\mathrm{DIF}}G_p}}\right) \\ &= \log\left(\frac{\mu_{pi}-c_i-c_{i\mathrm{DIF}}G_p}{d_i+d_{i\mathrm{DIF}}G_p-\mu_{pi}}\right), \end{aligned} \tag{1.24}$$

where the parameters c_i, $c_{i\mathrm{DIF}}$, d_i, and $d_{i\mathrm{DIF}}$ are unknown. The mean function is then determined by $\mu_{pi} = \pi_{pi}$ as given by (1.3) and (1.1) with a linear predictor

$$(a_i + a_{i\mathrm{DIF}}G_p)\,(X_p - b_i - b_{i\mathrm{DIF}}G_p)\,.$$

The topic of the parametric link functions has been extensively discussed in literature in the last decades by many authors including Basu and Rathouz (2005), Flach (2014), and Scallan, Gilchrist, and Green (1984). For example, Pregibon (1980) in his work proposed the maximum likelihood estimation of the link parameters using a weighted least squares algorithm. McCullagh and Nelder (1989) adapted this approach and presented an algorithm in which several models with the fixed link functions were fitted. Further, Kaiser (1997) proposed a modified scoring algorithm to perform simultaneous maximum likelihood estimation of all parameters. Scallan et al. (1984) proposed an iterative two-stage algorithm building on work by Richards (1961).

In this part we propose a new two-stage algorithm to estimate parameters $\boldsymbol{\gamma}_i = \{a_i, a_{i\mathrm{DIF}}, b_i, b_{i\mathrm{DIF}}, c_i, c_{i\mathrm{DIF}}, d_i, d_{i\mathrm{DIF}}\}$ in the model (1.1). Let $\boldsymbol{\gamma}_{i1} = \{\gamma_{ik}\}_{k=1}^{4} = \{a_i, a_{i\mathrm{DIF}}, b_i, b_{i\mathrm{DIF}}\}$, $\boldsymbol{\gamma}_{i2} = \{\gamma_{ik}\}_{k=5}^{8} = \{c_i, c_{i\mathrm{DIF}}, d_i, d_{i\mathrm{DIF}}\}$ be the sets of the first four and the last four parameters. Further, let $\widehat{\boldsymbol{\gamma}}_{i1}$ and $\widehat{\boldsymbol{\gamma}}_{i2}$ be their estimates. This algorithm is designed to gain the maximum likelihood estimates of the item parameters, as was also the case of the EM algorithm described in Section 1.2.3.

We use the same notation of the logistic regression curve in case that respondents were not guessing or were not inattentive as before:

$$\phi_{pi} = \frac{e^{(a_i+a_{i\mathrm{DIF}}G_p)(X_p-b_i-b_{i\mathrm{DIF}}G_p)}}{1+e^{(a_i+a_{i\mathrm{DIF}}G_p)(X_p-b_i-b_{i\mathrm{DIF}}G_p)}}$$

with $\widehat{\phi}_{pi}$ being its estimator. To simplify the formulae in this section, we further set

$$\begin{aligned} c_{iG_p} &= c_i + c_{i\mathrm{DIF}}G_p, \\ d_{iG_p} &= d_i + d_{i\mathrm{DIF}}G_p. \end{aligned}$$

Note that the sets of terms c_{iG_p} and d_{iG_p} consist only of two parameters each, which further depend on the group membership variable G_p. In other words,

$$\begin{aligned} c_{i0} &= c_i, \quad c_{i1} = c_i + c_{i\mathrm{DIF}}, \\ d_{i0} &= d_i, \quad d_{i1} = d_i + d_{i\mathrm{DIF}}. \end{aligned}$$

First step. At the first step, conditionally on current estimates $\widehat{c}_i$, $\widehat{c}_{i\mathrm{DIF}}$, $\widehat{d}_i$, and $\widehat{d}_{i\mathrm{DIF}}$ of the parametric link function (1.24), the estimates of parameters a_i, $a_{i\mathrm{DIF}}$, b_i, and $b_{i\mathrm{DIF}}$ maximize the following log-likelihood function:

$$\begin{aligned} l_{i1}^{\mathrm{PL}}(\boldsymbol{\gamma}_{i1}|\widehat{\boldsymbol{\gamma}}_{i2}) = \sum_{p=1}^{n} \Big\{ & Y_{pi} \log(\widehat{c}_{iG_p} + (\widehat{d}_{iG_p} - \widehat{c}_{iG_p})\phi_{pi}) \\ & + (1 - Y_{pi}) \log(1 - \widehat{c}_{iG_p} - (\widehat{d}_{iG_p} - \widehat{c}_{iG_p})\phi_{pi}) \Big\}. \end{aligned} \tag{1.25}$$

The log-likelihood function l_{i1}^{PL} (1.25) has a similar form to the log-likelihood function (1.17) using the maximum likelihood method, however, parameters $\boldsymbol{\gamma}_{i2}$ are here replaced by their current estimates $\widehat{\boldsymbol{\gamma}}_{i2}$. The next steps are then analogous to those in the Section 1.2.2. The scores of the log-likelihood function l_{i1}^{PL} (1.25) have the following form:

$$\frac{\partial l_{i1}^{\mathrm{PL}}}{\partial \gamma_{i1k}} = \sum_{p=1}^{n} \frac{\left(\widehat{d}_{iG_p} - \widehat{c}_{iG_p}\right) \left[Y_{pi} - \widehat{c}_{iG_p} - (\widehat{d}_{iG_p} - \widehat{c}_{iG_p})\phi_{pi}\right]}{\left[\widehat{c}_{iG_p} + (\widehat{d}_{iG_p} - \widehat{c}_{iG_p})\phi_{pi}\right] \left[1 - \widehat{c}_{iG_p} - (\widehat{d}_{iG_p} - \widehat{c}_{iG_p})\phi_{pi}\right]} \frac{\partial \phi_{pi}}{\partial \gamma_{i1k}},$$

for $k = 1, \ldots, 4$, where $\frac{\partial \phi_{pi}}{\partial \gamma_{ik}}$ are given by (1.14).

Nevertheless, even in this case, the solution of the equations $\frac{\partial l_{i1}^{\mathrm{PL}}}{\partial \gamma_{i1k}} = 0$, $k = 1, \ldots, 4$ does not have a closed form and appropriate numerical approaches, such as iteratively re-weighted least squares, need to be applied.

Second step. At the second step, estimates $\widehat{c}_i$, $\widehat{c}_{i\mathrm{DIF}}$, $\widehat{d}_i$, and $\widehat{d}_{i\mathrm{DIF}}$ of the parametric link function (1.24) are calculated conditionally on the current estimates of the parameters $\widehat{\boldsymbol{\gamma}}_{i1} = \left\{\widehat{a}_i, \widehat{a}_{i\mathrm{DIF}}, \widehat{b}_i, \widehat{b}_{i\mathrm{DIF}}\right\}$ as the arguments of the maxima of the following log-likelihood function

$$\begin{aligned} l_{i2}^{\mathrm{PL}}(\boldsymbol{\gamma}_{i2}|\widehat{\boldsymbol{\gamma}}_{i1}) = \sum_{p=1}^{n} \Big\{ & Y_{pi} \log(c_{iG_p} + (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi}) \\ & + (1 - Y_{pi}) \log(1 - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi}) \Big\}. \end{aligned}$$

Again, the parameters a_i, $a_{i\mathrm{DIF}}$, b_i, and $b_{i\mathrm{DIF}}$ are replaced by their estimates $\widehat{a}_i$, $\widehat{a}_{i\mathrm{DIF}}$, $\widehat{b}_i$, and $\widehat{b}_{i\mathrm{DIF}}$ in (1.10)–(1.13) and ϕ_{pi} is replaced by $\widehat{\phi}_{pi}$. The scores then take the following form:

$$\begin{aligned} \frac{\partial l_{i2}^{\mathrm{PL}}}{\partial c_i} &= \sum_{p=1}^{n} \frac{\left[Y_{pi} - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left(1 - \widehat{\phi}_{pi}\right)}{\left[c_{iG_p} + (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left[1 - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right]}, \\ \frac{\partial l_{i2}^{\mathrm{PL}}}{\partial c_{i\mathrm{DIF}}} &= \sum_{p=1}^{n} \frac{\left[Y_{pi} - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left(1 - \widehat{\phi}_{pi}\right) G_p}{\left[c_{iG_p} + (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left[1 - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right]}, \\ \frac{\partial l_{i2}^{\mathrm{PL}}}{\partial d_i} &= \sum_{p=1}^{n} \frac{\left[Y_{pi} - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \widehat{\phi}_{pi}}{\left[c_{iG_p} + (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left[1 - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right]}, \\ \frac{\partial l_{i2}^{\mathrm{PL}}}{\partial d_{i\mathrm{DIF}}} &= \sum_{p=1}^{n} \frac{\left[Y_{pi} - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \widehat{\phi}_{pi} G_p}{\left[c_{iG_p} + (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right] \left[1 - c_{iG_p} - (d_{iG_p} - c_{iG_p})\widehat{\phi}_{pi})\right]}. \end{aligned}$$

As before, the solution of the equations $\frac{\partial l_{i2}^{\text{PL}}}{\partial \gamma_{i2k}} = 0$, $k = 1, \ldots, 4$, does not have a closed form and numerical algorithms, e.g., iteratively re-weighted least squares, need to be applied.

Estimate of asymptotic variance. Given the fact that the newly proposed algorithm is designed to gain the maximum likelihood estimates of the item parameters, the estimate of the asymptotic covariance matrix is again the inverse of the Hessian matrix (1.20), closer described in Section 1.2.2.

In summary, the division into the two sets of parameters makes the algorithm based on parametric link function easy to implement in the `R` software and can take advantage of its existing functions. As the algorithm is designed to gain the maximum likelihood estimates, their asymptotic properties are the same as was described in Section 1.2.2. The implementation of the proposed estimation method based on parametric link function into the `R` software is described in Section 1.4.4. The empirical evidence of the convergence of the estimates given by the newly proposed algorithm is offered in Section 1.6.

1.3 DIF detection

The nonlinear model (1.1) can be utilized to detect DIF in a simple way by comparing the two nested models for item i, where in the submodel M_{i0} some parameters are set the same for both groups, while in the model M_{i1} parameters are freely estimated for both groups.

It should be noted that the following tests are not intended to test, for example, whether the 4PL nonlinear model (1.1) is superior over the 3PL or the 2PL models, as the values 0 and 1 of the asymptote parameters are on the boundary of the parametric space, as was discussed in Sections 1.2.1 and 1.2.2. However, it is, for example, possible to test whether the lower asymptotes are the same for the two groups while fixing the upper asymptotes on value of 1, i.e., assuming the 3PL model.

F-test. The F-test statistic measures the distance between the larger model M_{i1} with df_{i1} being a number of parameters and its submodel M_{i0} with df_{i0} being a number of parameters for item i as the difference between the RSS (1.4) of the model M_{i1} relative to the RSS of the submodel M_{i0}. The formula is the same as for the linear models:

$$F_i = \frac{\frac{\text{RSS}_{i0} - \text{RSS}_{i1}}{\text{df}_{i0} - \text{df}_{i1}}}{\frac{\text{RSS}_{i1}}{n - \text{df}_{i1}}},$$

where n is a number of respondents. However, in the nonlinear models the $\mathcal{F}$-distribution under the submodel M_{i0} holds only approximately (Ritz & Streibig, 2008):

$$F_i \overset{\text{app.}}{\sim} \mathcal{F}(\text{df}_{i1} - \text{df}_{i0}, n - \text{df}_{i1}). \tag{1.26}$$

While it has not been shown that the distribution of the test statistic F_i holds for binary data as in our case, simulations showed its reasonable behaviour (see Section 1.5).

Likelihood-ratio test. The likelihood-ratio test measures the difference between the log-likelihood l_{i1} of the larger model M_{i1} and the log-likelihood l_{i0} of its submodel M_{i0} for the item i:

$$LR_i = -2\left(l_{i0} - l_{i1}\right).$$

The LR_i statistic has an asymptotic χ^2-distribution under the submodel M_{i0}:

$$LR_i \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(\mathrm{df}_{i1} - \mathrm{df}_{i0}). \tag{1.27}$$

Wald test. Another option is to use the Wald test with the null hypothesis $H_{i0} : (a_{i\mathrm{DIF}}, b_{i\mathrm{DIF}}, c_{i\mathrm{DIF}}, d_{i\mathrm{DIF}}) \supseteq \boldsymbol{\gamma}_i^* = \mathbf{0}$ vs. alternative hypothesis $H_{i1} : \boldsymbol{\gamma}_i^* \neq \mathbf{0}$. The test statistic is then given by

$$W_i = \hat{\boldsymbol{\gamma}}_i^* \widehat{\mathbb{V}}_{in}^{-1} \hat{\boldsymbol{\gamma}}_i^{*\top},$$

where the matrix $\widehat{\mathbb{V}}_{in}$ is an estimate of the covariance matrix under the larger model restricted to parameters $\boldsymbol{\gamma}_i^*$ and $\hat{\boldsymbol{\gamma}}_i^*$ are estimates of the item parameters $\boldsymbol{\gamma}_i^*$ under the larger model. If H_{i0} holds, then the W_i statistic has an asymptotic χ^2 distribution:

$$W_i \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(\mathrm{df}_{i1} - \mathrm{df}_{i0}). \tag{1.28}$$

1.4 Implementation

In this part, we discuss an implementation of different estimation methods considered in Section 1.2 into the statistical software **R** (R Core Team, 2020). First, we introduce an **R** package – **difNLR** (Hladká & Martinková, 2020). This package offers an estimation using the nonlinear least squares described in Section 1.2.1 and the maximum likelihood method showed in Section 1.2.2. It provides DIF detection based on the F-test (1.26), the likelihood-ratio test (1.27), and the Wald test (1.28) including several features, from data generation to a graphical representation of the results. We also present an interactive implementation of the method within the **ShinyItemAnalysis** package and an online application (Martinková & Drabinová, 2018) which offers some functionalities of the **difNLR** package in Section 1.4.2. Then, we show an implementation of the EM algorithm covered by Section 1.2.3 and also the algorithm based on parametric link function described in Section 1.2.4.

1.4.1 The difNLR package

In this part, we discuss an implementation of the nonlinear model (1.1) for DIF detection into the **difNLR** package. This section has been adapted from Hladká and Martinková (2020) and it also comprises some new functionalities implemented in the newest version of the package – version 1.3.7. The newest development version can be downloaded from the GitHub repository using the **devtools** package (Wickham, Hester, & Chang, 2020) and the following commands:

```
devtools::install_github("adelahladka/difNLR")
library(difNLR)
```

The nonlinear model (1.1) can be fitted via the `difNLR()` function which offers a wide range of functionalities for DIF detection among dichotomous data. The full syntax of the `difNLR()` function is

```
difNLR(
  Data, group, focal.name, model, constraints, type = "all",
  method = "nls", match = "zscore", anchor = NULL, purify = FALSE,
  nrIter = 10, test = "LR", alpha = 0.05, p.adjust.method = "none",
  start, initboot = TRUE, nrBo = 20
)
```

Description of the arguments of the function can be found in Table A.1. To detect DIF using the `difNLR()` function, the user always needs to provide four pieces of information: 1. the binary data set, 2. the group membership vector, 3. the indication of the focal group, and 4. the model.

Data. `Data` is a `matrix` or a `data.frame` with rows representing dichotomously scored respondents' answers (1 correct, 0 incorrect) and columns which correspond to the items. In addition, `Data` may contain the vector of a group membership variable. If so, the `group` is a column identifier of the `Data`. Otherwise, the `group` must be a dichotomous vector of the same length as the number of rows (respondents) in `Data`. The name of the focal group is specified in the `focal.name` argument.

Data generation. To run a simulation study or to create an illustrative example, the **`difNLR`** package contains a data generator `genNLR()`, which can be used to generate dichotomous, ordinal, or nominal data. The type of items to be generated can be specified via the `itemtype` argument: `itemtype = "dich"` for dichotomous items, `"ordinal"` for ordinal items, and `"nominal"` for nominal items.

For the generation of the dichotomous items, discrimination and difficulty parameters need to be specified within the `a` and `b` arguments in the form of matrices with the two columns. The first column stands for the reference group and the second one for the focal group. Each row of matrices corresponds to one item. Additionally, one can provide guessing and inattention parameters via the arguments `c` and `d` in the same way as for the discriminations and difficulties. By default, values of the guessing parameters are set to 0 in both groups, and the values of the inattention parameters to 1 in both groups.

Distribution of the underlying latent trait is considered to be Gaussian. The user can specify its mean and standard deviation via arguments `mu` and `sigma` respectively. By default, mean is 0 and standard deviation is 1 and they are the same for both groups.

Furthermore, the user needs to provide a sample size (`N`) and the ratio of respondents in the reference and focal group (`ratio`). The latent trait for both groups is then generated and together with the item parameters it is used to generate item data. Output of the `genNLR()` function is a `data.frame` with items represented by columns and responses to them represented by rows. The last column is a group indicator, where 0 stands for a focal group and 1 indicates a reference group.

To illustrate generation of the dichotomously scored items and to exemplify basic DIF detection with the `difNLR()` function, we create an example dataset. We choose discrimination a, difficulty b, guessing c, and inattention d parameters for 15 items. Parameters are then set the same for both groups.

```
# discrimination
a <- matrix(rep(c(1.00, 1.12, 1.45, 1.25, 1.32, 1.38, 1.44, 0.89, 1.15,
                  1.30, 1.29, 1.46, 1.16, 1.26, 0.98), 2),
            ncol = 2)
# difficulty
b <- matrix(rep(c(1.34, 0.06, 1.62, 0.24, -1.45, -0.10, 1.76, 1.96,
                  -1.53, -0.44, -1.67, 1.91, 1.62, 1.79, -0.21), 2),
            ncol = 2)
# guessing
c <- matrix(rep(c(0.00, 0.00, 0.00, 0.00, 0.00, 0.17, 0.18, 0.05, 0.10,
                  0.11, 0.15, 0.20, 0.21, 0.23, 0.24), 2),
            ncol = 2)
# inattention
d <- matrix(rep(c(1.00, 1.00, 1.00, 0.92, 0.87, 1.00, 1.00, 0.88, 0.93,
                  0.94, 0.81, 0.98, 0.87, 0.96, 0.85), 2),
            ncol = 2)
```

For items 5, 8, 11, and 15 we introduce DIF caused by various sources: In item 5, DIF is caused by a difference in difficulty; in item 8 by discrimination; in item 11, the reference and focal groups differ in inattention, and in item 15 in guessing.

```
b[5, 2] <- b[5, 2] + 1
a[8, 2] <- a[8, 2] + 1
d[11, 2] <- 1
c[15, 2] <- 0
```

We generate dichotomous data with 500 observations in the reference group and 500 in the focal group. We assume that the underlying latent trait comes from a standard normal distribution for both groups (default setting). The output is a `data.frame` where the first 15 columns are dichotomously scored answers of 1,000 respondents and the last column is a group membership variable.

```
set.seed(42)
df <- genNLR(N = 1000, a = a, b = b, c = c, d = d)
head(df[, c(1:5, 16)])
  Item1 Item2 Item3 Item4 Item5 group
1     0     1     1     1     1     0
2     0     1     1     0     1     0
3     0     1     0     0     1     0
4     1     1     1     0     1     0
5     1     1     0     1     1     0
6     0     1     0     0     1     0

DataDIF <- df[, 1:15]
groupDIF <- df[, 16]
```

Model. The last necessary input of the `difNLR()` function is a specification of the model to be estimated. This can be made by the `model` argument. There are several predefined models, all of them based on the 4PL non-IRT model stated in equation (1.1) (see Table 1.1).

Table 1.1: Predefined models for the `model` argument in the `difNLR()` function.

Model annotation	Description
`"4PL"`	4PL model
`"4PLcdg"`, `"4PLc"`	4PL model with an inattention parameter set equal for the two groups
`"4PLcgd"`, `"4PLd"`	4PL model with a guessing parameter set equal for the two groups
`"4PLcgdg"`	4PL model with a guessing and an inattention parameters set equal for the two groups
`"3PLd"`	3PL model with an inattention parameter and $c = 0$
`"3PLc"`, `"3PL"`	3PL model with a guessing parameter and $d = 1$
`"3PLdg"`	3PL model with an inattention parameter set equal for the two groups
`"3PLcg"`	3PL model with a guessing parameter set equal for the two groups
`"2PL"`	Logistic regression model, i.e., $c = 0$ and $d = 1$
`"1PL"`	1PL model with a discrimination parameter set equal for the two groups
`"Rasch"`	1PL model with a discrimination parameter fixed on value 1 for the two groups

We are now able to perform the basic DIF detection with the 4PL model for all the items on a generated example dataset `DataDIF`.

```
(fit1 <- difNLR(DataDIF, groupDIF, focal.name = 1, model = "4PL"))
Detection of all types of differential item functioning
using generalized logistic regression model

Generalized logistic regression likelihood ratio chi-square statistics
based on 4PL model

Parameters were estimated with nonlinear least squares

Item purification was not applied
No p-value adjustment for multiple comparisons

       Chisq-value P-value
Item1   6.2044      0.1844
Item2   0.2802      0.9911
Item3   2.7038      0.6086
Item4   5.8271      0.2124
Item5  48.0052      0.0000 ***
Item6   7.2060      0.1254
Item7   3.2390      0.5187
```

```
Item8  16.8991      0.0020 **
Item9   2.1595      0.7064
Item10  4.6866      0.3210
Item11 69.5328      0.0000 ***
Item12  8.1931      0.0848 .
Item13  2.5850      0.6295
Item14  2.9478      0.5666
Item15 20.6589      0.0004 ***

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Detection thresholds: 9.4877 (significance level: 0.05)

Items detected as DIF items:
 Item5
 Item8
 Item11
 Item15
```

The output returns values of the test statistics (likelihood ratio test is used as a default option, see below) for DIF detection, corresponding *p*-values, and set of items which are detected as functioning differently. All items (5, 8, 11, and 15) are correctly identified.

The `difNLR()` function offers two techniques to estimate item parameters of the generalized logistic regression model (1.1). With a default option `method = "nls"`, the nonlinear least square estimation, as described in Section 1.2.1, is applied using a `nls()` function from the **stats** package (R Core Team, 2020). With an option `method = "likelihood"`, the maximum likelihood method, as described in Section 1.2.2, is used via an `optim()` function again from the **stats** package. Moreover, the user can specify what test of a submodel should be used to test for DIF. Options for testing are either a likelihood-ratio test (1.27) (`test = "LR"`, default), Wald test (1.28) (`test = "W"`), or an F-test (1.26) (`test = "F"`).

Estimates of the item parameters can be viewed with the `coef()` method. Method `coef()` returns a list of parameters, which can be simplified to a matrix by setting `simplify = TRUE`. Each row then corresponds to one item and columns indicate parameters of the estimated model.

```
round(coef(fit1, simplify = TRUE), 3)
           a      b     c     d   aDif   bDif  cDif   dDif
Item1  1.484  1.294 0.049 1.000  0.000  0.000 0.000  0.000
Item2  1.176  0.153 0.000 1.000  0.000  0.000 0.000  0.000
Item3  1.281  1.766 0.001 1.000  0.000  0.000 0.000  0.000
Item4  1.450  0.421 0.000 1.000  0.000  0.000 0.000  0.000
Item5  1.965 -1.147 0.000 0.868 -0.408  0.769 0.023 -0.006
Item6  1.458 -0.527 0.000 0.954  0.000  0.000 0.000  0.000
Item7  0.888  1.392 0.000 1.000  0.000  0.000 0.000  0.000
Item8  1.162  1.407 0.000 0.866 -0.117  0.974 0.007  0.134
Item9  1.482 -1.337 0.000 0.928  0.000  0.000 0.000  0.000
Item10 1.375 -0.570 0.007 0.967  0.000  0.000 0.000  0.000
```

```
Item11 1.071 -1.027 0.000 0.969  1.173 -0.499 0.000  0.011
Item12 1.051  1.560 0.080 1.000  0.000  0.000 0.000  0.000
Item13 1.009  1.348 0.084 1.000  0.000  0.000 0.000  0.000
Item14 1.093  1.659 0.141 1.000  0.000  0.000 0.000  0.000
Item15 0.875 -0.565 0.000 0.945  0.205  0.348 0.000 -0.142
```

The user can also print standard errors of the estimates using an option `SE = TRUE`. With the nonlinear least squares estimation (`method = "nls"`), two types of standard errors are available. The first and default option is standard errors returned by the `nls()` function, i.e.,

$$\widehat{\mathbb{V}}_{in} = \widehat{\sigma}^2 \left([\nabla \mathrm{RSS}_i(\hat{\gamma}_i)]^\top [\nabla \mathrm{RSS}_i(\hat{\gamma}_i)] \right)^{-1},$$

$$\widehat{\sigma}^2 = \frac{1}{n-8} \sum_{p=1}^{n} \left(Y_{pi} - \widehat{Y}_{pi} \right)^2 .$$

For example, estimated difference in difficulty between the reference and the focal groups in item 5 is 0.769 with standard error of 0.483.

```
round(coef(fit1, SE = TRUE)[[5]], 3)
              a      b     c     d   aDif  bDif  cDif   dDif
estimate 1.965 -1.147 0.000 0.868 -0.408 0.769 0.023 -0.006
SE       0.844  0.404 0.307 0.044  1.045 0.483 0.345  0.093
```

The second option is the sandwich estimator (1.15) available via an argument `sandwich = TRUE` in the `difNLR()` function.

```
fit1_sandwich <- difNLR(DataDIF, groupDIF, focal.name = 1,
                        model = "4PL", type = "all", sandwich = TRUE)
round(coef(fit1_sandwich, SE = TRUE)[[5]], 3)
              a      b     c     d   aDif  bDif  cDif   dDif
estimate 1.965 -1.147 0.000 0.868 -0.408 0.769 0.023 -0.006
SE       1.146  0.300 0.231 0.058  1.267 0.426 0.273  0.083
```

The `difNLR()` function provides a visual representation of the ICCs using the **`ggplot2`** package (Wickham, 2016) and its graphical environment. Curves are always based on the results of a DIF detection procedure – when an item displays DIF, two curves are plotted, one for the reference and one for the focal group. Curves are accompanied by points representing empirical probabilities, i.e., proportions of correct answers with respect to the level of matching criterion and the group membership variable. Size of the points is determined by a number of respondents at this ability level. The ICCs may simply be rendered with the method `plot()` and by specifying items to be plotted. We show here the ICCs for the DIF items only (Figure 1.4).

```
plot(fit1, item = fit1$DIFitems)
```

Besides predefined models (Table 1.1), all parameters of the model can be further constrained using the argument `constraints` specifying which parameters should be set equally for the two groups. For example, choice `"ac"` in the 4PL

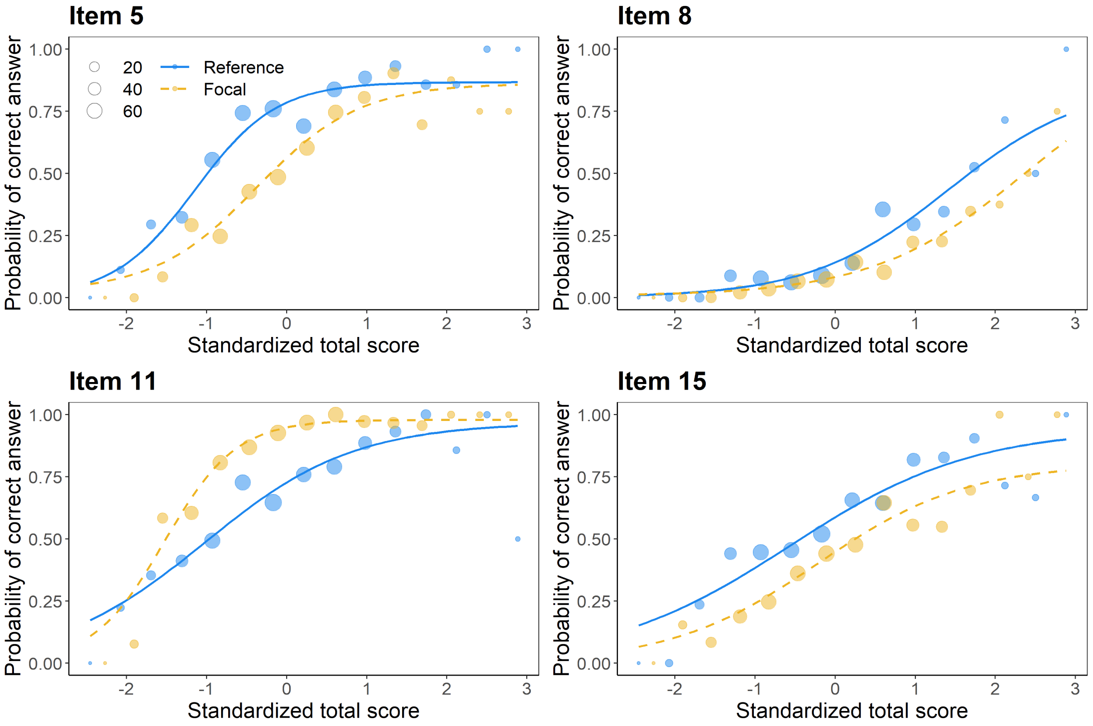


Figure 1.4: The ICC of the DIF items by the **difNLR R** package with an ability being estimated by the standardized total score.

model means that the discrimination parameter a and the pseudo-guessing parameter c are set equally for the two groups while the remaining parameters (b and d) are not. In addition, both arguments `model` and `constraints` are item-specific, meaning that a single value for all items can be introduced as well as a vector specifying the setting for each item. While the model specification can be challenging, this offers a wide range of models for DIF detection which goes hand in hand with the complexity of the offered method.

Furthermore, via `type` argument one can specify which type of DIF to test. Default option `type = "all"` allows one to test the difference in any parameter which is not constrained to be the same for both groups. Uniform DIF (difference in difficulty b only) can be tested by setting `type = "udif"`, while nonuniform DIF (difference in discrimination a, difficulty b being freely estimated for the two groups) by setting `type = "nudif"`. With the argument `type = "both"`, the differences in both parameters (a and b) are tested. Moreover, to identify DIF in more detail, one can determine in which parameter the difference should be tested. The argument `type` is also item-specific.

```
# item-specific model
model <- c("1PL", rep("2PL", 2), rep("3PL", 2),
           rep("3PLd", 2), rep("4PL", 8))
fit2 <- difNLR(DataDIF, groupDIF, focal.name = 1, model = model,
               type = "all")
fit2$DIFitems
[1]  5  8 11 15

# item-specific type
```

```
type <- rep("all", 15)
type[5] <- "b"; type[8] <- "a"; type[11] <- "c"; type[15] <- "d"
fit3 <- difNLR(DataDIF, groupDIF, focal.name = 1, model = model,
               type = type)
fit3$DIFitems
[1] 5

# item-specific constraints
constraints <- rep(NA, 15)
constraints[5] <- "ac"; constraints[8] <- "bcd";
constraints[11] <- "abd"; constraints[15] <- "abc"
fit4 <- difNLR(DataDIF, groupDIF, focal.name = 1, model = model,
               constraints = constraints, type = type)
fit4$DIFitems
[1]  5  8 11 15
```

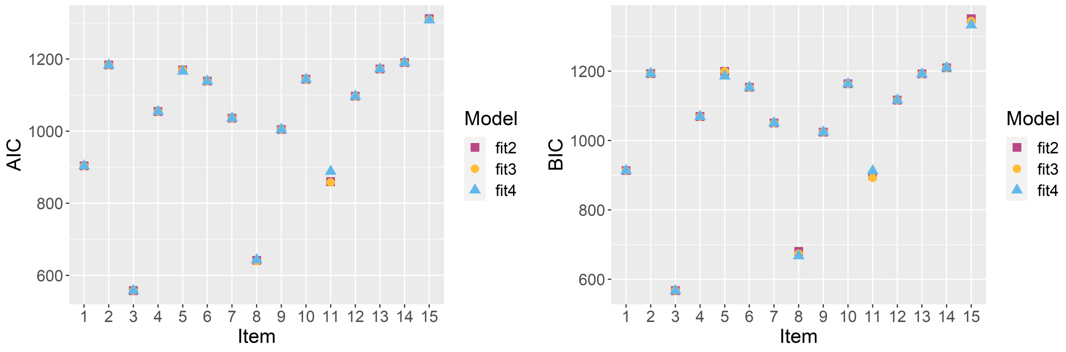
In `fit2` we allowed different models for items. In `fit3`, when items were intended to function differently, we tested only the difference in those parameters which were selected to be a source of DIF when we generated data, while using the same item-specific models as for `fit2`. Finally, in items which were intended to function differently we constrained all other parameters to be the same for both groups in `fit4`. As expected, models `fit2` and `fit4` correctly identified all DIF items, while `fit3` detected only item 5.

Fit of the selected models can be examined using so called information criteria, specifically Akaike's Information Criterion (AIC) (Akaike, 1974) and Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978). Information criteria for the best fitting model resulting from the DIF detection can be computed using methods `AIC()` and `BIC()`. We plot both criteria for all items using the **`ggplot2`** package and its function `ggplot()` (Wickham, 2016), see Figure 1.5.

```
df <- data.frame(AIC = c(AIC(fit2), AIC(fit3), AIC(fit4)),
                 BIC = c(BIC(fit2), BIC(fit3), BIC(fit4)),
                 Model = paste0("fit", rep(2:4, each = 15)),
                 Item = as.factor(rep(1:15, 3)))

ggplot(df, aes(x = Item, y = AIC, col = Model, shape = Model)) +
  geom_point(size = 3) +
  scale_color_manual(values = c("#cc79a7", "#ffbe33", "#4aaee8")) +
  scale_shape_manual(values = c(15, 16, 17))
ggplot(df, aes(x = Item, y = BIC, col = Model, shape = Model)) +
  geom_point(size = 3) +
  scale_color_manual(values = c("#cc79a7", "#ffbe33", "#4aaee8")) +
  scale_shape_manual(values = c(15, 16, 17))
```

While there is, not surprisingly, no difference between both information criteria of the three models for non-DIF items, a distinction may be observed in DIF items. AIC suggests that the model `fit3` fits best to items 8 and 11 and the model `fit4` to items 5 and 15, while BIC indicates that for item 8 the model `fit4` is the most suitable. However the differences are small (Figure 1.5). Fit measures can also be displayed for specific items.



(A) AIC.

(B) BIC.

Figure 1.5: Information criteria for item models.

```
logLik(fit3, item = 8)
'log Lik.' -312.7227 (df=7)
logLik(fit4, item = 8)
'log Lik.' -316.4998 (df=5)
```

Fitted values and residuals can be calculated with the methods `fitted()` and `residuals()`, again for all items or for those specified via the `item` argument. This also holds for predicted values and the method `predict()`. Predictions for any new respondents can be obtained by the `group` and `match` arguments representing the group membership and the value of the matching criterion (e.g., standardized total score) of the new respondent. For example, with `fit1` in item 5, new respondents with an average performance (`match = 0`) have approximately a 22% lower probability of a correct answer if they come from the focal rather than the reference group.

```
predict(fit1, item = 5, group = c(0, 1), match = 0)
   item match group      prob
1 Item5     0     0 0.7851739
2 Item5     0     1 0.5624883
```

This can also be observed when comparing the ICCs for the reference and the focal group in item 5 (see upper left Figure 1.4).

Further features

The **difNLR** covers user-friendly features that are common in standard DIF software – various matching criteria, anchor items, item purification, and *p*-value adjustments.

Matching criterion. By default, the underlying latent trait is estimated as a standardized total score in the `difNLR()` function. However, this estimate can be changed using the `match` argument. Besides default option `"zscore"` (standardized total score), it can also be the total test score (`match = "score"`) or any numeric vector of the same length as the number of respondents. It is hence possible to use, for instance, latent trait estimates provided by some IRT models, or to use a pre-test score instead of the total score of the current test to be examined.

Anchor items and item purification. Including DIF items into the calculation of the matching criterion can lead to a potential bias and misidentification of DIF and non-DIF items (see also Chapter 4). With an argument `anchor`, one can specify which items are supposed to be used for the calculation of the matching criterion.

In the following examples, for illustration purposes, we take only items 1–6 of the `DataDIF` dataset and we apply some features with the 4PL model. The matching criterion is now calculated as a total test score based on items 1–6.

We start with not specifying the anchor items. This indicates that any item can be considered as DIF one.

```
fit8a <- difNLR(DataDIF[, 1:6], groupDIF, focal.name = 1,
                match = "score", model = "4PL", type = "all")
fit8a$DIFitems
[1] 5 6
```

Initial fit `fit8a` detected items 5 a 6 as functioning differently. Now we can set all items excluding these two as the anchors.

```
fit8b <- difNLR(DataDIF[, 1:6], groupDIF, focal.name = 1,
                match = "score", model = "4PL", type = "all",
                anchor = 1:4)
fit8b$DIFitems
[1] 5
```

With a test score based only on DIF-free items 1–4 (i.e., excluding potentially unfair items 5 and 6 detected in previous run from calculation of the total score), we detected only item 5 as functioning differently. We could again fit the model excluding only item 5 from the calculation of the matching criterion.

The process of including and omitting DIF and potentially unfair items could be demanding and time consuming. However, this process can be applied iteratively and automatically. This procedure is called *item purification* (Lord, 1980; Marco, 1977, see also Section 4.1.1) and it has been shown that it can improve DIF detection. Item purification can be accessed with a `purify` argument. This can only be done when the matching criterion is either the total score or Z-score. The maximal number of iterations is determined by the `nrIter` argument, where the default value is 10.

```
fit9 <- difNLR(DataDIF[, 1:6], groupDIF, focal.name = 1,
               match = "score", model = "4PL", type = "all",
               purify = TRUE)
```

Item purification was run with 2 iterations plus one initial step. The process of including and excluding items into the calculation of the matching criterion can be found in the `difPur` element of the output.

```
fit9$difPur
      Item1 Item2 Item3 Item4 Item5 Item6
Step0     0     0     0     0     1     1
Step1     0     0     0     0     1     0
Step2     0     0     0     0     1     0
```

In the initial step, items 5 and 6 were identified as DIF as it was shown with `fit8a`. The matching criterion was then calculated as the sum of the correct answers in items 1–4 as demonstrated by `fit8b`. In the next step, only item 5 was identified as DIF and the matching criterion was based on items 1–4 and 6. The result of the DIF detection procedure was the same in the next step and the item purification process thus ended.

Multiple comparison corrections. As the DIF detection procedure is done item by item, corrections for multiple comparisons may be considered (Kim & Oshima, 2013, see also Section 4.1.2). For example, applying Holm's adjustment (Holm, 1979) results in item 5 being detected as DIF.

```
fit10 <- difNLR(DataDIF[, 1:6], groupDIF, focal.name = 1,
                match = "score", model = "4PL", type = "all",
                p.adjust.method = "holm")
fit10$DIFitems
[1] 5
```

And of course, item purification and multiple comparison corrections can be combined in a way that the p-value adjustment is applied for a final run of the item purification.

```
fit11 <- difNLR(DataDIF[, 1:6], groupDIF, focal.name = 1,
                match = "score", model = "4PL", type = "all",
                p.adjust.method = "holm", purify = TRUE)
fit11$DIFitems
[1] 5
```

While all three approaches correctly identify item 5 as a DIF item, the significance level varies:

```
round(fit9$pval, 3)
[1] 0.144 0.974 0.244 0.507 0.000 0.126
round(fit10$adj.pval, 3)
[1] 1.000 1.000 1.000 0.747 0.000 0.137
round(fit11$adj.pval, 3)
[1] 0.629 1.000 0.733 1.000 0.000 0.629
```

Troubleshooting

In this part, we focus on several issues which can be encountered when fitting the generalized logistic regression models and using the features offered in the **difNLR** package.

Convergence issues. First, there is no guarantee that the estimation process in the `difNLR()` function will always end successfully. For instance, in the case of a small sample size, convergence issues may appear.

The easiest way to fix such issues is to specify different starting values. Various starting values can be applied via a `start` argument as a list with the named numeric vectors as its elements. Each element needs to include values for the parameters `a`, `b`, `c`, and `d` of the reference group and the differences between the reference and focal groups denoted by `aDif`, `bDif`, `cDif`, and `dDif`. However, there is no need to determine initial values manually. In the instance of convergence issues, the initial values are by default automatically re-calculated based on bootstrapped samples and applied only to models that failed to converge. This is also performed when starting values were initially introduced via a `start` argument. This feature can be turned off by setting `initboot = FALSE`. In such a case, no estimates are obtained for items that failed to converge. To demonstrate described situations, we now use a sample of our original simulated data set.

```
# sampled data
set.seed(42)
sam <- sample(1:1000, 420)
# using re-calculation of starting values
fit12a <- difNLR(DataDIF[sam, ], groupDIF[sam], focal.name = 1,
                 model = "4PL", type = "all")
Starting values were calculated based on bootstraped samples.

# turn off option of re-calculating starting values
fit12b <- difNLR(DataDIF[sam, ], groupDIF[sam], focal.name = 1,
                 model = "4PL", type = "all", initboot = FALSE)
Warning message:
Convergence failure in item 3
Convergence failure in item 14
```

With an option `initboot = TRUE` in `fit12a`, starting values were re-calculated and no convergence issue occurred. When setting `initboot = FALSE` in `fit12b` we observed convergence failures in items 3 and 14.

The re-calculation process is by default performed up to twenty times, but the number of runs can be increased via the `nrBo` argument.

Another option is to apply the maximum likelihood method (see Section 1.2.2) instead of the nonlinear least squares (Section 1.2.1) to estimate item parameters.

```
fit13 <- difNLR(DataDIF[sam, ], groupDIF[sam], focal.name = 1,
                model = "4PL", type = "all", method = "likelihood")
```

There is no convergence issue in `fit13` using the maximum likelihood estimation in contrast to `fit12b` and the nonlinear least squares option.

Item purification. Issues may also occur when applying an item purification. Although this is rare in practice, there is no guarantee that the process will end successfully. This can be observed, for instance, when we use the `DataDIF` dataset with the first 12 items only.

```
fit14 <- difNLR(DataDIF[, 1:12], groupDIF, focal.name = 1,
               model = "4PL", type = "all", purify = TRUE)
Warning message:
Item purification process not converged after 10 iterations.
Results are based on the last iteration of the item purification.
```

The maximum number of item purification iterations can be increased using the `nrIter` argument. However, in our example this would not necessarily lead to success as the process was not able to decide whether or not to include item 1 in the calculation of the matching criterion.

```
fit14$difPur
          I1   I2   I3   I4   I5   I6   I7   I8   I9  I10  I11  I12
Step0      0    0    0    0    1    0    0    1    0    0    1    0
Step1      1    0    0    0    1    0    0    1    0    0    1    0
Step2      0    0    0    0    1    0    0    1    0    0    1    0
Step3      1    0    0    0    1    0    0    1    0    0    1    0
Step4      0    0    0    0    1    0    0    1    0    0    1    0
Step5      1    0    0    0    1    0    0    1    0    0    1    0
Step6      0    0    0    0    1    0    0    1    0    0    1    0
Step7      1    0    0    0    1    0    0    1    0    0    1    0
Step8      0    0    0    0    1    0    0    1    0    0    1    0
Step9      1    0    0    0    1    0    0    1    0    0    1    0
Step10     0    0    0    0    1    0    0    1    0    0    1    0
```

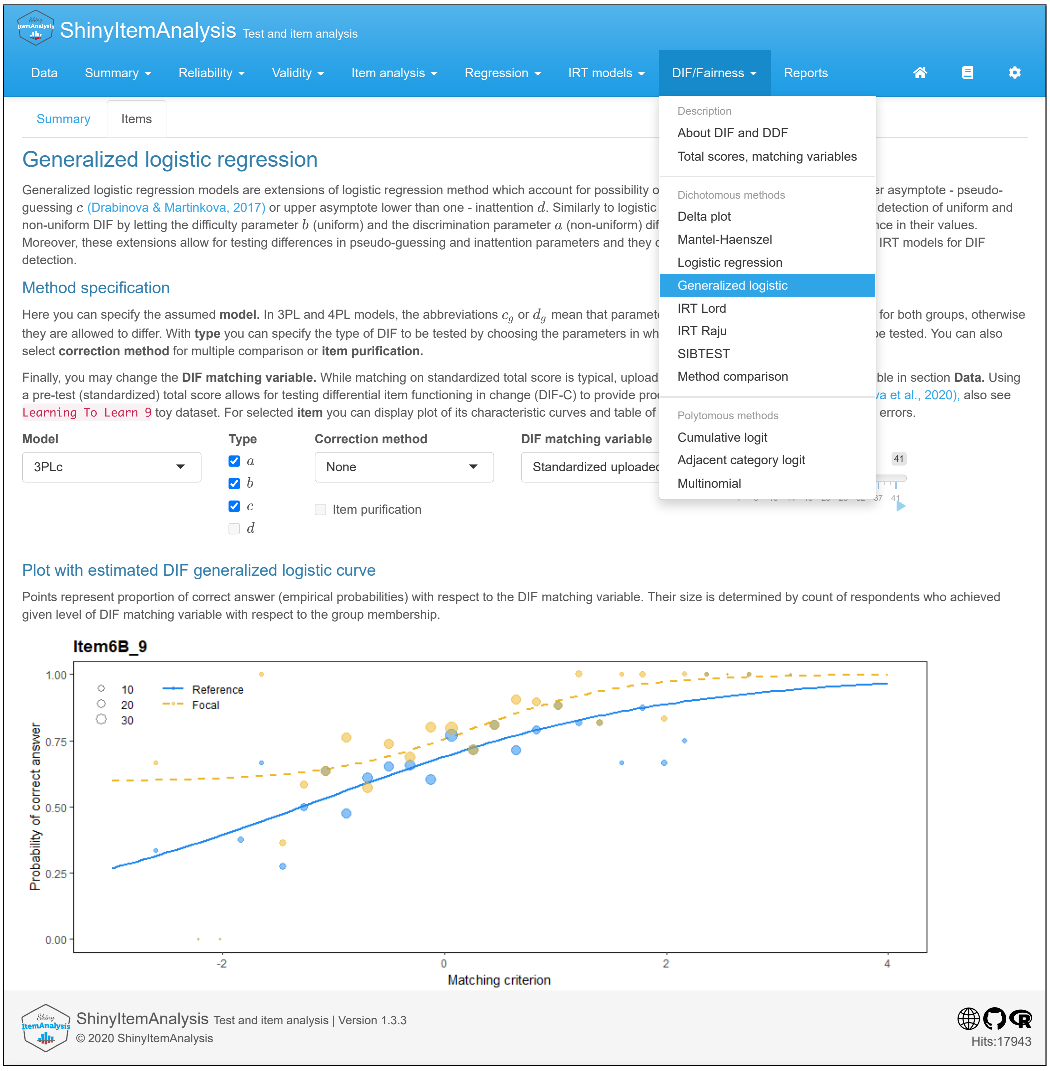
In this context, we advise considering such items as DIF to be on the safe side. As a general rule, any suspicious item should be reviewed by content experts. Not every DIF item is necessarily unfair, however even in such a case, understanding the reasons behind DIF may inform educators and help provide the best assessment and learning experience to all individuals involved.

In summary, the **`difNLR`** package provides two estimation methods for generalized logistic model (1.1) (see Sections 1.2.1 and 1.2.2), three test statistics (see Section 1.3), and number of features such as item purification or corrections for multiple comparisons (see Chapter 4).

1.4.2 The `ShinyItemAnalysis` package and application

Some functionalities of the **`difNLR`** package are exploited by another **`R`** package and interactive online application – **`ShinyItemAnalysis`** (Martinková & Drabinová, 2018). This includes various options of the model selection, matching criteria, parameters to be tested, and also further features such as multiple comparison corrections or item purification. Application provides summary table with the likelihood ratio test statistics and corresponding p-values while the plot of the ICCs may be displayed for each item (Figure 1.6).

The **`ShinyItemAnalysis`** further offers, besides fairness analysis, wide range of psychometric methods including traditional item analysis, regression models, and IRT models. Software also provides possibility to upload user's data and generate reports. Moreover, it offers equations of the models, parameter estimates with their interpretation, and selected **R** code and therefore it can serve as a springboard for more detailed analysis within the **difNLR** package.



Generalized logistic regression

Generalized logistic regression models are extensions of logistic regression method which account for possibility o … er asymptote - pseudo-guessing c (Drabinova & Martinkova, 2017) or upper asymptote lower than one - inattention d. Similarly to logistic … detection of uniform and non-uniform DIF by letting the difficulty parameter b (uniform) and the discrimination parameter a (non-uniform) dif … nce in their values. Moreover, these extensions allow for testing differences in pseudo-guessing and inattention parameters and they c … IRT models for DIF detection.

Method specification

Here you can specify the assumed **model.** In 3PL and 4PL models, the abbreviations c_g or d_g mean that paramete … for both groups, otherwise they are allowed to differ. With **type** you can specify the type of DIF to be tested by choosing the parameters in wh … e tested. You can also select **correction method** for multiple comparison or **item purification.**

Finally, you may change the **DIF matching variable.** While matching on standardized total score is typical, upload … ble in section **Data.** Using a pre-test (standardized) total score allows for testing differential item functioning in change (DIF-C) to provide pro … a et al., 2020), also see `Learning To Learn 9` toy dataset. For selected **item** you can display plot of its characteristic curves and table of … errors.

Model: 3PLc

Type: ☑ a ☑ b ☑ c □ d

Correction method: None

□ Item purification

DIF matching variable: Standardized uploaded

Plot with estimated DIF generalized logistic curve

Points represent proportion of correct answer (empirical probabilities) with respect to the DIF matching variable. Their size is determined by count of respondents who achieved given level of DIF matching variable with respect to the group membership.

Figure 1.6: Generalized logistic model for DIF detection (1.1) implemented in the interactive **ShinyItemAnalysis** application.

1.4.3 EM algorithm

In this part we show an implementation of the EM algorithm described in Section 1.2.3 to the statistical software **R** (R Core Team, 2020). For the illustrative purposes we generate responses to one binary item with the parameters $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 0.5$, $c = 0.2$, $c_{\text{DIF}} = 0.1$, $d = 1$, and $d_{\text{DIF}} = -0.1$ based on group membership variable `g` and observed ability `x` which has standard normal distribution.

```
set.seed(42)
x <- rnorm(1000)
g <- rep(c(0, 1), each = 500)
```

```
p <- 0.2 + 0.1 * g + (1 - 0.1 * g - 0.2 - 0.1 * g) /
  (1 + exp(0 - x + 1 * g - 0.5 * x * g))
y <- rbinom(1000, 1, p)
```

To run E-step of the EM algorithm, we first need to set initial values of the item parameters. Here we used classical intercept-slope parametrization for the calculation purposes. For illustration, we set fixed initial values to be close to true parameters:

```
b0_new <- 0.1
b1_new <- 0.85
b2_new <- -1.1
b3_new <- 0.6
c_new <- 0.15
cDif_new <- 0.15
d_new <- 0.95
dDif_new <- -0.05
```

The implementation of the E-step via function `expectation()` is a straightforward application of the formulae for estimates of the latent variables $\widehat{Z}_{pi1}$, $\widehat{Z}_{pi2}$, $\widehat{Z}_{pi3}$, and $\widehat{Z}_{pi4}$:

```
# Expectation step
# Calculates estimates of latent variables Z1, Z2, Z3, and Z4
# Arguments:
#    y = outcome (binary vector)
#    x = observed ability (numeric vector)
#    g = group membership variable (binary vector)
#    b0, b1, b2, b3 = parameters of logistic curve without asymptotes
#    c, cDif = group specific lower asymptotes of logistic curve
#    d, dDif = group specific upper asymptotes of logistic curve
expectation <- function(y, x, g,
                        b0, b1, b2, b3,
                        c, cDif, d, dDif) {
  expit <- function(x) {
    return(exp(x)/(1 + exp(x)))
  }
  phi <- as.vector(expit(c(b0, b1, b2, b3)
                         %*% t(cbind(1, x, g, x * g))))

  z1 <- y * (c + cDif * g) /
    (c + cDif * g + (d + dDif * g - c - cDif * g) * phi)
  z2 <- y - z1
  z4 <- (1 - y) * (1 - d - dDif * g) /
    (1 - c - cDif * g - (d + dDif * g - c - cDif * g) * phi)
  z3 <- 1 - y - z4

  return(list(z1 = z1, z2 = z2, z3 = z3, z4 = z4))
}
```

The `expectation()` function returns list of estimates of four latent variables based on current estimates of the item parameters.

```
Z <- expectation(y, x, g, b0, b1, b2, b3, c, cDif, d, dDif)
lapply(Z, summary)
$z1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.1933  0.2266  0.3824  0.9945

$z2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.1785  0.3094  0.6560  0.8325

$z3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3905  0.8470  0.9364

$z4
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.07346 0.14896 0.87161
```

Now we are ready to proceed with the M-step to estimate the item parameters. As was noted earlier, the corresponding log-likelihood function (1.22) (with ϕ_{pi} defined using the intercept-slope parametrization) has a form of the logistic regression, which makes parameters β_{i0}, β_{i1}, β_{i2}, and β_{i3} easy to estimate using a standard **R** function `glm()` with a quasi-binomial family (to account for the fact that $Z_{pi2} + Z_{pi3} \leq 1$):

```
(fit1 <- glm(cbind(Z$z2, Z$z3) ˜ x + g + x:g,
  family = binomial()))
Coefficients:
(Intercept)            x            g          x:g
     0.1619       0.9193      -1.1614       0.5749

Degrees of Freedom: 999 Total (i.e. Null);  996 Residual
Null Deviance:          960.9
Residual Deviance: 778.9        AIC: 855.4
```

As the second part of the log-likelihood function (1.23) has a form of the multinomial regression, specialized package to fit such model may be applied. Here we use the function `multinom()` from the **nnet** package (Venables & Ripley, 2002).

```
library(nnet)
(fit2 <- multinom(cbind(Z$z2 + Z$z3, z$z1, z$z4) ˜ g))
Coefficients:
  (Intercept)         g
2   -1.653768 0.9611856
3   -2.830450 1.0360393

Residual Deviance: 1507.611
AIC: 1515.611
```

To calculate estimated probabilities of the $Z_{pij}, j = 1, \ldots, 4$, and therefore asymptote parameters, one can extract fitted values from the model. It should be noted that the latent variable Z_{pi4} accounts for the probability of answer to be not missed due inattention, therefore, related parameter is subtracted from 1. The estimates of the asymptote parameters are then $c_i = 0.153$, $c_{i\mathrm{DIF}} = 0.147$, $d_i = 0.953$, and $d_{i\mathrm{DIF}} = -0.053$.

```
# extracting probabilities per groups
par_asympt <- as.data.frame(unique(cbind(g, fitted(fit2))))
# calculating upper asymptotes for the two groups
par_asympt$V4 <- 1 - par_asympt$V4
# differences in parameters between focal and reference group
par_asympt[3, ] <- par_asympt[par_asympt$g == 1, ] -
  par_asympt[par_asympt$g == 0, ]
par_asympt[c(1, 3), c(1, 3, 4)]
  g        V3          V4
1 0 0.1530237  0.95282282
3 1 0.1471744 -0.05256752
```

The E-step and M-step are repeated till the convergence criterion is met or till the maximum number of iteration is reached. The iterations converge when

$$\frac{\left|\mathrm{dev}_{i(j)}^{\mathrm{EM}} - \mathrm{dev}_{i(j-1)}^{\mathrm{EM}}\right|}{\left|\mathrm{dev}_{i(j)}^{\mathrm{EM}}\right| + 0.1} < \epsilon,$$

where $\mathrm{dev}_{i(j)}^{\mathrm{EM}}$ is a deviance of the model for item i in the j-th iteration and ϵ is a positive convergence tolerance.

In our illustrative example we set $\epsilon = 10^{-6}$. After 287 iterations, we received convergence with the final estimates

```
# final parameter estimates
      b0      b1       b2      b3       c     cDif       d     dDif
-0.17602 0.98603 -0.87166 0.85062 0.21989 0.096286 0.99945 -0.14934
```

which are close to the true parameters considering limited sample size. Standard errors of the final parameter estimates can be calculated using the inverse of the Hessian matrix implemented within the `covariance.matrix()` function:

```
# Covariance matrix
# Calculates estimate of covariance matrix
# Arguments:
#    y = outcome (binary vector)
#    x = observed ability (numeric vector)
#    g = group membership variable (binary vector)
#    par = vector of parameters of the nonlinear model:
#          b0, b1, b2, b3, c, cDif, d, dDif
covariance.matrix <- function(x, y, g, par) {
  # formula for log-likelihood
  f <- "y * log(c + cDif * g + (d + dDif * g - c - cDif * g) /
  (1 + exp(- (b0 + b1 * x + b2 * g + b3 * x * g)))) +
  (1 - y) * log(1 - (c + cDif * g + (d + dDif * g - c - cDif * g) /
  (1 + exp(- (b0 + b1 * x + b2 * g + b3 * x * g)))))"
  # calculating Hessian
  hess <- hessian(
    f = f,
    var = c("b0", "b1", "b2", "b3", "c", "cDif", "d", "dDif")
  )
  # Hessian as a function of data and parameters
  hess.fun <- eval(
    parse(text = paste0(
      "function(",
      paste(c("y", "x", "g",
              "b0", "b1", "b2", "b3", "c", "cDif", "d", "dDif"),
            collapse = ", "), ") {
      return(list(", paste(as.list(hess),
            collapse = ", "), "))}"
    ))
  )
  # evaluating Hessian function and creating matrix
  n <- length(x)
  return(solve(-n * matrix(
    sapply(
      do.call(
        hess.fun,
        append(list(y = y, x = x, g = g), par)
      ),
      mean
    ),
    ncol = 8, nrow = 8
  )))
}
```

For the final estimates we get:

```
# standard errors of the estimates
sqrt(diag(covariance.matrix(x, y, g, par[nrow(par), ])))
[1] 0.79370 0.44492 1.00681 0.91258 0.20527 0.21341 0.16233 0.19429
```

The full **R** script for the EM algorithm can be found in Appendix A.3.

1.4.4 Parametric link function

In this part we show an implementation of the algorithm based on parametric link function described in Section 1.2.4 to the statistical software **R** (R Core Team, 2020). The crucial part of the implementation covers specification of the parametric link function:

```
# specification of parametric logit link
plogit <- function(c, cDif, d, dDif, g) {
  cp <- c + cDif * g
  dp <- d + dDif * g
  logitint <- function(p, cp, dp){
    log(ifelse((p - cp)/(dp - p) <= 0, 0.00001, (p - cp) / (dp - p)))
  }

  # link function
  linkfun <- function(mu) logitint(mu, cp, dp)

  # the inverse of the link function
  linkinv <- function(eta)
    cp + (dp - cp) * exp(eta) / (1 + exp(eta))

  # derivative of the inverse-link function with respect
  # to the linear predictor
  mu.eta <- function(eta)
    (dp - cp) * exp(eta) / (1 + exp(eta))^2

  # TRUE if eta is in the domain of linkinv
  valideta <- function(eta) TRUE

  name <- "plogit"
  structure(list(linkfun = linkfun,
                 linkinv = linkinv,
                 valideta = valideta,
                 mu.eta = mu.eta,
                 name = name),
            class = "link-glm")
}
```

Then, we need to specify a function to compute the log-likelihood which is supposed to be maximized to estimate parameters of the asymptotes:

```
# likelihood for asymptote parameters, when parameters of the logistic
# curve are fixed
param.likel.cd <- function(theta){
  param.expit <- function(x, g, c0, c1, d0, d1){
    c0 * (1 - g) + c1 * g +
      (d0 * (1 - g) + d1 * g - c0 * (1 - g) - c1 * g) /
      (1 + exp(-x))
  }
  n <- nrow(X)
  c0 <- theta[1]
  c1 <- theta[2]
  # c1 <- c0 + cDif
  d0 <- theta[3]
  d1 <- theta[4]
  # d1 <- d0 + dDif

  h <- param.expit(X %*% c(b0_new, b1_new, b2_new, b3_new),
                   g, c0, c1, d0, d1)
  l <- -(1 / n) * sum((y * log(h)) + ((1 - y) * log(1 - h)))
  return(l)
}
```

For illustration, we use the same data as in previous section and we use also the same initial values:

```
set.seed(42)

x <- rnorm(1000)
g <- rep(c(0, 1), each = 500)
p <- 0.2 + 0.1 * g + (1 - 0.1 * g - 0.2 - 0.1 * g) /
  (1 + exp(0 - x + 1 * g - 0.5 * x* g))
y <- rbinom(1000, 1, p)

b0_new <- 0.1
b1_new <- 0.85
b2_new <- -1.1
b3_new <- 0.6
c_new <- 0.15
cDif_new <- 0.15
d_new <- 0.95
dDif_new <- -0.05
```

We are now ready to proceed with the first step of the algorithm – fitting generalized linear model with the parametric link function `plogit()`:

```
(fit_glm <- glm(y ˜ x + g + x:g,
                family = binomial(
                  link = plogit(c_new, cDif_new, d_new, dDif_new, g)
                ),
                start = c(b0_new, b1_new, b2_new, b3_new)))
Coefficients:
(Intercept)            x            g          x:g
     0.1746       0.9379      -1.2016       0.6259

Degrees of Freedom: 999 Total (i.e. Null);  996 Residual
Null Deviance:            1381
Residual Deviance: 1276         AIC: 1284

b0_old <- b0_new
b1_old <- b1_new
b2_old <- b2_new
b3_old <- b3_new

b0_new <- coef(fit_glm)[1]
b1_new <- coef(fit_glm)[2]
b2_new <- coef(fit_glm)[3]
b3_new <- coef(fit_glm)[4]
```

Using the new estimated parameters of the logistic regression curve being fixed, we proceed with the second step by fitting model to estimate asymptote parameters:

```
# bound for asymptotes
c0_max <- max(min(fitted(fit_glm)[g == 0], na.rm = TRUE), 0)
c1_max <- max(min(fitted(fit_glm)[g == 1], na.rm = TRUE), 0)
d0_min <- min(max(fitted(fit_glm)[g == 0], na.rm = TRUE), 1)
d1_min <- min(max(fitted(fit_glm)[g == 1], na.rm = TRUE), 1)

(fit_cd <- optim(fn = param.likel.cd,
                 par = setNames(
                   c((c_new + c0_max) / 2,
                     (c_new + cDif_new + c1_max) / 2,
                     (d0_min + d_new) / 2,
                     (d_new + dDif_new + d1_min) / 2),
                   c("c0", "c1", "d0", "d1")),
                 method = "L-BFGS-B",
                 lower = c(0, 0, d0_min, d1_min),
                 upper = c(c0_max, c1_max, 1, 1)))
$par
    c0     c1     d0     d1
0.1329 0.3011 0.9738 0.8930
```

Similarly as in Section 1.4.3, the two steps are repeated till the convergence criterion is met or till the maximum number of iterations is reached. In our illustrative example we again set $\epsilon = 10^{-6}$. The standard errors of the final parameter estimates can be again calculated using the `covariance.matrix()` function. After 51 iterations, we received convergence with the final estimates

```
# final parameter estimates
      b0      b1       b2      b3       c    cDif       d     dDif
-0.15238 0.97252 -0.89734 0.86781 0.21307 0.10343 1.00000 -0.15020
# standard errors of the estimates
sqrt(diag(covariance.matrix(x, y, g, par[nrow(par), ])))
[1] 0.78199 0.43279 0.99895 0.91238 0.20541 0.21339 0.16474 0.19620
```

The full **R** script for the algorithm based on parametric link function can be found in Appendix A.4.

1.5 Simulation study – properties of the method

In this part we include a simulation study evaluating properties of DIF detection procedure based on the restricted form of the nonlinear model (1.1). This section was adapted from Drabinová and Martinková (2017). Note that while we described the theoretical properties for the 4PL non-IRT model in previous sections, this early simulation study was limited to the case of inattention parameter fixed at value of 1 (i.e., $d_{iG_p} = 1$), that is

$$\mathsf{P}(Y_{pi} = 1 | X_p, G_p) = c_{iG_p} + (1 - c_{iG_p}) \frac{e^{a_{iG_p}(X_p - b_{iG_p})}}{1 + e^{a_{iG_p}(X_p - b_{iG_p})}}, \tag{1.29}$$

also termed here as the 3PL non-IRT model.

In the simulation study, we compared DIF detection method based on the restricted model (1.29) to other commonly used DIF detection approaches such as the logistic regression model (7) (Swaminathan & Rogers, 1990), the Mantel-Haenszel test (4) (Mantel & Haenszel, 1959), and the Lord's test (2) (Lord, 1980) based on the 3PL IRT model:

$$\mathsf{P}(Y_{pi} = 1|\theta_p) = c_i + (1 - c_i)\frac{e^{a_i(\theta_p - b_i)}}{1 + e^{a_i(\theta_p - b_i)}}. \tag{1.30}$$

The simulation study evaluated convergence behavior, power rate (i.e., the proportion of true positives), and rejection rate (i.e., the proportion of false positives; type I error).

1.5.1 Study design

In this part we describe design of the simulation study including data generation, DIF detection procedures, and evaluation of the results.

Data and DIF generation

The dichotomously scored data are generated with the 3PL IRT model (1.30) as follows: examinees' knowledge θ_p is assumed to follow the standard normal distribution, i.e., $\theta_p \sim \mathcal{N}(0, 1)$. All parameters are set to be the same for both reference and focal group unless the item is a DIF item, in which case the item parameters of the focal group are manipulated (see below). To reflect realistic values of item parameters and to be in line with previous simulation studies (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1996; Jodoin & Gierl, 2001; Güler & Penfield, 2009; Kim & Oshima, 2013), the simulation study is based on item parameters according to 20-item data set from the 1985 problem solving of the GMAT (Kingston, Leary, & Wightman, 1985, p. 47). Probabilities of correct answers are calculated based on true values of items and examinees parameters and dichotomous responses are then generated from Bernoulli distribution with these calculated probabilities.

Out of 20 items, one, or three first items are manipulated to perform DIF caused by difference in difficulty parameter b_{iG_p} (here referenced as uniform DIF), or in discrimination parameter a_{iG_p} (here referenced as non-uniform DIF), or in guessing parameter c_{iG_p}. The thresholds for DIF effect size of DIF items, represented by the Area Measure (AM) between the two ICCs (defined as UA in (3)), respectively by the Weighted Area Measure (WAM) in case of varying pseudo-guessing parameter when the AM is weighted by density of normal distribution (Siebert, 2013), are determined by values 0.4 (low), 0.6 (moderate), and 0.8 (large) (respectively 0.09, 0.12 and 0.14 for the WAM) following Swaminathan and Rogers (1990), Narayanan and Swaminathan (1996), and Siebert (2013). When one DIF item is considered, the large size of DIF is chosen. Mixture of DIF sizes is considered for the larger proportion of DIF items.

When uniform DIF is considered, the discrimination parameters for the focal and the reference group are kept the same and fixed at value 1. The differences in difficulty between the reference and the focal group are set to 0.5 (low), 0.75 (moderate), and 1 (large) (see Table 1.2, DIF source *b*).

When simulating non-uniform DIF, the difficulty parameters for both groups are kept the same and fixed at value 0 and the discrimination parameters are chosen according to Narayanan and Swaminathan (1996, p. 264) (see Table 1.2, DIF source *a*).

For DIF caused by varying guessing among groups, the discrimination is fixed at value 1 and the difficulty at value 0 for both groups. Guessing parameter is manipulated for both groups to achieve desired DIF size level (see Table 1.2, DIF source *c*). The parameters of remaining (19 or 17) non-DIF items are selected from the problem solving 1985 of the GMAT as reported in Kingston et al. (1985, p. 47). To evaluate rejection rates of procedures also simulations without DIF items are considered.

Table 1.2: Item parameters used to generate DIF items.

DIF source	Item	DIF effect size		Reference group			Focal group		
		AM	WAM	*a*	*b*	*c*	*a*	*b*	*c*
b	1	0.8	0.16	1.00	0	0.20	1.00	1.00	0.20
b	1	0.4	0.08	1.00	0	0.20	1.00	0.50	0.20
	2	0.6	0.12	1.00	0	0.20	1.00	0.75	0.20
	3	0.8	0.16	1.00	0	0.20	1.00	1.00	0.20
a	1	0.8	0.12	0.56	0	0.20	1.79	0	0.20
a	1	0.4	0.09	0.90	0	0.20	2.01	0	0.20
	2	0.6	0.12	0.70	0	0.20	1.97	0	0.20
	3	0.8	0.12	0.56	0	0.20	1.79	0	0.20
c	1	–	0.14	1.00	0	0.10	1.00	0	0.38
c	1	–	0.09	1.00	0	0.10	1.00	0	0.28
	2	–	0.12	1.00	0	0.10	1.00	0	0.34
	3	–	0.14	1.00	0	0.10	1.00	0	0.38

Note. AM = area measure between curves, WAM = weighted area measure between curves. The area between two curves with different *c* among groups would be infinite.

The above described scenarios are investigated on various levels of the total sample size. Larger sample sizes are determined to yield satisfactory convergence levels especially for the IRT models (Kim & Oshima, 2013; Siebert, 2013). Specifically, five levels of sample size are considered; three with the same proportion in groups: 1,000 (500 per group), 2,000 (1,000 per group) and 4,000 (2,000 per group), and two with the proportion of 1/2: 1,500 (500 reference, 1,000 focal) and 3,000 (1,000 reference, 2,000 focal) as inspired by Narayanan and Swaminathan (1996) and Jodoin and Gierl (2001).

DIF identification

Four distinct methods for DIF detection are selected: The restricted model (1.29), the logistic regression procedure (7) (Swaminathan & Rogers, 1990), the Mantel-Haenszel test (4) (Mantel & Haenszel, 1959), and the Lord's test (2) (Lord, 1980) based on the 3PL IRT model (1.30). As suggested by (Kim & Oshima, 2013), Benjamini-Hochberg multiple comparison correction is applied to all methods.

Evaluation of the results

Due to numerical estimation procedures in the nonlinear model (1.29) and in the IRT-based methods, convergence issues can be observed. It should be noted that large proportions of convergence failures may have significant impact on power and rejection rates. For items that fail to converge no results are obtained and no conclusion about DIF detection can be drawn. To make simulations comparable for all procedures, runs with the convergence issues are excluded and the proportion of these events is scored. Convergence failure rate is calculated as a ratio of items with the convergence issues and total number of generated items (that is total number of generated data sets times number of items). Rejection and power rate analyses are based only on 1,000 simulation iterations without convergence issues. All tests are performed at $\alpha = 0.05$ significance level.

Implementation

For all analyses, software **R** (version 3.3.2) is used (R Core Team, 2020). The nonlinear model (1.29) is fitted using the **difNLR R** package (Hladká & Martinková, 2020, see also Section 1.4). To specify suitable initial values, we consider approach based on linear approximation. Mean values of the standardized total score of the first and the third tertiles are spaced by line $\tilde{p}(x) = kx + q$, where x stands for the standardized total score (For discussion about starting values see also Section 1.6). Guessing parameter c stands for asymptotic minimum $p(-\infty)$ but taking into account a linear approximation $\tilde{p}$, this value would be $-\infty$ (considering only positive values of parameter k). Initial value of the guessing parameter is set as $\tilde{p}(-4)$ considering this value to be sufficient. Only non-negative values are taken into consideration and negative values are set to zero. The guessing parameter influences the difficulty and discrimination parameters. For cases with zero probability of guessing, difficulty parameter b is defined as $p(b) = \frac{1}{2}$. When considering positive guessing $c \in (0, 1)$, condition $p(b) = \frac{1+c}{2}$ holds instead. Hence initial value of b based on linear approximation $\tilde{p}$ is set to $b = \frac{\frac{1+c}{2} - q}{k}$. With zero probability of guessing, discrimination parameter a is defined as $p'(b) = \frac{a}{4}$, the slope in inflection point b divided by 4. With the positive guessing $c \in (0, 1)$, formula $p'(b) = \frac{a(1-c)}{4}$ is applied. Therefore, by using linear approximation, initial estimation of a is set to $a = \frac{4k}{1-c}$. To test for DIF presence, the F-test (1.26) is used.

The logistic regression procedure is implemented via function `glm()` from the **stats** package (R Core Team, 2020). To detect DIF, likelihood ratio test is performed (Agresti, 2010). The **difR R** package (Magis et al., 2010) is used to perform the Mantel-Haenszel test via the `difMH()` function and Lord's statistics for the 3PL IRT model are calculated using the `difLord()` function.

For uniform and non-uniform DIF, the 3PL IRT and the nonlinear regression model with the same guessing for both groups are considered: For the IRT approach, the 3PL IRT model for all data is fitted with the function `itemParEst()` and vector of common guessing parameters is estimated. Then, the 3PL IRT models for both groups are fitted with the fixed estimated guessing parameter and the `difLord()` function is applied. In case of the same guessing, the nonlinear model (1.29) is fitted using restriction $d_{iG_p} = 1$ and further compared to model with no group effect. In case of DIF caused by varying pseudo-guessing parameter models allowing different guessing are considered for the IRT as well as for the nonlinear model procedure.

1.5.2 Results

Convergence issues

Due to numerical estimation procedures in the nonlinear model (1.29) and in the Lord's test, convergence issues occur. Considering only DIF with the same guessing parameters for groups, the number of generated data sets rapidly decreases with increasing sample size and becomes stable. The average number of generated data sets for sample size of 1,000 is 1,391 (range $1,327 - 1,460$), for sample size of 1,500 it is 1,089 ($1,036 - 1,133$), for sample size of 2,000 it is 1,064 ($1,045 - 1,095$), for sample size of 3,000 it is 1,081 ($1,055 - 1,100$) and for sample size of 4,000 it is 1,085 ($1,059 - 1,121$). The trend is similar in case of DIF caused by varying pseudo-guessing parameters among groups, however the average number is 32,153 ($22,646 - 60,316$) which is several times higher than in previous scenarios. This is primarily due to high number of convergence issues in fitting 3PL IRT model with various guessing (Table 1.4).

Considering common guessing parameters, both for uniform and non-uniform scenario, the Lord's test results in a large proportion of convergence problematic items, however, with increasing number of examinees proportion of convergence failures declines rapidly (see Table 1.3). Similar tendency can be observed in the procedure using the nonlinear model, however the proportion of convergence failures is less than 1% (0.08–0.68%) for all scenarios in contrast to the Lord's test where the proportion reaches up over 10% (0.39–10.35%).

Considering DIF caused by different guessing among groups, method based on the nonlinear model performs slightly larger proportion of the convergence issues than in previous scenarios, however, it still remains under 1% (0.40–0.95%). This is not the case of the IRT model where the convergence failure rate greatly increases and remains on high level, above 19% (19.02–49.61%), even for the large sample sizes (Table 1.4).

Rejection rates

For almost all scenarios, the rejection rates (i.e., false positives) of the nonlinear model based DIF detection method and also for the Mantel-Haenszel test and the logistic regression procedure maintain below the 5% nominal level. The nominal value is exceeded only when 3 uniform DIF items and larger sample sizes ($>$ 2,000) are considered (Part C of Table 1.3) and in case of 3 DIF items caused by varying guessing and sample size of 4,000 (Part G of Table 1.4).

High rejection rates exceeding nominal level of 5% are apparent in the Lord's test in all studied scenarios with small sample sizes ($<$ 2,000). Nevertheless with the increasing sample size the rejection rates stabilize and reach the nominal value considering uniform and non-uniform DIF (Table 1.3). Similarly as for non-IRT methods, rejection rates are mildly exceeded for larger sample sizes in case of three uniform DIF items (Part C of Table 1.3). Supposing DIF caused by varying guessing the Lord's detection procedure is not able to control rejection rate disregarding sample size or proportion of DIF items (Table 1.4).

The situation is similar in the case where no DIF item is present in data set. All non-IRT procedures including the nonlinear model are able to control type I error. The nominal value of 5% is exceeded only by the Lord's test in case of sample sizes smaller than 3,000 (Part A of Table 1.3).

Power rates

When uniform DIF is considered, all three non-IRT procedures (the nonlinear model, the Mantel-Haenszel test, and the logistic regression) yield satisfactory high power rate (over 80%) in almost all scenarios. Although the power analysis shows superiority of the Mantel-Haenszel test, the differences between non-IRT methods are negligible, especially for smaller proportion of DIF items. While the Lord's test yields lower power in almost all uniform DIF scenarios, it gains satisfactory power on low rejection rate for sample sizes larger than 2,000 and also for sample size of 2,000 when one uniform DIF item was present (Parts B and C of Table 1.3).

For non-uniform DIF and sample size less than 2,000, no method achieves satisfactory power rates regardless of DIF items proportion (Parts D and E of Table 1.3). However, with the increasing sample size power rates increase rapidly. The logistic regression procedure outperforms other methods in terms of power at low rejection rate in almost all scenarios with power rates ranging from 36.13% to 100%. When one non-uniform DIF item is considered, the nonlinear model outmatches the Lord's test. For larger proportion of DIF items, the power rates of the nonlinear model and the Lord's test are comparable.

Supposing DIF caused by different guessing among groups, all non-IRT procedures including the nonlinear model gain satisfactory power in almost all scenarios. The only exception are the cases of small sample size (< 2,000) and large proportion of DIF items, where power rates are below value of 80% (Part G of Table 1.4). Otherwise the differences between non-IRT methods are inconsequential, as they are all close to 100% (see Table 1.4).

The strong increasing trend of power rates with the increasing sample size is pattern in almost all DIF procedures regardless of proportion of reference and focal group. The only exception is the Mantel-Haenszel test which is not able to detect non-uniform DIF not even in large sample sizes or in presence of three DIF items (see Parts D and E of Table 1.3).

Table 1.3: Rejection rates (RR), power rates (PR) and proportion of convergence failures (CF) for the DIF detection procedures.

		Sample size = 1,000 (500 per group)			Sample size = 1,500 (1,000 foc., 500 ref.)			Sample size = 2,000 (1,000 per group)			Sample size = 3,000 (2,000 foc., 1,000 ref.)			Sample size = 4,000 (2,000 per group)		
		RR	PR	CF	RR	PR	CF	RR	PR	CF	RR	PR	CF	RR	PR	CF
A. None DIF item																
	MH	0.19		0.00	0.24		0.00	0.18		0.00	0.27		0.00	0.24		0.00
	LR	0.24		0.00	0.24		0.00	0.23		0.00	0.29		0.00	0.25		0.00
	NLR	0.38		0.52	0.40		0.19	0.32		0.16	0.40		0.11	0.34		0.08
	LORD	10.93*		5.77	8.89*		1.28	5.10*		0.51	3.44		0.79	2.17		0.66
B. One uniform DIF item																
	MH	0.66	97.10°	0.00	0.62	99.90°	0.00	0.88	100.00°	0.00	0.95	100.00°	0.00	1.21	100.00°	0.00
	LR	0.64	96.00	0.00	0.66	99.50	0.00	0.87	100.00°	0.00	0.74	100.00°	0.00	0.96	100.00°	0.00
	NLR	0.82	96.50	0.68	0.87	99.50	0.22	1.08	100.00°	0.20	1.03	100.00°	0.14	1.34	100.00°	0.12
	LORD	9.99*	77.10	10.35	9.56*	95.20	1.11	4.11	99.80	0.80	4.19	100.00	0.60	2.93	100.00	0.65
C. Three uniform DIF items																
	MH	1.76	63.80°	0.00	2.61	73.27°	0.00	4.24	86.07°	0.00	5.79*	91.83	0.00	10.24*	97.77	0.00
	LR	1.53	58.50	0.00	2.06	69.37	0.00	3.15	82.40	0.00	4.34	89.07°	0.00	7.52*	96.40	0.00
	NLR	1.95	60.13	0.56	2.78	69.87	0.29	3.99	83.13	0.20	5.11*	89.30	0.08	9.00*	96.60	0.10
	LORD	11.03*	47.50	8.29	10.89*	66.77	0.91	5.23*	75.70	0.97	6.08*	86.73	0.40	6.51*	93.70	0.43
D. One non-uniform DIF item																
	MH	0.24	0.40	0.00	0.20	0.33	0.00	0.23	0.20	0.00	0.30	0.20	0.00	0.21	0.20	0.00
	LR	0.47	46.40°	0.00	0.54	69.50°	0.00	0.58	88.50°	0.00	0.58	96.70°	0.00	0.56	100.00°	0.00
	NLR	0.60	36.70	0.62	0.72	60.33	0.27	0.72	81.50	0.24	0.72	93.40	0.20	0.71	99.50	0.24
	LORD	10.63*	35.00	10.27	9.36*	55.36	1.21	3.58	72.50	0.70	4.01	94.70	0.49	2.51	99.70	0.81
E. Three non-uniform DIF items																
	MH	0.18	0.13	0.00	0.21	0.30	0.00	0.20	0.17	0.00	0.28	0.27	0.00	0.34	0.33	0.00
	LR	0.69	36.63°	0.00	1.03	55.70°	0.00	1.38	78.23°	0.00	1.91	89.60°	0.00	2.64	97.83°	0.00
	NLR	0.78	28.20	0.56	1.11	47.13	0.27	1.66	69.17	0.38	2.24	84.93	0.13	2.94	96.30	0.14
	LORD	9.77*	35.80	5.69	10.49*	59.47	0.55	3.92	77.83	0.59	4.87	91.70	0.50	3.04	98.50	0.39

Note. MH = Mantel-Haenszel test, LR = logistic regression, NLR = nonlinear model, LORD = Lord's test. An asterisk * indicates that the rejection rate exceeds nominal value of 5% and thus corresponding power is meaningless. A circle ° indicates the highest power at rejection rate lower than nominal value of 5%.

Table 1.4: Rejection rates (RR), power rates (PR) and proportion of convergence failures (CF) for the DIF detection procedures.

	Sample size = 1,000			Sample size = 1,500			Sample size = 2,000			Sample size = 3,000			Sample size = 4,000		
	(500 per group)			(1,000 foc., 500 ref.)			(1,000 per group)			(2,000 foc., 1,000 ref.)			(2,000 per group)		
	RR	PR	CF	RR	PR	CF	RR	PR	CF	RR	PR	CF	RR	PR	CF
F. One DIF item with varying guessing															
MH	0.56	93.40°	0.00	0.58	98.80°	0.00	0.78	99.80	0.00	0.81	100.00°	0.00	1.12	100.00°	0.00
LR	0.63	91.60	0.00	0.67	98.40	0.00	0.75	99.90°	0.00	0.75	100.00°	0.00	1.03	100.00°	0.00
NLR	0.45	89.50	0.91	0.55	98.00	0.70	0.56	99.90°	0.63	0.62	100.00°	0.44	0.71	100.00°	0.40
LORD	35.63*	77.70	49.61	34.91*	89.00	39.91	34.21*	90.60	23.40	36.09*	94.80	20.66	37.26*	96.40	19.05
G. Three DIF items with varying guessing															
MH	1.96	66.60°	0.00	2.62	79.97°	0.00	4.08	91.13°	0.00	5.76*	96.87	0.00	9.25*	99.43	0.00
LR	1.76	61.70	0.00	2.52	75.50	0.00	3.37	88.63	0.00	4.79	95.23°	0.00	7.67*	99.10	0.00
NLR	1.32	57.83	0.95	1.81	72.83	0.75	2.78	86.87	0.69	3.48	94.63	0.48	5.52*	98.73	0.48
LORD	33.24*	71.03	49.38	33.84*	80.13	36.68	33.24*	84.10	23.29	33.15*	89.47	20.69	34.71*	91.53	19.02

Note. MH = Mantel-Haenszel test, LR = logistic regression, NLR = nonlinear model, LORD = Lord's test. An asterisk * indicates that the rejection rate exceeds nominal value of 5% and thus corresponding power is meaningless. A circle ° indicates the highest power at rejection rate lower than nominal value of 5%.

1.5.3 Discussion and conclusion

In this section we presented results by Drabinová and Martinková (2017) who examined properties of the nonlinear model (1.29). The model is a restricted version of the model (1.1) with inattention parameter being the same for both groups and fixed at the value of 1, i.e., $d_{iG_p} = 1$. Model (1.29) is also a natural generalization of the logistic regression method (Swaminathan & Rogers, 1990) by allowing nonzero probability of guessing and by allowing different guessing for groups. In this simulation study, we showed pleasant properties of the nonlinear model (1.29) including low rate of convergence failures and in most cases sufficient power and low rejection rate. Thus as the only non-IRT method that accounts for guessing, the nonlinear model not only fills logical gap in DIF detection methodology but it also seems to be an useful alternative to other methods.

Obvious advantage of the nonlinear model method over IRT-model-based approaches is a pleasant behavior even in small sample sizes (1,000). Despite our assumption that sample size of 500 in each group would be sufficient for item calibration (Kim & Oshima, 2013; Siebert, 2013), Lord's test shows large proportion of convergence failures (Table 1.3). In this simulation study we report the convergence failures rate as a proportion of items that fail to converge. It should be noted that proposed nonlinear model performs the DIF detection procedure item by item. Thus the convergence issue in one item does not prevent the DIF testing in other items. In contrast, in the Lord's method the DIF test statistics is calculated for all items simultaneously and convergence issue in one item would cause its calculation to fail. A possible solution would be to fit the IRT model while excluding convergence problematic items. Besides that this approach takes an extra effort, it may also cause bias in estimated knowledge and hence possibly wrong conclusion in DIF detection. Summarizing, in practical implementation less time and effort is needed to fit models and to test for DIF with procedure based on the nonlinear model (1.29) than with the Lord's test.

Poor control of rejection rates for the Lord's test when considering multiple DIF items is consistent with the finding of Battauz (2019) and Wang and Yeh (2003). As expected, strong and consistent increasing trend in power rates with the increasing sample size is obvious in almost all DIF detection procedures and all studied scenarios. For sample size of 4,000 power of all other procedures is almost 100%. With the increasing sample size, the differences between methods decreases and IRT models become easy to fit. IRT-based approaches then bring more precise model and added value in terms of estimates of latent trait while the nonlinear model and the logistic regression model are only proxies to the 3PL and 2PL IRT models.

Looking closer at non-IRT approaches, although the Mantel-Haenszel test yields excellent results in uniform DIF detection, its poor performance for non-uniform DIF detection (i.e., power rate close to zero) makes it a limited tool in a study field, which is in line with findings by Swaminathan and Rogers (1990). For these reasons, it seems that nonlinear model, together with the logistic regression procedure, can be seen as useful alternatives to IRT methods, especially in smaller sample sizes ($< 2{,}000$), where the nonlinear model and the logistic regression procedures outperform other methods in terms of power.

Moreover, in uniform DIF detection, the nonlinear model achieves slightly better results than the logistic regression approach. This may suggest that the nonlinear method profits from more precise model by introducing guessing parameter c into the logistic regression procedure. In non-uniform DIF detection, the logistic regression procedure is superior to other methods, but achieved power rates remain on low level. One explanation may be that we consider only non-uniform DIF items with the same difficulty parameter for both groups. Moreover, for all methods, Benjamini-Hochberg multiple comparison correction is applied. Negative effect of using such procedures can be decrease of power, as noted by Kim and Oshima (2013), which may be the case in non-uniform DIF detection in small sample sizes. For further discussion, see Chapter 4. In case of DIF caused by varying pseudo-guessing parameters among groups, the DIF detection procedures gain satisfactory power in almost all scenarios even though the Mantel-Haenszel test and logistic regression method are not able to capture the fact that probability of guessing varies by group. In that case, both methods seem to project the difference of probabilities of guessing into difference of difficulties which is understandable since we assumed only difference in pseudo-guessing parameters (and not in discrimination).

Proposed nonlinear model (1.29) procedure fills a logical gap in DIF detection methodology. While in IRT-based DIF detection methods the third parameter is often taken into account, the logistic regression accounts only for the two parameters. The nonlinear model extends the logistic regression procedure in this way and to our best knowledge it is the only non-IRT method for DIF detection with the third, guessing parameter. Moreover it is the only non-IRT procedure which can test for difference in the pseudo-guessing parameters among groups and hence explore nature of DIF in more detail.

Proposed nonlinear model (1.29) method models the probability of correct answer with respect to (standardized) total score, which can be seen as inadequate estimate of knowledge. The main difference between the 3PL IRT-based methods and the nonlinear model approach is that in the IRT-based procedures the knowledge of examinees is modeled as an unobserved latent variable with standard normal distribution. Although the nonlinear model method can be viewed as less precise, our simulation study shows that for the task of DIF detection this proxy is ample: The nonlinear model procedure has low rejection rate, sufficient power and less convergence issues than the IRT approach, especially for smaller samples. Its good properties as well as easier implementation and interpretation predetermines the nonlinear model method to be a handy tool in identification of DIF.

As noted in Introduction, the common way of applying 3PL IRT model in DIF detection, when considering the same probability of guessing for both groups, is to fit the model on all data and estimate the common guessing parameter. Fixed estimate of guessing parameter is then applied into two separate models for focal and reference group (Magis et al., 2010). Further, the estimated parameters are re-scaled (Candell & Drasgow, 1988; Lautenschiager & Park, 1988) and then Lord's statistic is calculated. It should be noted that this approach can lead to biased standard errors and consequently to biased estimates (Battauz, 2019). Simultaneous estimation of parameters for both groups including guessing parameter is offered, e.g., in the **mirt R** package (Chalmers, 2012), however fitting without convergence issues in small sample sizes seems to be nearly impossible. Our procedures uses the simultaneous parameter estimation and as non-IRT approach does not encounter as many convergence issues.

We believe that also the currently proposed nonlinear procedure may benefit from further improvements. Better specification of initial values could lead to smaller proportion of convergence issues (see also Section 1.6). Also other estimating procedures can be implemented to provide more accurate estimates, such as weighted non-linear least squares or Bayesian methods.

As we showed at the beginning of this chapter, model can be extended by allowing upper asymptote to be smaller than one and thus introduce an non-IRT alternative to four-parameter IRT model (Barton & Lord, 1981). Besides, more than two groups can be taken into account (Magis, Raîche, Béland, & Gérard, 2011). The nonlinear DIF detection method can be also refined by implementing iterative purification similarly as for logistic regression (Zumbo, 1999) or IRT-based methods (Candell & Drasgow, 1988; Wang & Yeh, 2003, see also Chapter 4).

The current simulation study is limited to the investigated conditions as test length, nature and proportion of DIF items, and especially sample size. It should be noted that only large sample sizes are considered, which is not a necessarily realistic condition. Another restriction is that we consider only average difficulty items in designs of non-uniform DIF and DIF caused by varying guessing, where difficulty is the same for both groups. In the latter design we also considered the same value of discrimination for both groups.

Despite its limitations, this study demonstrates pleasant properties of the nonlinear model. Sufficient power rate and low rejection rate even in small sample sizes predetermine the nonlinear model to be an attractive and user friendly alternative to other procedures used in DIF detection. As the only one non-IRT approach, the nonlinear model allows for incorporation of guessing parameter into the model while it keeps the simplicity of the logistic regression procedure.

1.6 Simulation study – estimation procedures

The second study focused on comparison of various procedures to estimate parameters in the nonlinear model (1.1) which were described in Sections 1.2.1–1.2.4, including the nonlinear least squares, the maximum likelihood method, the EM algorithm, and the newly proposed algorithm based on parametric link function.

1.6.1 Starting values

The crucial part of the estimation is the specification of starting values for item parameters as it may have a great impact on the speed of the estimation process and also on its precision. Starting values which are far from the true item parameters may lead to the situation that algorithm yields only local extreme or even that algorithm does not converge. Therefore, we propose two different approaches to specify starting values described below – first based on CTT and the second based on a grid search. Both approaches are designed to compute starting values when standardized total score is used as the matching criterion. Both procedures may be used even in case when grouping variable G_p is present in data. In such a case, the initial values are computed separately for both groups and the differences in item parameters between the two groups are calculated.

Starting values based on CTT

The first approach is based on CTT which uses empirical probabilities to estimate item characteristics. This method is an updated and improved version of the algorithm which was described in Section 1.5 and which is currently used by the `startNLR()` function of the **`difNLR`** package. This newly proposed approach accounts for the variability of the matching criterion.

To explore average probabilities of the weakest and the strongest respondents, we first split respondents into the three groups based on tertiles of the matching criterion X_p. We then estimate the asymptotes: The estimate of the lower asymptote c_i for item i is calculated as the mean empirical probability of correct answer to item i of those $n_{(1)}$ respondents who have their matching criterion X_p lower than the average matching criterion of respondents from the group 1 (specified by the first tertile) $\bar{X}_{(1)}$ minus the sample standard deviation of X_p divided by 2, i.e.,

$$\begin{aligned}
\bar{X}_{(1)} &= \frac{3}{n}\sum_{p=1}^{\frac{n}{3}} X_{(p)}, \quad n_{(1)} = \sum_{p=1}^{n} \mathbf{1}_{\left[X_p < X_{(1)} - \frac{\mathrm{SD}(X)}{2}\right]} \\
\mathrm{SD}(X) &= \sqrt{\frac{1}{n-1}\sum_{p=1}^{n}\left(X_p - \bar{X}\right)^2}, \\
\widehat{c}_i &= \frac{1}{n_{(1)}}\sum_{p=1}^{n} Y_{pi}\mathbf{1}_{\left[X_p < X_{(1)} - \frac{\mathrm{SD}(X)}{2}\right]}.
\end{aligned}$$

Analogously, the estimate of the upper asymptote d_i for item i is calculated as the mean empirical probability of correct answer to item i of those $n_{(3)}$ who have their matching criterion X_p larger than the average matching criterion of respondents from the group 3 (specified by the third tertile) $\bar{X}_{(3)}$ plus sample standard deviation of X_p divided by 2, i.e.,

$$\begin{aligned}
\bar{X}_{(3)} &= \frac{3}{n}\sum_{p=\frac{2n}{3}+1}^{n} X_{(p)}, \quad n_{(3)} = \sum_{p=1}^{n} \mathbf{1}_{\left[X_p > X_{(3)} + \frac{\mathrm{SD}(X)}{2}\right]} \\
\widehat{d}_i &= \frac{1}{n_{(3)}}\sum_{p=1}^{n} Y_{pi}\mathbf{1}_{\left[X_p > X_{(3)} + \frac{\mathrm{SD}(X)}{2}\right]}.
\end{aligned}$$

The slope of the ICC b_{i1} for item i is estimated as a difference between mean empirical probabilities of correct answer to item i of the groups 3 and 1 multiplied by 4. Multiplication by 4 comes from the derivative of the logistic function in its inflection point (see Section 1.1.1).

The intercept b_{i0} is estimated as follows: The center point between the two asymptotes for item i is calculated $\dot{Y}_i = \frac{\widehat{c}_i + \widehat{d}_i}{2}$ and then we look for the level of the matching criterion X_p, which would correspond to this empirical probability $\dot{Y}_i$. In more detail, we proceed as follows: First, the empirical probability $\dot{Y}_i$ is calculated. Second, we round the matching criterion to the one decimal place and we calculate relevant empirical probabilities of correct answer to item i. Third, we compute absolute distances between these empirical probabilities and the value $\dot{Y}_i$, while weighting them with division of proportions of the rounded matching criterion X_p, i.e., we get larger values for the empirical values which are more distant from $\dot{Y}_i$ and for those which are calculated based on less observations. Fourth, weighted distances are smoothed by accounting for the neighbor values with the weight of 0.1. Fifth, we compute the values of the matching criterion which give the minimal weighted and smoothed distance of empirical probabilities. Sixth, we take minimum, respectively maximum, to which we add, respectively subtract, sample standard deviation of the matching criterion divided by two. Finally, we take an average value of the matching criterion which is in rage specified by these values. The $\widehat{b}_{i0}$ is then specified as minus this value multiplied by the estimate $\widehat{b}_{i1}$.

The full syntax of the `startCTT()` function which computes starting values based on CTT described here can be found in Appendix A.5.

Starting values using grid search

Another approach which we consider here is based on grid search. We start with the starting values computed based on CTT as described above. Then we create a 4 dimensional grid which covers values of item parameter estimates in the neighbourhood of those specified by CTT method. For each combination of parameters, the value of the log-likelihood function is calculated. The final item estimate of set of parameters is then the one which corresponds to the maximum value of the computed log-likelihoods values.

The full syntax of the `startGRID()` function can be found in Appendix A.6.

Comparison of approaches

In this part we directly compare both approaches to calculate starting values in terms of precision in small simulation study accounting only for one group. We evaluate their performance in sense of bias, i.e., the mean/median difference between the starting values and the true values of parameters, in Mean Squared Error (MSE), i.e., mean/median square difference between the starting values and the true parameters, and in sense of whether their use leads to convergence issues or not when applying four different estimation procedures – the nonlinear least squares, the maximum likelihood, the EM algorithm, and the algorithm based on parametric link function.

To generate data, we used the following parameters: The intercept $b_0 = 0$ and the slope $b_1 = 2$. We considered several values for the asymptote parameters: $c \in \{0, 0.1, 0.2, 0.3\}$ and $d \in \{0.7, 0.8, 0.9, 1\}$. The matching criterion was generated from standard normal distribution. Binary responses were then generated from the Bernoulli distribution with the calculated probabilities based on the nonlinear model (1.2). Sample size was set to 500. Each scenario was replicated 1,000 times.

Results. Mean and median biases of both approaches were similar and close to zero for all four parameters. Approach based on CTT yielded slightly more precise starting values of parameters c, d, and b_1 in terms of MSE, while grid search provided more precise estimate of intercept parameter b_0 (Table 1.5).

The percentage of crashed estimation procedures was similar in both approaches to calculating starting values. Slightly lower rate was observed for the nonlinear least squares, the maximum likelihood method, and the algorithm based on parametric link function when approach based on grid search was used,

Table 1.5: Mean and median bias and MSE over all parameters choices and over all methods for the two different approaches to calculate starting values – based on CTT and based on grid search.

Method/parameter	Bias		MSE	
	Mean	Median	Mean	Median
CTT				
c	0.016	0.006	0.004	0.001
d	-0.011	0.000	0.004	0.001
b_0	-0.007	-0.008	0.310	0.136
b_1	0.043	0.060	0.243	0.138
grid				
c	-0.008	0.000	0.005	0.001
d	0.007	0.000	0.005	0.001
b_0	0.014	0.015	0.176	0.065
b_1	0.140	0.093	0.403	0.155

while the EM algorithm benefited more from the method based on CTT (Table 1.6). In total, when using the approach based on CTT, 0.297% of all estimating procedures crashed while for the grid search it was 0.250%.

Table 1.6: Percentage of crashed estimating procedures with respect to method used to calculate starting values – based on CTT and based on grid search.

Method	Crashed [%]
CTT	
Nonlinear least squares	0.562
Maximum likelihood	0.188
EM algorithm	0.375
Parametric link function	0.062
grid	
Nonlinear least squares	0.500
Maximum likelihood	0.062
EM algorithm	0.438
Parametric link function	0.000

To conclude, differences between the two approaches to calculate starting values were small in all four parameters c, d, b_0, and b_1 and also in all four estimation methods. Since the initial run of the grid search is based on the CTT approach and then it goes through four dimensional grid, it is clear that this method is more computationally demanding. As it did not provide notably improved starting values, we will further use approach based on CTT.

1.6.2 Comparison of the estimation methods

To compare the four estimation procedures and to illustrate differences between them in terms of precision, convergence status, and time consumption, small simulation study is performed in this section.

Design of simulation study

Generation of the data. To generate data, we used the same parameters as in illustrating implementation of the EM algorithm (see Section 1.4.3) and of the algorithm based on parametric link function (see Section 1.4.4): $b_0 = 0$, $b_1 = 1$, $b_2 = -1$, $b_3 = 0.5$, $c = 0.2$, $c_{\text{DIF}} = 0.1$, $d = 1$, $d_{\text{DIF}} = -0.1$. The matching criterion was again generated from the standard normal distribution. Binary responses were generated from the Bernoulli distribution with the calculated probabilities based on the nonlinear model (1.2), true parameters, and the matching criterion variable. Sample size was set to 1,000 and 10,000, i.e., 500 and 5,000 per each group respectively. Each scenario was replicated 1,000 times.

Estimation methods and implementation. Four estimation methods were considered: The nonlinear least squares with the sandwich estimator of the asymptotic covariance matrix (1.15) using the `nls()` function and `"port"` algorithm, the maximum likelihood estimation with the `optim()` function with the `"L-BFGS-B"` algorithm, the EM algorithm as was shown in Section 1.4.3, and the algorithm based on parametric link function as was described in Section 1.4.4. The maximum number of iterations was set to 2,000 for all methods and convergence criterion was set to 10^{-6} where possible. For each simulation run, the starting values were calculated using the `startCTT()` function and applied for all four methods.

Simulation evaluation. To compare estimation methods, we first computed mean and median number of iteration runs together with the convergence status of the methods, i.e., percentage of converged simulation runs, of those which crashed, and those which reached the maximum number of iterations without convergence. We then calculated mean and median parameter estimates together with the mean bias, i.e., mean difference between estimates and true values. We also included model-based standard errors, i.e., mean of the standard errors of the parameter estimates, and empirical standard deviations, i.e., standard deviations of the parameter estimates. Finally, we measured time necessary to complete the estimation process.

Results

Convergence status. All four methods had low percentage of iterations that crashed (caused error when fitting). The highest rate was observed for the nonlinear least squares and sample size of 1,000, where it was 3.1%, followed by the algorithm based on parametric link function with 1.9%. Both the maximum likelihood method and the EM algorithm had no crashed simulation runs which was also the case for the algorithm based on parametric link and sample size of 10,000 (Table 1.7). While the nonlinear least squares and the maximum likelihood estimation did not reach the maximum number of iterations in any simulation runs, it was not the case for the EM algorithm where 21.8% and 31.3% of all simulation runs ended after reaching this limit without convergence for sample sizes of 1,000 and 10,000 respectively. Method based on parametric link function reached the maximum limit of 2,000 iterations only in 0.2% of simulation runs when smaller sample size was considered (Table 1.7).

Table 1.7: Convergence status and number of iterations for the four estimation methods.

Method	Convergence status [%]			Number of iterations*	
	Converged	Crashed	DNF	Mean	Median
			$n = 1,000$		
NLS	96.90	3.10	0.00	13.00 / 13.00	11.00 / 11.00
MLE	100.00	0.00	0.00	88.64 / 88.64	80.00 / 80.00
EM	78.20	0.00	21.80	1065.15 / 804.54	972.00 / 737.00
PLF	97.90	1.90	0.20	59.87 / 55.91	32.00 / 32.00
			$n = 10,000$		
NLS	99.80	0.20	0.00	6.34 / 6.34	6.00 / 6.00
MLE	100.00	0.00	0.00	77.38 / 77.38	77.00 / 77.00
EM	68.70	0.00	31.30	1430.20 / 1170.60	1449.00 / 1148.00
PLF	100.00	0.00	0.00	24.42 / 24.42	23.00 / 23.00

Note. * the first value calculated from all simulation runs, the second value calculated from converged simulation runs only, DNF = did not finish, NLS = nonlinear least squares, MLE = maximum likelihood estimation, EM = expectation-maximization algorithm, PLF = algorithm based on parametric link function.

Number of iterations. Methods also differed in a number of iterations until the estimation process ended. The EM algorithm yielded the largest mean and median number of iterations which was overestimated by simulation runs which did not finish. However, both the mean and median number of iterations remained on a very high level even in case when only successfully converged simulation runs were taken into account and when sample size increased. The least number of iterations till convergence was needed for the nonlinear least squares followed by the algorithm based on parametric method (Table 1.7).

Parameter estimates. Estimation methods seemed to be sensitive to specification of the starting values and although the estimation process converged, the estimates were far from the true parameters, especially for the smaller sample size of 1,000. These extreme estimates have great impact especially on mean values, bias, model-based standard errors, and empirical standard deviations (Table A.4). This was especially the case for the maximum likelihood method and for the EM algorithm. Therefore, estimates below the 1$^{\text{st}}$ and above the 99$^{\text{th}}$ percentiles of each parameter were removed from simulations for sample size of 1,000 and evaluating statistics were based on remaining values only.

For sample size of 1,000, the algorithm based on parametric link function yielded the best precision in terms of the mean bias for all four parameters b_0, b_1, b_2, and b_3, followed by the EM algorithm and maximum likelihood method (left part of Table 1.8). The only difference can be seen in parameter b_3, where the second most precise estimation was performed by the nonlinear least squares. The model-based standard errors were closest to the empirical standard deviations when the algorithm based on parametric link function was used in parameters b_0, b_2, and b_3. For the parameter b_1 it was by the nonlinear least squares and the maximum likelihood method.

For the smaller sample size, the mean and median estimates were close to the true values of parameters c, c_{DIF}, d, and d_{DIF} for all four estimation methods and the differences between methods were small (right part of Table 1.8). The most precise estimates of parameters c and c_{DIF} were obtained with the nonlinear least squares. The least biased estimate of parameter d was provided by the maximum likelihood while for parameter d_{DIF} it was by the EM algorithm. The model-based standard errors were closest to the empirical standard deviations when the nonlinear least squares were applied for all four parameters.

For the larger sample size of 10,000, all parameter estimates were less biased and the model-based standard errors were closer to the empirical standard deviations. Differences between estimation methods were small (Table 1.9).

Time consumption. The mean and median time to finish estimating procedure varied among methods. The nonlinear least square approach was the fastest method with the mean/median time of 0.103/0.095 seconds (range 0.001–0.477) for smaller sample size and 0.447/0.440 seconds (range 0.010–1.054) for larger sample size, followed by the maximum likelihood method with the times of 0.318/0.293 seconds (0.132–1.419) and 2.134/2.116 seconds (1.150–3.528). For the algorithm based on parametric link function the time consumption was higher but remained on user-friendly level of 1.711/0.878 seconds (0.004–142.472) for smaller sample size and of 5.387/4.861 seconds (2.128–28.804) for larger sample size. On the other hand, the EM algorithm was very slow with the average time of 19.518/18.749 seconds (0.170–52.818) for sample size of 1,000 and even 260.489/262.128 seconds (15.452–440.295) for sample size of 10,000, which was closely connected to the number of iterations necessary till the estimation ended.

1.6.3 Summary

In this part we performed two small simulation studies. The first was intended to compare two approaches to calculate starting values for the estimation procedures. This included approach based on CTT and approach which includes grid search to evaluate the best estimate. Due to time consumption of the grid search and not significantly improved starting values we decided to further use the approach based on CTT.

The second simulation study was performed to illustrate differences between the four estimation methods in the nonlinear model (1.1): The nonlinear least squares, the maximum likelihood method, the EM algorithm, and the newly proposed algorithm based on parametric link function. It seems that estimation procedures are sensitive to specification of starting values especially for smaller

Table 1.8: Mean and median parameter estimates with the bias, model based standard error, and empirical standard deviation for $n = 1,000$.

	NLS	MLE	EM	PLF		NLS	MLE	EM	PLF
b_0					c				
Count	946	975	977	972	Count	960	987	989	973
Mean	0.063	0.013	0.010	0.008	Mean	0.216	0.218	0.223	0.221
Median	0.101	0.036	0.034	0.026	Median	0.237	0.239	0.244	0.247
Bias	0.063	0.013	0.010	0.008	Bias	0.016	0.018	0.023	0.021
MBSE	0.632	0.621	0.611	0.612	MBSE	0.181	0.206	0.191	0.198
ESD	0.509	0.515	0.508	0.516	ESD	0.138	0.137	0.131	0.132
b_1					c_{DIF}				
Count	936	977	985	972	Count	951	972	982	965
Mean	1.447	1.411	1.408	1.392	Mean	0.073	0.071	0.069	0.065
Median	1.258	1.238	1.255	1.245	Median	0.066	0.065	0.065	0.056
Bias	0.447	0.411	0.408	0.392	Bias	−0.027	−0.029	−0.031	−0.035
MBSE	0.862	0.799	0.784	0.780	MBSE	0.217	0.239	0.223	0.232
ESD	0.710	0.647	0.609	0.557	ESD	0.163	0.161	0.156	0.157
b_2					d				
Count	954	970	972	974	Count	955	990	990	975
Mean	−1.378	−1.342	−1.340	−1.197	Mean	0.942	0.951	0.949	0.950
Median	−1.218	−1.202	−1.194	−1.146	Median	0.965	0.994	0.979	0.983
Bias	−0.378	−0.342	−0.340	−0.197	Bias	−0.058	−0.049	−0.051	−0.050
MBSE	1.322	1.221	1.184	1.116	MBSE	0.114	0.147	0.141	0.145
ESD	1.441	1.378	1.456	1.143	ESD	0.064	0.062	0.060	0.060
b_3					d_{DIF}				
Count	948	971	974	977	Count	947	975	979	969
Mean	0.790	0.804	0.818	0.589	Mean	−0.043	−0.045	−0.047	−0.042
Median	0.451	0.451	0.447	0.424	Median	−0.020	−0.005	−0.024	−0.012
Bias	0.290	0.304	0.318	0.089	Bias	0.057	0.055	0.053	0.058
MBSE	2.026	1.765	1.713	1.663	MBSE	0.208	0.255	0.241	0.255
ESD	2.363	2.330	2.421	1.909	ESD	0.112	0.111	0.109	0.108

Note. NLS = nonlinear least squares, MLE = maximum likelihood estimation, EM = expectation-maximization algorithm, PLF = method based on parametric link function, Count = number of parameter estimates excluding 1$^{\text{st}}$ and 99$^{\text{th}}$ percentile and crashed simulation runs, MBSE = model based standard error, ESD = empirical standard deviation.

sample sizes. While the estimation process converged, sometimes the final estimates were far from the true parameters, which was especially the case for the maximum likelihood method and the EM algorithm. With the increasing sample size, these issues disappeared and the differences between the methods were negligible.

The second simulation study suggested that all four methods similarly precisely estimated asymptote parameters for the both group, while the nonlinear least squares performed slightly better than other methods when smaller sample size was considered. On the other hand, the algorithm based on parametric link function seemed to slightly outperform other methods when estimating parameters b_0, b_1, b_2, and b_3 for smaller sample size.

It should be noted that the second simulation study was performed rather to illustrate estimating procedures than to offer their complex comparison. To compare estimating methods in more detail, simulation study including several sets of parameters and more levels of sample size should be designed.

Table 1.9: Mean and median parameter estimates with the bias, model based standard error, and empirical standard deviation for $n = 10,000$.

	NLS	MLE	EM	PLF		NLS	MLE	EM	PLF
b_0					c				
Count	998	1000	1000	1000	Count	998	1000	1000	1000
Mean	0.023	0.007	0.006	−0.007	Mean	0.206	0.209	0.212	0.218
Median	0.023	0.005	0.005	−0.009	Median	0.210	0.212	0.215	0.219
Bias	0.023	0.007	0.006	−0.007	Bias	0.006	0.009	0.012	0.018
MBSE	0.164	0.161	0.161	0.161	MBSE	0.064	0.064	0.062	0.061
ESD	0.160	0.159	0.158	0.147	ESD	0.056	0.054	0.052	0.048
b_1					c_{DIF}				
Count	998	1000	1000	1000	Count	998	1000	1000	1000
Mean	1.066	1.068	1.079	1.090	Mean	0.092	0.089	0.086	0.080
Median	1.042	1.052	1.065	1.070	Median	0.090	0.086	0.084	0.079
Bias	0.066	0.068	0.079	0.090	Bias	−0.008	−0.011	−0.014	−0.020
MBSE	0.193	0.191	0.191	0.194	MBSE	0.070	0.070	0.068	0.067
ESD	0.151	0.139	0.136	0.138	ESD	0.061	0.060	0.057	0.054
b_2					d				
Count	998	1000	1000	1000	Count	998	1000	1000	1000
Mean	−1.040	−1.025	−1.024	−0.999	Mean	0.982	0.985	0.982	0.981
Median	−1.039	−1.015	−1.016	−0.993	Median	1.000	1.000	0.991	0.995
Bias	−0.040	−0.025	−0.024	0.001	Bias	−0.018	−0.015	−0.018	−0.019
MBSE	0.256	0.252	0.252	0.253	MBSE	0.046	0.047	0.046	0.046
ESD	0.249	0.250	0.249	0.244	ESD	0.025	0.023	0.022	0.025
b_3					d_{DIF}				
Count	998	1000	1000	1000	Count	998	1000	1000	1000
Mean	0.474	0.465	0.459	0.454	Mean	−0.078	−0.079	−0.078	−0.080
Median	0.452	0.448	0.439	0.444	Median	−0.083	−0.085	−0.081	−0.084
Bias	−0.026	−0.035	−0.041	−0.046	Bias	0.022	0.021	0.022	0.020
MBSE	0.349	0.339	0.338	0.343	MBSE	0.072	0.073	0.071	0.071
ESD	0.311	0.304	0.293	0.287	ESD	0.057	0.057	0.053	0.054

Note. NLS = nonlinear least squares, MLE = maximum likelihood estimation, EM = expectation-maximization algorithm, PLF = method based on parametric link function, Count = number of parameter estimates excluding crashed simulation runs, MBSE = model based standard error, ESD = empirical standard deviation.

2. Generalized logistic regression models for polytomous items

The logistic regression procedure which estimates the probability of the correct answer can be extended to estimate the probability of partial credit scores or option choices. This chapter is adapted from Hladká and Martinková (2020) and we review generalized logistic regression models for DIF and DDF detection among polytomous items. In Section 2.1, we introduce group-specific cumulative logit and adjacent category logit models for DIF detection among ordinal data. In Section 2.2, we introduce group-specific multinomial model for DDF detection among nominal data. In both sections, besides the model specification, we newly describe in more detail the maximum likelihood method for estimation of item parameters and related DIF and DDF detection procedures. Finally, we show implementation of the models within the **R** software (R Core Team, 2020) and the **difNLR** package as was described in Hladká and Martinková (2020).

2.1 Group-specific models for ordinal items

When the responses are ordinal, the item response patterns can be described by the cumulative logit model (see, e.g., Agresti, 2010, Section 7.2) or by the adjacent category logit model (see, e.g., Agresti, 2010, Section 7.4). Both models can be naturally used for DIF detection by introducing a group membership variable G_p and its interaction with the matching criterion X_p.

2.1.1 Cumulative logit model

Probably the most popular logit model which reflects an ordinal nature of data is the cumulative logit model (Agresti, 2010, Section 7.2). This model can be used to describe functioning of the items based on observed respondent's ability, group membership variable, and their mutual interaction.

In contrast to Chapter 1, here we start with the classical intercept-slope parametrization, as it is more convenient here. Considering $K_i + 1$ ordered outcome categories for the item i, i.e., $Y_{pi} \in \{0, 1, \dots, K_i\}$, the probability of gaining at least $k = 1, \dots, K_i$ points on item i by respondent p is given by the logit model for DIF detection as

$$\mathsf{P}(Y_{pi} \geq k | X_p, G_p) = \frac{e^{\beta_{i0k} + \beta_{i1} X_p + \beta_{i2} G_p + \beta_{i3} X_p{:}G_p}}{1 + e^{\beta_{i0k} + \beta_{i1} X_p + \beta_{i2} G_p + \beta_{i3} X_p{:}G_p}}, \tag{2.1}$$

for $k = 1, \dots, K_i$. The K_i parameters β_{i0k} are k category-specific intercepts within item i, ordered by the category index k, i.e., $\beta_{i01} < \beta_{i02} < \dots < \beta_{i0K_i}$. The slope parameter β_{i1} is connected to the effect of the observed ability X_p on probability of gaining at least $k = 1, \dots, K_i$ points in item i and it is assumed to be the same for all categories $k = 1, \dots, K_i$. The parameter β_{i2} is related to the effect of group membership. Finally, the parameter β_{i3} is associated with the effect of interaction between the matching criterion X_p and the group membership variable G_p. In total, we end up with $K_i + 3$ parameters to estimate in item i. Note that

$$\mathsf{P}(Y_{pi} \geq 0 | X_p, G_p) = 1, \ \forall i, \ \forall p,$$

which means that no parameters are estimated for category $k = 0$.

The category probability, i.e., the probability of gaining exactly k points in item i, is then calculated as a difference between cumulative probabilities of the two adjacent categories:

$$\mathsf{P}(Y_{pi} = k | X_p, G_p) = \mathsf{P}(Y_{pi} \geq k | X_p, G_p) - \mathsf{P}(Y_{pi} \geq k + 1 | X_p, G_p),$$

for categories $k = 0, \ldots, K_i - 1$, while

$$\mathsf{P}(Y_{pi} = K_i | X_p, G_p) = \mathsf{P}(Y_{pi} \geq K_i | X_p, G_p).$$

Parametrization. Analogously as for the nonlinear model (1.1), the group-specific cumulative logit model (2.1) can be defined using the IRT notation with the difficulty and discrimination parameters:

$$\mathsf{P}(Y_{pi} \geq k | X_p, G_p) = \frac{e^{a_{iG_p}(X_p - b_{ikG_p})}}{1 + e^{a_{iG_p}(X_p - b_{ikG_p})}}. \tag{2.2}$$

The $2K_i$ parameters $b_{ikG_p} = b_{ik} + b_{ik\mathrm{DIF}} G_p$, where $G_p \in \{0, 1\}$ and $k = 1, \ldots, K_i$, indicate the level of the matching criterion X_p for which the respondents from group G_p have probability of 0.5 to gain at least k points in item i. The parameter $b_{ik\mathrm{DIF}}$ can be interpreted as the difference in difficulty of item i between the focal and the reference group for category $k = 1, \ldots, K_i$. The parameter b_{ik} can be seen as category-k-specific difficulty of item i for the reference group. The two discrimination parameters $a_{iG_p} = a_i + a_{i\mathrm{DIF}} G_p$ describe the slope of the logistic curve of the cumulative probabilities for the two groups, however, they are assumed to be the same for all categories $k = 1, \ldots, K_i$. In total, the group-specific cumulative logit model using the IRT parametrization includes $2K_i + 2$ parameters for item i, while only $K_i + 3$ are freely estimated as described below.

Similarly as for the nonlinear model (1.1), the respondent's p ability X_p can be described by their standardized total test score or other observed ability variable. Therefore, the cumulative logit model (2.2) can be seen as a proxy to a graded response IRT model (Samejima, 1969) or more precisely, its group-specific extension.

Mutual relationship between the parameters using the intercept-slope notation (2.1) and the parameters using the IRT notation (2.2) is then given as follows:

$$\begin{aligned} \beta_{i0k} &= -a_i b_{ik}, \\ \beta_{i1} &= a_i, \\ \beta_{i2} &= -a_{i\mathrm{DIF}} b_{ik} - a_i b_{ik\mathrm{DIF}} - a_{i\mathrm{DIF}} b_{ik\mathrm{DIF}}, \\ \beta_{i3} &= a_{i\mathrm{DIF}}. \end{aligned} \tag{2.3}$$

In other words, for the IRT parametrization we have

$$b_{ik} = -\frac{\beta_{i0k}}{\beta_{i1}},$$

$$b_{ik\text{DIF}} = \frac{\beta_{i1}\beta_{i2} - \beta_{i0k}\beta_{i3}}{\beta_{i1}\left(\beta_{i1} + \beta_{i3}\right)},$$

therefore both sets b_{ik} and $b_{ik\text{DIF}}$ depend on category k as they include category-specific intercept β_{i0k}, which results in more parameters in the IRT notation than in the classical intercept-slope parametrization. In what follows, we work with the classical intercept-slope parametrization as it is used in estimation and implementation. To provide item parameters in the IRT notation, the $K_i + 3$ parameters are estimated using the classical intercept-slope parametrization and the delta method (Doob, 1935) is applied to calculate new parameter values and their standard errors.

Estimation and asymptotic properties

The natural way how to estimate item parameters is to use the maximum likelihood method. For item i and respondent p, let $(Y_{pi0}, \ldots, Y_{piK_i})$ be the binary indicators of the response patterns defined as $Y_{pik} = 1$ when respondent p gained k points in item i and $Y_{pik} = 0$ otherwise. The likelihood function is then

$$\begin{aligned} L_i &= \prod_{p=1}^{n}\prod_{k=0}^{K_i} \mathsf{P}(Y_{pi} = k|X_p, G_p)^{Y_{pik}} \\ &= \prod_{p=1}^{n}\prod_{k=0}^{K_i-1} \left[\mathsf{P}(Y_{pi} \geq k|X_p, G_p) - \mathsf{P}(Y_{pi} \geq k+1|X_p, G_p)\right]^{Y_{pik}}, \end{aligned}$$

where $\mathsf{P}(Y_{pi} \geq k|X_p, G_p)$ is given by (2.1). The log-likelihood function takes the following form:

$$\begin{aligned} l_i &= \sum_{p=1}^{n}\sum_{k=0}^{K_i-1} Y_{pik}\log\left(\mathsf{P}(Y_{pi} \geq k|X_p, G_p) - \mathsf{P}(Y_{pi} \geq k+1|X_p, G_p)\right) \\ &= \sum_{p=1}^{n}\sum_{k=0}^{K_i-1} Y_{pik}\log\left(\frac{e^{\beta_{i0k}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}{1+e^{\beta_{i0k}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}\right. \\ &\qquad\qquad \left. - \frac{e^{\beta_{i0(k+1)}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}{1+e^{\beta_{i0(k+1)}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p}}\right), \end{aligned}$$

which can be shown to be a concave function (Pratt, 1981).

Parameter estimates are then obtained as a solution of estimating equations $\frac{\partial l_i}{\partial \gamma_{ij}} = 0$, $j = 1, \ldots, K_i + 3$, where $\boldsymbol{\gamma}_i = (\beta_{i01}, \ldots, \beta_{i0K_i}, \beta_{i1}, \beta_{i2}, \beta_{i3})$ is the set of $K_i + 3$ parameters. It is easy to see that the solution does not have a closed form. Therefore, numerical approaches, such as iteratively re-weighted least squares or Fisher scoring algorithm (Bliss, 1935), need to be applied.

Maximum likelihood estimators of the item parameters have pleasant properties such as consistency and asymptotic normality. Estimated asymptotic covariance matrix is given by the inverse of the Hessian matrix.

2.1.2 Adjacent category logit model

Another option how to model ordinal group-specific item responses and to test for DIF is the adjacent category logit model (see, e.g., Agresti, 2010, Section

7.4). This model focuses on moving from category $k-1$ to category k and models their log odds by a linear predictor. In other words, for K_i+1 outcome ordinal categories we have

$$\log \frac{\mathsf{P}(Y_{pi}=k|X_p,G_p)}{\mathsf{P}(Y_{pi}=k-1|X_p,G_p)} = \beta_{i0k}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p,$$

where $k=1,\ldots,K_i$ and category-specific intercepts are ordered by the category index k, i.e., $\beta_{i01}<\beta_{i02}<\cdots<\beta_{i0K_i}$. Using recursive formula, it can be shown that the category probability takes the following form:

$$\mathsf{P}(Y_{pi}=k|X_p,G_p) = \frac{e^{\sum_{l=0}^{k}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}}{\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}}, \tag{2.4}$$

where $k=0,\ldots,K_i$ and the terms for $k=0$ are set to zero, i.e., $\beta_{i00}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p=0$, and hence

$$\mathsf{P}(Y_{pi}=0|X_p,G_p) = \frac{1}{\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}}.$$

The K_i parameters β_{i0k} are the intercepts for category $k=1,\ldots,K_i$ within item i. The slope parameter β_{i1} is then connected with the effect of the matching criterion X_p on probability of gaining k points in item i and it is assumed to be the same for all categories $k=1,\ldots,K_i$. The parameter β_{i2} is associated with the effect of group membership. Finally, the parameter β_{i3} is related to the effect of interaction between the matching criterion X_p and the group membership variable G_p. In summary, we have K_i+3 parameters in total to estimate for item i.

Parametrization. Similarly as for the nonlinear model (1.1) and the group-specific cumulative logit model detection (2.2), the adjacent category logit model for DIF detection (2.4) can be defined using the IRT notation:

$$\mathsf{P}(Y_{pi}=k|X_p,G_p) = \frac{e^{\sum_{l=0}^{k} a_{iG_p}\left(X_p-b_{ilG_p}\right)}}{\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j} a_{iG_p}\left(X_p-b_{ilG_p}\right)}}, \tag{2.5}$$

for $k=1,\ldots,K_i$. The $2K_i$ parameters $b_{ikG_p}=b_{ik}+b_{ik\mathrm{DIF}}G_p$, $G_p\in\{0,1\}$, indicate the level of the matching criterion (or other observed ability) for which the respondents from group G_p have the same probability to gain $k-1$ and k points. In other words, b_{ikG_p} are group-specific locations of the response probability intersections which can be interpreted as levels of the matching criterion required to transition from category $k-1$ to category k. The two discrimination parameters $a_{iG_p}=a_i+a_{i\mathrm{DIF}}G_p$ denote group-specific item slopes which are assumed to be the same for all categories $k=1,\ldots,K_i$. In total, the group-specific adjacent category logit model using the IRT notation includes $2K_i+2$ parameters for item i. Analogously as for the cumulative logit model, the classical intercept-slope notation is used for estimation and implementation and K_i+3 parameters are freely estimated. Mutual relationship between the parameters using the classical intercept-slope notation (2.4) and those using the notation based on IRT

(2.5) is the same as for the cumulative logit model as stated in (2.3). Therefore, parameters from the IRT notation are calculated using (2.3) and the standard errors are obtained using the delta method (Doob, 1935).

Similarly as for the previous models, the matching criterion X_p can be described by respondent's standardized total test score or other observed ability variable. Therefore, the adjacent category logit model (2.5) can be seen as a proxy to the generalized partial credit IRT model (Muraki, 1992), extended to the group-specific version for DIF detection.

Estimation and asymptotic properties

Let again $(Y_{pi0}, \dots, Y_{piK_i})$ be binary indicators of the responses to item i by respondent p, with $Y_{pik} = 1$ when respondent p gained k points in item i and $Y_{pik} = 0$ otherwise. The likelihood function is

$$L_i = \prod_{p=1}^{n} \prod_{k=0}^{K_i} \mathsf{P}(Y_{pi} = k | X_p, G_p)^{Y_{pik}},$$

where $\mathsf{P}(Y_{pi} = k|X_p, G_p)$ is given by (2.4). The log-likelihood function then takes the following form:

$$\begin{aligned}
l_i = & \sum_{p=1}^{n}\sum_{k=0}^{K_i} Y_{pik} \log\left(\mathsf{P}(Y_{pi} = k|X_p, G_p)\right) \\
= & \sum_{p=1}^{n}\sum_{k=0}^{K_i} Y_{pik} \log\left(\frac{e^{\sum_{l=0}^{k}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}}{\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}}\right) \\
= & \sum_{p=1}^{n}\sum_{k=0}^{K_i} Y_{pik} \left[\sum_{l=0}^{k}(\beta_{i0l} + \beta_{i1}X_p + \beta_{i2}G_p + \beta_{i3}X_p : G_p)\right. \\
& \left. - \log\left(\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}\right)\right] \\
= & \sum_{p=1}^{n}\sum_{k=0}^{K_i}\left[Y_{pik}k\left(\beta_{i1}X_p + \beta_{i2}G_p + \beta_{i3}X_p : G_p\right) + Y_{pik}\sum_{l=1}^{k}\beta_{i0l}\right. \\
& \left. - \log\left(\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}\right)\right] \\
= & \sum_{p=1}^{n}(\beta_{i1}X_p + \beta_{i2}G_p + \beta_{i3}X_p : G_p)\sum_{k=0}^{K_i} Y_{pik}k - \sum_{p=1}^{n}\sum_{k=0}^{K_i} Y_{pik}\sum_{l=0}^{k}\beta_{i0l} \\
& - \sum_{p=1}^{n}\sum_{k=0}^{K_i}\log\left(\sum_{j=0}^{K_i} e^{\sum_{l=0}^{j}(\beta_{i0l}+\beta_{i1}X_p+\beta_{i2}G_p+\beta_{i3}X_p:G_p)}\right).
\end{aligned}$$

Analogously as in Section 2.1.1, parameter estimates are obtained as a solution of estimating equations $\frac{\partial l_i}{\partial \gamma_{ij}} = 0$, for $j = 1, \dots, K_i + 3$, where $\boldsymbol{\gamma}_i = (\beta_{i01}, \dots, \beta_{i0K_i}, \beta_{i1}, \beta_{i2}, \beta_{i3})$ is set of $K_i + 3$ parameters of item i. Again, the solution does not have a closed form and numerical approaches, such as iteratively re-weighted least squares or Fisher scoring algorithm, need to be applied.

As was noted before, also in this case, the maximum likelihood estimators of the item parameters have pleasant properties including consistency and asymptotic normality. Asymptotic covariance matrix and standard errors can be estimated using the inverse of the Hessian matrix.

2.1.3 DIF detection

The natural way of testing for DIF using ordinal regression models discussed in this section is to use the likelihood ratio test. Analogously as described in Section 1.3, the likelihood-ratio test measures difference in the log-likelihood l_{i1} of model M_{i1} and the log-likelihood l_{i0} of its submodel M_{i0} for the item i. The submodel M_{i0} sets some group-specific parameters, i.e., β_{i2} and/or β_{i3}, to zeros, while these parameters are freely estimated in the larger model M_{i1} . The corresponding test statistic has the following form:

$$LR_i = -2\left(l_{i0} - l_{i1}\right).$$

The LR_i statistic asymptotically has χ^2-distribution under the submodel M_{i0}:

$$LR_i \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(\mathrm{df}_{i1} - \mathrm{df}_{i0}), \tag{2.6}$$

where df_{i1} and df_{i0} are numbers of parameters in the larger model M_{i1} and its submodel M_{i0}.

2.1.4 Implementation

Group-specific logit models for DIF detection among ordinal responses (2.2) and (2.5) are implemented in the **difNLR R** package via function `difORD()`. The full syntax of the `difORD()` function is

```
difORD(
  Data, group, focal.name, model = "adjacent", type = "both",
  match = "zscore", anchor = NULL, purify = FALSE, nrIter = 10,
  p.adjust.method = "none", parametrization = "irt", alpha = 0.05
)
```

Description of the arguments of the function can be found in Table A.2 and we describe some of them here in more detail. To detect DIF among ordinal data using the `difORD()` function, the user needs to provide four pieces of information: 1. the ordinal data set, 2. the group membership vector, 3. the indication of the focal group, and 4. the model to be fitted.

Data. `Data` takes a similar format as used for the `difNLR()` function (see Section 1.4), however, rows represent ordinaly scored respondents' answers instead of dichotomous, such as Likert scale 1–5, represented by numerical values. Specification of the `group` and `focal.name` arguments remains the same.

Data generation. Data generator `genNLR()` is able to generate ordinal data using the adjacent category logit model (2.5) by setting `itemtype = "ordinal"`. For polytomous items (ordinal or nominal), sets of parameters `a` and `b` have the form of matrices as it was the case for dichotomous data but each column now represents parameters of partial scores (or distractors). For example, to generate an item with 4 partial scores (i.e., 0–3), the user needs to provide 3 sets of discrimination and difficulty parameters. As was noted in models specifications, the parameters for minimal partial scores (i.e., 0; or correct answer in the case of nominal data) do not need to be specified because their probabilities are calculated as a complement to the sums of the partial scores probabilities.

To illustrate usage of the `difORD()` function, we created an example ordinal dataset with 5 items, each scored with a range of 0–4. We first generated discrimination parameters a and difficulties b from a uniform distribution for partial scores $k = 1, \ldots, 4$ for each item. In an adjacent category logit model (2.5), parameter b_{ik} corresponds to an ability level for which the response categories k and $k-1$ intersect in item i. For this reason and to create well-functioning items, parameters b_{ik} are sorted so that $b_{ik} < b_{i(k+1)}$. The parameters are set the same for both the reference and the focal group.

```
set.seed(42)
# discrimination
a <- matrix(rep(runif(5, 0.25, 1), 8), ncol = 8)
# difficulty
b <- t(sapply(1:5, function(i) rep(sort(runif(4, -1, 1)), 2)))
```

For the first two items we introduce uniform and non-uniform DIF respectively.

```
b[1, 5:8] <- b[1, 5:8] + 0.1
a[2, 5:8] <- a[2, 5:8] - 0.2
```

Using parameters `a` and `b` of the adjacent category logit model, we generate ordinal data with a total sample size of 1,000 (500 observations per group). The first 5 columns of dataset `DataORD` represent ordinaly scored items, while the last column represents a group membership variable.

```
DataORD <- genNLR(N = 1000, itemtype = "ordinal", a = a, b = b)
summary(DataORD)
Item1   Item2   Item3   Item4   Item5       group
 0:488   0:376   0:417   0:530   0:556   Min.   :0.0
 1:229   1:237   1:331   1:226   1:253   1st Qu.:0.0
 2:150   2:195   2:170   2:129   2:123   Median :0.5
 3: 93   3:114   3: 71   3: 83   3: 47   Mean   :0.5
 4: 40   4: 78   4: 11   4: 32   4: 21   3rd Qu.:1.0
                                         Max.   :1.0
```

Model. The last input of the `difORD()` function which needs to be specified is `model`. It offers two possibilities. With an option `model = "cumulative"` the cumulative logit model (2.2) is fitted, while with an option `model = "adjacent"` (default) DIF detection is performed using the adjacent category logit model

(2.5). The parameters for both models are estimated via `vglm()` function from the **VGAM** package (Yee, 2010) using the maximum likelihood estimation with the iteratively re-weighted least squares algorithm.

DIF detection with the cumulative logit model

In this part we exemplify usage of the `difORD()` function to fit the group-specific cumulative logit model among ordinal data and to test for DIF. The `group` argument is introduced here by specifying the name of the group membership variable in the `DataORD` dataset, i.e., `group = "group"`. Knowledge is estimated as a standardized total score, i.e., standardized sum of all item scores.

```
(fit5 <- difORD(DataORD, group = "group", focal.name = 1,
                model = "cumulative"))
Detection of both types of Differential Item
Functioning for ordinal data using cumulative logit
regression model

Likelihood-ratio Chi-square statistics

Item purification was not applied
No p-value adjustment for multiple comparisons

      Chisq-value P-value
Item1  7.4263      0.0244 *
Item2 13.4267      0.0012 **
Item3  0.6805      0.7116
Item4  5.6662      0.0588 .
Item5  2.7916      0.2476

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Items detected as DIF items:
 Item1
 Item2
```

Output provides test statistics for the likelihood ratio test (2.6), corresponding p-values, and the set of items which were detected as functioning differently. Items 1 and 2 are correctly identified as DIF items.

Similarly as was shown for the `difNLR()` function, the ICCs can be imaged with the `plot()` method. Besides the ICCs, the method `plot()` for the cumulative logit model also offers the plot of item cumulative probabilities. This can be achieved using `plot.type = "cumulative"`, while with `plot.type = "category"` the ICCs are shown. The plot of cumulative probabilities shows only 4 partial scores and does not show the cumulative probability of $\mathsf{P}(Y_{pi} \geq 0)$ since it is always equal to 1. Note that category probability of the highest score corresponds to its cumulative probability (Figure 2.1).

```
plot(fit5, item = "Item1", plot.type = "cumulative")
plot(fit5, item = "Item1", plot.type = "category")
```

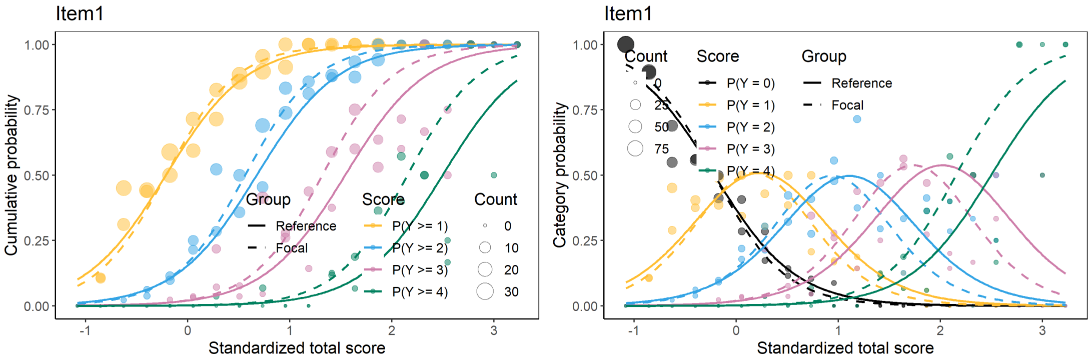


Figure 2.1: Cumulative probabilities and the ICCs of item 1 under the cumulative logit model.

DIF detection with adjacent logit model

We illustrate here the fitting of the adjacent category logit model for DIF detection using the `difORD()` function. The group argument is now introduced by specifying the identifier of a group membership variable in `Data` (i.e., `group = 6`).

```
(fit6 <- difORD(DataORD, group = 6, focal.name = 1,
                model = "adjacent"))
Detection of both types of Differential Item
Functioning for ordinal data using adjacent category
logit model

Likelihood-ratio Chi-square statistics

Item purification was not applied
No p-value adjustment for multiple comparisons

      Chisq-value P-value
Item1  8.9024      0.0117 *
Item2 12.9198      0.0016 **
Item3  1.0313      0.5971
Item4  4.3545      0.1134
Item5  2.3809      0.3041

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Items detected as DIF items:
 Item1
 Item2
```

Output again provides test statistics for the likelihood ratio test, corresponding p-values, and the set of items which were detected as functioning differently. Items 1 and 2 are correctly identified as DIF items. The ICCs can again be rendered using the `plot()` method (Figure 2.2).

```
plot(fit6, item = fit6$DIFitems)
```

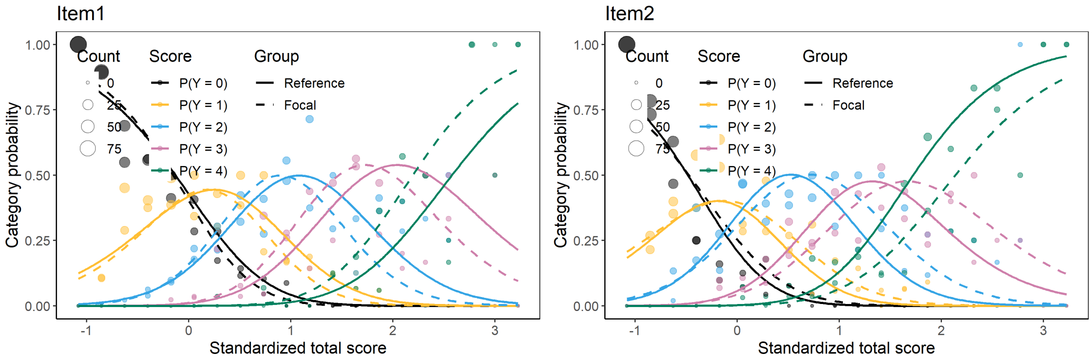


Figure 2.2: The ICCs of DIF items with the adjacent category logit model.

Common features

The `difORD()` function offers the possibility to specify parametrization of the item parameters using an argument `parametrization`. By default, the IRT parametrization as stated in (2.2) and (2.5) is utilized using `parametrization = "irt"`, but the classical intercept-slope parametrization (`parametrization = "classic"`) as stated in (2.1) and (2.4) may be applied, i.e., $\beta_{i0k} + \beta_{i1}X_p + \beta_{i2}G_p + \beta_{i3}X_p : G_p$ instead of $a_{iG_p}(X_p - b_{ikG_p})$. The DIF detection is the same as with the IRT parametrization, the only difference can be found in parameter estimates and estimates of their standard errors:

```
fit6a <- difORD(DataORD, group = 6, focal.name = 1, model = "adjacent",
                parametrization = "classic")
# coefficients with IRT parametrization
round(coef(fit6, SE = TRUE)[[1]], 3)
            b1    b2    b3    b4     a
estimate 0.013 0.603 1.500 2.500 1.776
SE       0.064 0.073 0.094 0.141 0.114
          bDIF1  bDIF2  bDIF3  bDIF4  aDIF
estimate -0.042 -0.121 -0.240 -0.374 0.273
SE        0.058  0.055  0.081  0.130 0.115
# coefficients with classic intercept-slope parametrization
round(coef(fit6a, SE = TRUE)[[1]], 3)
(Interc):1 (Interc):2 (Interc):3 (Interc):4      x  group  x:group
estimate        -0.023        -1.070        -2.664        -4.441
SE               0.114         0.141         0.212         0.316
             x group x:group
estimate 1.776 0.082   0.273
SE       0.114 0.109   0.115
```

Note that estimated discrimination for the reference group (parameter `a`) corresponds to the effect of the matching criterion `x`, and in both cases their value is 1.776 for item 1. The same holds for the difference in discrimination and the effect of interaction between the matching criterion and the group membership variable.

Similarly to the `difNLR()` function presented in Section 1.4.1, the `difORD()` function also offers fit measures provided by `AIC()`, `BIC()`, and `logLik()` S3 methods. Function `difORD()` further provides item purification via an argument `purify = TRUE` or corrections for multiple comparisons of user's choice specified via the argument `p.adjust.method` (see also Chapter 4).

2.2 Group-specific model for nominal items

When responses are nominal (e.g., multiple choice), DDF detection can be performed with the multinomial model, which is a baseline category logit model (see, e.g., Agresti, 2010, Section 7.1). Considering that $k = 0, \dots, K_i$ possible option choices are offered, with $k = 0$ being the correct answer and other ones distractors, the multinomial model pairs each response category with a baseline category (for example, with the correct answer in case of multiple-choice items):

$$\log \frac{\mathsf{P}(Y_{pi} = k|X_p, G_p)}{\mathsf{P}(Y_{pi} = 0|X_p, G_p)} = \beta_{i0k} + \beta_{i1k} X_p + \beta_{i2k} G_p + \beta_{i3k} X_p : G_p,$$

where $k = 1, \dots, K_i$. The K_i parameters β_{i0k} are the intercepts for category k within item i. The K_i parameters β_{i1k} are the slopes for category k withing item i. The K_i parameters β_{i2k} are associated with the effect of group membership on probability of selecting option k, while the K_i parameters β_{i3k} are related to the effect of category-specific interaction between this group membership variable and the matching criterion X_p. In contrast to ordinal models, all parameters are category-specific. Therefore, the group-specific multinomial model for DDF detection includes $4K_i$ item parameters in total.

Using recursive formula, it can be shown that the probability of choosing distractor k is given by

$$\mathsf{P}(Y_{pi} = k|X_p, G_p) = \frac{e^{\beta_{i0k} + \beta_{i1k} X_p + \beta_{i2k} G_p + \beta_{i3k} X_p : G_p}}{\sum_{l=0}^{K_i} e^{\beta_{i0l} + \beta_{i1l} X_p + \beta_{i2l} G_p + \beta_{i3l} X_p : G_p}}, \tag{2.7}$$

while for the correct answer $k = 0$ the parameters are set to zero, i.e., $\beta_{i00} + \beta_{i10} X_p + \beta_{i20} G_p + \beta_{i30} X_p : G_p = 0$, and thus

$$\mathsf{P}(Y_{pi} = 0|X_p, G_p) = \frac{1}{\sum_{l=0}^{K_i} e^{\beta_{i0l} + \beta_{i1l} X_p + \beta_{i2l} G_p + \beta_{i3l} X_p : G_p}}.$$

Parametrization. Similarly as before, reparametrization using the IRT notation can be applied:

$$\mathsf{P}(Y_{pi} = k|X_p, G_p) = \frac{e^{a_{ikG_p}(X_p - b_{ikG_p})}}{\sum_{l=0}^{K_i} e^{a_{ilG_p}(X_p - b_{ilG_p})}}, \tag{2.8}$$

for $k = 1, \dots, K_i$. The $2K_i$ parameters $b_{ikG_p} = b_{ik} + b_{ik\mathrm{DIF}} G_p$, $G_p \in \{0, 1\}$, are group-specific locations of the response probability curves intersections with the response probability curve of the baseline category, i.e., the probability curve of the correct answer. The $2K_i$ parameters $a_{ikG_p} = a_{ik} + a_{ik\mathrm{DIF}} G_p$, $G_p \in \{0, 1\}$, are group-specific slopes of the category probabilities. Mutual relationship between parameters using the classical intercept-slope notation (2.7) and the notation based on IRT (2.8) is similar as for the group-specific ordinal models for DIF detection in (2.3), however, now all the parameters are category-specific. As before, we will further stick with the classical intercept-slope parametrization as it is more convenient for estimation and implementation, noting that parameters using the IRT notation together with standard errors can be again obtained with the delta method (Doob, 1935).

2.2.1 Estimation and asymptotic properties

For item i and respondent p, let $(Y_{pi0}, \dots, Y_{piK})$ be binary indicators of the responses defined as $Y_{pik} = 1$ when respondent p selected option k as an answer in item i and $Y_{pik} = 0$ otherwise. The corresponding likelihood function has the following form

$$L_i = \prod_{p=1}^{n} \prod_{k=0}^{K_i} \left\{ \mathsf{P}(Y_{pi} = k | X_p, G_p) \right\}^{Y_{pik}},$$

where $\mathsf{P}(Y_{pi} = k|X_p, G_p)$ is given by (2.7). The corresponding log-likelihood function is then given as:

$$\begin{aligned} l_i &= \sum_{p=1}^{n} \sum_{k=0}^{K_i} Y_{pik} \log\left(\mathsf{P}(Y_{pi} = k|X_p, G_p)\right) \\ &= \sum_{p=1}^{n} \sum_{k=0}^{K_i} Y_{pik} \log\left(\frac{e^{\beta_{i0k}+\beta_{i1k}X_p+\beta_{i2k}G_p+\beta_{i3k}X_p:G_p}}{\sum_{l=0}^{K_i} e^{\beta_{i0l}+\beta_{i1l}X_p+\beta_{i2l}G_p+\beta_{i3l}X_p:G_p}}\right) \\ &= \sum_{p=1}^{n} \sum_{k=0}^{K_i} Y_{pik} \left(\beta_{i0k} + \beta_{i1k}X_p + \beta_{i2k}G_p + \beta_{i3k}X_p : G_p\right) \\ &\quad - \sum_{p=1}^{n} \sum_{k=0}^{K_i} Y_{pik} \log\left(\sum_{l=0}^{K_i} e^{\beta_{i0l}+\beta_{i1l}X_p+\beta_{i2l}G_p+\beta_{i3l}X_p:G_p}\right). \end{aligned}$$

The log-likelihood function l_i is concave and the solution of the related estimating equations gives maximum likelihood estimates of the item parameters (Agresti, 2010, Section 7.1.4). As the solution of the estimating equations does not have a closed form, numerical approaches, such as the Newton-Raphson method or neural networks (Venables & Ripley, 2002), need to be applied.

As noted in previous sections, the maximum likelihood estimators of the item parameters are consistent and have asymptotically normal distribution. To estimate the asymptotic covariance matrix, the inverse of the corresponding Hessian matrix may be used.

2.2.2 DDF detection

The natural way how to test for DDF, i.e., group differences in item parameters, is to perform the likelihood ratio test as was described in Section 2.1.3.

2.2.3 Implementation

The group-specific multinomial model for DDF detection (2.8) is provided in the **`difNLR`** package by function `ddfMLR()`. The full syntax of the `ddfMLR()` function is as follows:

```
ddfMLR(
  Data, group, focal.name, key, type = "both", match = "zscore",
  anchor = NULL, purify = FALSE, nrIter = 10,
  p.adjust.method = "none", parametrization = "irt", alpha = 0.05
)
```

Description of all arguments of the `ddfMLR()` function can be found in Table A.3. To detect DDF among nominal data using the `ddfMLR()` function, the user needs to provide four pieces of information: 1. the unscored data set, 2. the key of correct answers, 3. the group membership vector, and 4. the indication of the focal group. The parameters are estimated via `multinom()` function from the **`nnet`** package (Venables & Ripley, 2002).

Data. The format of `Data` argument is similar to previously described functions. However, rows here represent respondents' unscored answers (for example, in ABCD format or as numerical values without ordering). The `group` and `focal.name` arguments are specified as in `difNLR()` or `difORD()` functions.

Data generation. Data generator `genNLR()` can be used to generate nominal data using a multinomial model (2.8) by setting `itemtype = "nominal"`. Specification of arguments `a` and `b` is the same as for ordinal items, however, these now represent parameters for distractors (incorrect answers).

To create an illustrative example dataset of nominal data, we first generate discrimination a and difficulty b parameters from a uniform distribution for distractors of 10 items. The parameters are set equal for the reference and the focal group. For the first 5 items, we consider only two distractors (i.e., three item choices in total). For the last 5 items, we consider three distractors (i.e., four item choices in total).

```
set.seed(42)
# discrimination
a <- matrix(rep(runif(30, -2, -0.5), 2), ncol = 6)
a[1:5, c(3, 6)] <- NA
# difficulty
b <- matrix(rep(runif(30, -3, 1), 2), ncol = 6)
b[1:5, c(3, 6)] <- NA
```

For item 1, we introduce DDF by difference in discrimination and for item 6 by difference in difficulty.

```
a[1, 4] <- a[1, 1] - 1; a[1, 5] <- a[1, 2] + 1
b[6, 4] <- b[6, 1] - 1; b[6, 5] <- b[6, 2] - 1.5
```

Finally, we generate nominal data with 500 observations in each group, i.e., 1,000 in total. The first 10 columns of the generated `DataDDF` dataset represent the unscored answers of respondents and the last column describes a group membership variable.

```
DataDDF <- genNLR(N = 1000, itemtype = "nominal", a = a, b = b)
head(DataDDF)
  Item1 Item2 Item3 Item4 Item5 Item6 Item7 Item8 Item9 Item10 group
1     B     B     C     A     C     B     B     D     B      B     0
2     C     A     B     A     C     C     B     B     C      C     0
3     B     C     C     B     C     C     B     C     B      D     0
4     B     A     C     A     C     B     A     B     B      B     0
5     B     B     C     B     C     B     A     C     A      B     0
6     B     A     A     A     A     B     B     A     A      A     0
```

The correct answers in the generated dataset are denoted by A for each item; the key is hence a vector of As with a length of 10.

We now have all the necessary inputs to fit the multinomial model (2.8) using the `ddfMLR()` function. The `group` argument is introduced here by specifying the name of group membership variable in `Data` (i.e., `group` = ”group”). For the generated data, the total score is calculated as a number of correct answers (i.e., number of As on a given row) and the matching criterion is then its standardized value (Z-score).

```
(fit7 <- ddfMLR(DataDDF, group = "group", focal.name = 1,
                key = rep("A", 10)))
Detection of both types of Differential Distractor
Functioning using multinomial log-linear regression model

Likelihood-ratio chi-square statistics

Item purification was not applied
No p-value adjustment for multiple comparisons

       Chisq-value P-value
Item1  29.5508      0.0000 ***
Item2   1.1136      0.8921
Item3   1.0362      0.9043
Item4   4.1345      0.3881
Item5   7.4608      0.1134
Item6  47.0701      0.0000 ***
Item7   1.3285      0.9701
Item8   2.3629      0.8835
Item9  10.4472      0.1070
Item10  3.5602      0.7359

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Items detected as DDF items:
 Item1
 Item6
```

The output again summarizes the statistics of the likelihood ratio test of a submodel, corresponding *p*-values, and the set of items identified as DDF. As expected, items 1 and 6 are detected as DDF. Their ICCs can be displayed with a `plot()` method while the name of the reference and focal group can be modified via `group.names` argument (Figure 2.3). This option is also available for functions `difNLR()` and `difORD()` and their plotting methods.

```
plot(fit7, item = fit7$DDFitems, group.names = c("Group 1", "Group 2"))
```

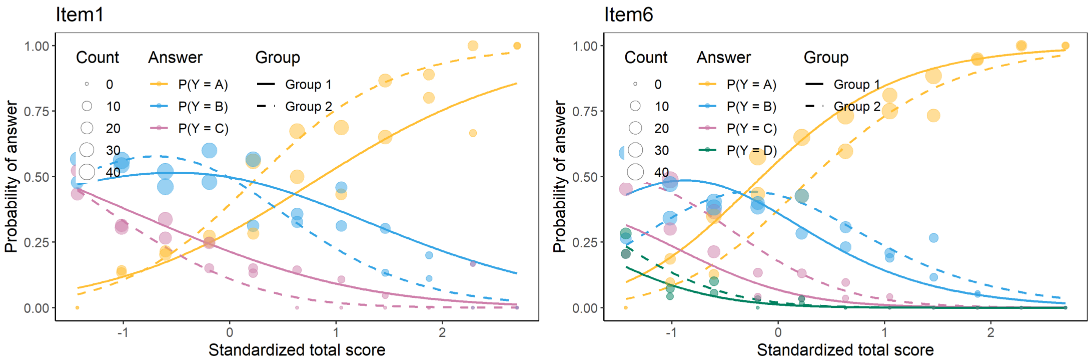


Figure 2.3: The ICCs of DDF items with the multinomial model.

Similarly as for the `difNLR()` and `difORD()` functions, item fit measures are offered via `AIC()`, `BIC()`, and `logLik()` S3 methods. Parameter estimates can be obtained using the method `coef()`. Parametrization can be set using the `parametrization` argument in the `ddfMLR()` function while the options are the same as for the `difORD()` function from Section 2.1.4. The `ddfMLR()` function also offers item purification and corrections for multiple comparisons.

3. Nonparametric comparison of regression curves

DIF and its detection is closely connected to a general problem of describing relationship between respondents' answers $\boldsymbol{Y_i} = (Y_{1i}, \ldots, Y_{ni})$ to item i and respondents' observed ability $\boldsymbol{X} = (X_1, \ldots, X_n)$. This relationship can be generally described by the regression function m_i for item i such as

$$\boldsymbol{Y_i} = m_i(\boldsymbol{X}) + \boldsymbol{\epsilon_i},$$

where $\mathsf{E}(\boldsymbol{\epsilon_i}|\boldsymbol{X}) = \boldsymbol{0}$. In this chapter, we focus on binary outcomes $\boldsymbol{Y_i}$, in which case this relationship can be reformulated as

$$\mathsf{E}(\boldsymbol{Y_i}|\boldsymbol{X}) = \mathsf{P}(\boldsymbol{Y_i}|\boldsymbol{X}) = m_i(\boldsymbol{X}).$$

Chapters 1 and 2 considered parametric model for the probability of correct answer or partial score. However, any parametric approach runs the risk of simplifying the true underlying model and omitting important information. It may be the case, that the underlying model is too complicated and/or no parametric model can be assumed. In such cases, the nonparametric smoothing methods may be a flexible tool in analyzing unknown regression function $m_i(x)$ (Härdle, 1990).

In context of multi-item measurements, Mokken (1971, Chapter 4) proposed two nonparametric IRT models for binary items: The monotone homogeneity model and the double-monotonicity model. Ramsay (1991) suggested a kernel smoothing method to estimate the ICCs using Nardaraya-Watson weights and estimates of ability based on rank approach.

Several methods incorporate kernel smoothing into DIF detection procedure including for example kernel smoothed SIBTEST (Douglas, Stout, & DiBello, 1996) or TestGraf, a graphical DIF method with kernel smoothing for estimating the conditional probability of correct answers related to proficiency estimates (Bolt & Gierl, 2006; Ramsay, 2000).

In this chapter, we present a new approach for DIF detection using the kernel smoothing estimates of the ICCs and their comparison, as proposed in Hladká and Martinková (2021). While many authors deal with the topic of nonparametric comparison of the regression curves including Dette and Neumeyer (2001); Hall and Hart (1990); Neumeyer and Dette (2003), here we offer and study a method which is based on general statistic proposed by Srihera and Stute (2010). We adapted and expanded their approach to estimate the ICCs and to identify DIF.

3.1 Method specification

Building on the method suggested by Srihera and Stute (2010), we propose a kernel smoothing estimation of the ICCs of item i for the reference and focal group based on nearest neighbors and test statistic to detect DIF. Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be the sets of respondents from the reference and focal group of size n_0 and n_1, respectively, where n_0+n_1 is a total sample size. Let further Y_{pi} be a binary response on item i by respondent p and X_p his/her observed ability such as standardized total test score or other matching criterion. Further, let $\hat{F}_{i0}(x)$ and $\hat{F}_{i1}(x)$ be empirical distribution functions of the ability variable X_p for the reference and focal group, i.e.:

$$\hat{F}_0(x) = \frac{1}{n_0} \sum_{p \in \mathcal{P}_0} \mathbf{1}_{X_p \leq x}, \text{ and } \hat{F}_1(x) = \frac{1}{n_1} \sum_{p \in \mathcal{P}_1} \mathbf{1}_{X_p \leq x}.$$

The nearest neighbor estimate of the ICCs of item i for the reference and focal group then takes the following form:

$$\begin{aligned}
\hat{m}_{i0}(x) &= \sum_{p \in \mathcal{P}_0} Y_{pi} W_{pi0}(x), \text{ with weights } W_{pi0}(x) = \frac{K\left(\frac{\hat{F}_0(X_p) - \hat{F}_0(x)}{h}\right)}{\sum\limits_{k \in \mathcal{P}_0} K\left(\frac{\hat{F}_0(X_k) - \hat{F}_0(x)}{h}\right)}, \\
\hat{m}_{i1}(x) &= \sum_{p \in \mathcal{P}_1} Y_{pi} W_{pi1}(x), \text{ with weights } W_{pi1}(x) = \frac{K\left(\frac{\hat{F}_1(X_p) - \hat{F}_1(x)}{h}\right)}{\sum\limits_{k \in \mathcal{P}_1} K\left(\frac{\hat{F}_1(X_k) - \hat{F}_1(x)}{h}\right)},
\end{aligned} \tag{3.1}$$

where $K(u)$ is a kernel function which satisfies assumptions standard in literature (Srihera & Stute, 2010, p. 2042):

(i) K is symmetric and non-negative for $u \in \mathbb{R}$; K is non-decreasing for $u < 0$,

(ii) $\int K(u)\, \mathrm{d}u = 1$,

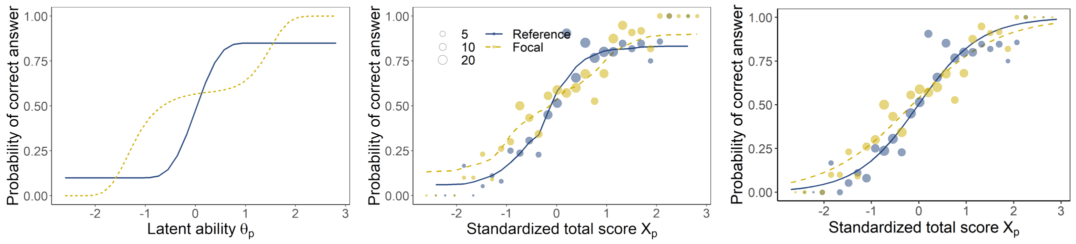
(iii) K has compact support and is twice continuously differentiable.

Several types of kernel functions may be used, for example, Epanechnikov $K(u) = \frac{3}{4}(1 - u^2), |u| \leq 1$ (Epanechnikov, 1969) or uniform $K(u) = \frac{1}{2}, |u| \leq 1$. A bandwidth parameter h needs to meet the assumptions of $nh^3 \rightarrow \infty$ and $nh^4 \rightarrow 0$ (see Srihera & Stute, 2010, p. 2042). Therefore, h should take the value of $n^{-\zeta}$, where $\zeta \in \left(\frac{1}{4}, \frac{1}{3}\right)$ and n has the order of n_0 and n_1.

The main advantage of the kernel smoothing estimations of the ICCs is that it does not assume underlying parametric model, thus it can be used even in situations when the ICCs are more complicated. For simple illustration, let's consider the ICCs with several inflection points (Figure 3.1A). Then, the estimate based on nearest neighbour method (3.1) using generated dichotomous data for the two groups (Figure 3.1B) better corresponds to the shape of the true curves than the estimate based on simple logistic regression method (7) (Figure 3.1C).

3.1.1 Test statistic

It is a common phenomenon that ability distributions for the reference and the focal group have a different support. Srihera and Stute (2010, p. 2040) suggested a way to deal with this issue by averaging all values of the matching criterion X_p and comparing the ICCs on this new support. Let $\bar{X}_{p_0 p_1} = \frac{X_{p_0} + X_{p_1}}{2}$ for $p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1$ be the mean value of the observed ability of respondent p_0



(A) Underlying ICCs. (B) Nearest neighbours. (C) Logistic regression.

Figure 3.1: Example of the nearest neighbour and logistic regression estimates of the ICCs. Curves are accompanied by points representing empirical probabilities.

from the reference group and of respondent p_1 from the focal group. The test statistic then takes the following form:

$$\hat{T}_i = \frac{1}{n_0 n_1} \sum_{p_0 \in \mathcal{P}_0} \sum_{p_1 \in \mathcal{P}_1} W_i\left(\bar{X}_{p_0 p_1}\right)\left[\hat{m}_{i0}\left(\bar{X}_{p_0 p_1}\right) - \hat{m}_{i1}\left(\bar{X}_{p_0 p_1}\right)\right], \tag{3.2}$$

where W_i is a twice continuously differentiable weight function of item i (see Srihera & Stute, 2010).

Asymptotic properties

The asymptotic variance of the test statistic. The asymptotic variance of the test statistic (3.2) under the null hypothesis, i.e., when there is no difference between the two regression curves m_{i0} and m_{i1}, is given by

$$\sigma_i^2 = (1-\lambda)\rho_{i0}^2 + \lambda\rho_{i1}^2, \tag{3.3}$$

where

$$\begin{aligned}
\rho_{i0}^2 &= \int \sigma_{i0}(x) W_i^2(x) \frac{e(x)}{f_0(x)} E(\mathrm{d}x) < \infty, \quad \sigma_{i0}(x) = m_{i0}(x)(1 - m_{i0}(x)),\\
\rho_{i1}^2 &= \int \sigma_{i1}(x) W_i^2(x) \frac{e(x)}{f_1(x)} E(\mathrm{d}x) < \infty, \quad \sigma_{i1}(x) = m_{i1}(x)(1 - m_{i1}(x)),
\end{aligned}$$

and $f_0(x)$, $f_1(x)$, and $e(x)$ are twice continuously differentiable densities of the observed ability for the reference and focal group and for their averaged values, while $E(x)$ is their cumulative distribution function. Finally, $\lambda \in (0,1)$ is determined by $\frac{n_0}{n_0+n_1} \to \lambda$ and $\frac{n_1}{n_0+n_1} \to 1-\lambda$ (see Srihera & Stute, 2010).

Further, Srihera and Stute (2010, p. 2044) proposed an estimate of the asymptotic variance of the test statistic (3.2) of the following form:

$$\begin{aligned}
\hat{\sigma}_i^2 = {} & \frac{1}{n_0+n_1} \sum_{p_0 \in \mathcal{P}_0} \left(Y_{p_0 i} - \hat{m}_{i0}(X_{p_0})\right)^2 \left[\sum_{k \in \mathcal{P}_0} \sum_{l \in \mathcal{P}_1} W_i\left(\bar{X}_{kl}\right) W_{p_0 i}\left(\bar{X}_{kl}\right)\right]^2\\
& + \frac{1}{n_0+n_1} \sum_{p_1 \in \mathcal{P}_1} \left(Y_{p_1 i} - \hat{m}_{i1}(X_{p_1})\right)^2 \left[\sum_{k \in \mathcal{P}_0} \sum_{l \in \mathcal{P}_1} W_i\left(\bar{X}_{kl}\right) W_{p_1 i}\left(\bar{X}_{kl}\right)\right]^2.
\end{aligned} \tag{3.4}$$

The $\hat{\sigma}_i^2$ is actually the sum of weighted estimates of the conditional variances $\operatorname{var}(Y_{p_0 i}|X_p)$ and $\operatorname{var}(Y_{p_1 i}|X_p)$. In case of binary data, it might be more suitable to replace the terms $(Y_{p_0 i} - \hat{m}_{i0}(X_{p_0}))^2$ and $(Y_{p_1 i} - \hat{m}_{i1}(X_{p_1}))^2$ by

$$\hat{\sigma}_{i0}(X_{p_0}) = \hat{m}_{i0}(X_{p_0})\left(1 - \hat{m}_{i0}(X_{p_0})\right),$$
$$\hat{\sigma}_{i1}(X_{p_1}) = \hat{m}_{i1}(X_{p_1})\left(1 - \hat{m}_{i1}(X_{p_1})\right),$$

respectively. Therefore, for the case of binary data, we propose a new estimate of the asymptotic variance of the test statistic in the following form:

$$\begin{aligned}\hat{\sigma}_i^2 = &\frac{1}{n_0+n_1}\sum_{p_0\in\mathcal{P}_0}\hat{\sigma}_{i0}(X_{p_0})\left[\sum_{k\in\mathcal{P}_0}\sum_{l\in\mathcal{P}_1}W_i\left(\bar{X}_{kl}\right)W_{p_0 i}\left(\bar{X}_{kl}\right)\right]^2\\ &+\frac{1}{n_0+n_1}\sum_{p_1\in\mathcal{P}_1}\hat{\sigma}_{i1}(X_{p_1})\left[\sum_{k\in\mathcal{P}_0}\sum_{l\in\mathcal{P}_1}W_i\left(\bar{X}_{kl}\right)W_{p_1 i}\left(\bar{X}_{kl}\right)\right]^2.\end{aligned} \tag{3.5}$$

To evaluate which estimate of the asymptotic variance, i.e., either (3.4) or (3.5), is more appropriate to use, a short simulation study was performed. Both estimates yielded similar results (not shown) which were comparable to the empirical variance of the test statistic. We decided to use (3.5) further as it takes into account binary nature of the data.

Asymptotic distribution. It can be shown, under the conditions specified above and when the null hypothesis holds, that the test statistic (3.2) normalized by $\hat{\sigma}_i$ asymptotically follows a standard normal distribution (for details see Srihera & Stute, 2010, Theorems 1 and 2), i.e.:

$$\frac{\sqrt{N}\hat{T}_i}{\hat{\sigma}_i} \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,1),$$

where $N = \frac{n_0 n_1}{n_0+n_1}$.

Support of the test statistic. The new support, defined by all combinations of the matching criterion from the two groups, is a set of size equal to $n_0 \cdot n_1$ which may slow down data manipulation in the statistical software and method then may become time challenging and memory demanding, especially for larger sample sizes. We thus propose and use another technique to calculate common support of the test statistic which is also based on averaged points. This method proceeds as follows: First, the common support is calculated as in original approach. Then, the empirical weights of unique values of averaged points are calculated. Finally, the new support is created by generating a fixed-sized random sample from unique values of the common support using these weights. This new approach provides the support of the test statistic covering ability distribution of both groups while it keeps the data manipulation effective. In case of using reduced support, it should be noted, that the original size of the product, i.e., $n_0 \cdot n_1$, needs to be replaced by the size of the newly defined support set.

3.1.2 Weight function

The choice of the weight function W_i is crucial as it may have a great impact on power of the test (Srihera & Stute, 2010). In this work, we consider several options of the weight function.

Fixed weights

First, we take non-informative fixed weights, that is

$$W_i(x) = 1 \ \ \forall x. \tag{3.6}$$

Optimal weights

Second, we consider optimal weight function $W_i(x)$ which maximizes the local power of the test statistic (3.2), as proposed by Srihera and Stute (2010). We adapt their approach for our case of binary data and for comparison of the ICCs.

Under the local alternative hypotheses, i.e., $m_{i0} = m_{i1} + \frac{cs_i}{N}, c \neq 0$, where s_i is a difference function, the following holds:

$$\frac{\sqrt{N}\hat{T}_i}{\hat{\sigma}_i} \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}\left(\frac{\mu_i}{\sigma_i}, 1\right),$$

where $\mu_i = -\int W_i(x)\left(m_{i0}(x) - m_{i1}(x)\right)E(\mathrm{d}x)$ and σ_i^2 is a variance of the test statistic (3.3) (Srihera & Stute, 2010, Theorem 2). The asymptotic power is then given by

$$\mathsf{P}\left(\left|\frac{\sqrt{N}\hat{T}_i}{\hat{\sigma}_i}\right| \geq q_{1-\frac{\alpha}{2}}\right) \simeq 1 - \phi\left(\frac{\mu_i}{\sigma_i} + q_{1-\frac{\alpha}{2}}\right) + \phi\left(\frac{\mu_i}{\sigma_i} - q_{1-\frac{\alpha}{2}}\right), \tag{3.7}$$

which is an increasing function of $\left|\frac{\mu_i}{\sigma_i}\right|$. Thus, the weight function which maximizes the asymptotic power (3.7) is the one which maximizes the term $\left|\frac{\mu_i}{\sigma_i}\right|$. This is equivalent to maximizing the term $\frac{\mu_i^2}{\sigma_i^2}$. Srihera and Stute (2010) further showed the form of $W_i(x)$ for which the term $\frac{\mu_i^2}{\sigma_i^2}$ is maximized.

For our case of binary data, the term $\frac{\mu_i^2}{\sigma_i^2}$ has form

$$\frac{\mu_i^2}{\sigma_i^2} = \frac{\left[\int W_i(x)s_i(x)E(\mathrm{d}x)\right]^2}{\int\left[(1-\lambda)\sigma_{i0}(x)\frac{e(x)}{f_0(x)} + \lambda\sigma_{i1}(x)\frac{e(x)}{f_1(x)}\right]W_i^2(x)E(\mathrm{d}x)}$$

and it is maximized for

$$W_i(x) = \frac{s_i(x)}{(1-\lambda)\sigma_{i0}(x)\frac{e(x)}{f_0(x)} + \lambda\sigma_{i1}(x)\frac{e(x)}{f_1(x)}}.$$

Differences between the ICCs cannot be satisfactorily described by a generic function such as a polynomial. Therefore, in our case, it is possible to use $s_i(x) = m_{i0}(x) - m_{i1}(x)$, i.e., the true difference between the two ICCs, which results into the optimal weights in form of

$$W_i(x) = \frac{m_{i0}(x) - m_{i1}(x)}{(1-\lambda)\sigma_{i0}(x)\frac{e(x)}{f_0(x)} + \lambda\sigma_{i1}(x)\frac{e(x)}{f_1(x)}}. \tag{3.8}$$

However, the optimal weights (3.8) cannot be directly used in practice, unless we know the true difference between the ICCs. Figure 3.2 depicts selected ICCs displaying DIF caused by different parameters and corresponding optimal weights. Note that weight functions do not necessary need to be non-negative, which is also not assumed by Srihera and Stute (2010). Negative values of the weights allow to detect differences between the ICCs even in case when the curves cross.

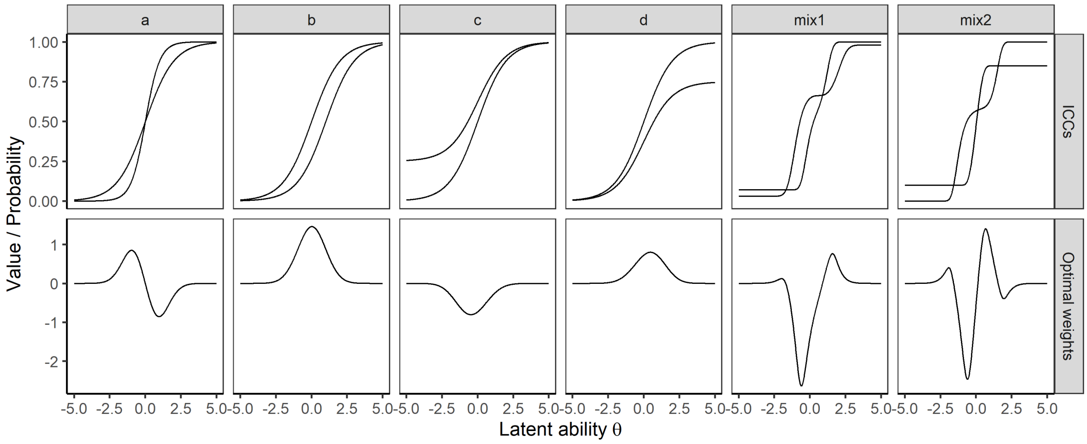


Figure 3.2: Examples of the ICCs and corresponding optimal weight functions (3.8) for DIF caused by various parameters a, b, c, and d in 4PL IRT model, and for logistic curves with several inflection points using normally distributed latent trait for both groups.

Estimates of optimal weights

Third, going beyond work of Srihera and Stute (2010), we propose and study estimates of the optimal weights (3.8). Srihera and Stute (2010) in their work considered only fixed weights (3.6) and optimal weight functions (3.8) for specific examples. However, typically, the true difference function $s_i(x)$ and the true ICCs $m_{i0}(x)$ and $m_{i1}(x)$ or densities $e(x)$, $f_0(x)$, and $f_1(x)$ are not known and need to be estimated.

As a natural estimate of the optimal weights (3.8) we consider $\hat{W}_i(x)$ which incorporates estimates of the densities and estimates of ICCs (3.1) and difference function:

$$\hat{W}_i(x) = \frac{\hat{m}_{i0}(x) - \hat{m}_{i1}(x)}{(1-\hat{\lambda})\hat{\sigma}_{i0}(x)\frac{\hat{e}(x)}{\hat{f}_0(x)} + \hat{\lambda}\hat{\sigma}_{i1}(x)\frac{\hat{e}(x)}{\hat{f}_1(x)}}. \tag{3.9}$$

Contrary to the cases of the fixed weights (3.6) or the optimal weights (3.8), when using the estimated weights $\hat{W}_i(x)$, it can be easily seen that the asymptotic normality is no longer met as the resulting test statistic for the item i has the following form:

$$\hat{T}_i = \frac{1}{n_0 n_1} \sum_{p_0 \in \mathcal{P}_0} \sum_{p_1 \in \mathcal{P}_1} \frac{\left[\hat{m}_{i0}\left(\bar{X}_{p_0 p_1}\right) - \hat{m}_{i1}\left(\bar{X}_{p_0 p_1}\right)\right]^2}{(1-\hat{\lambda})\hat{\sigma}_{i0}(\bar{X}_{p_0 p_1})\frac{\hat{e}(\bar{X}_{p_0 p_1})}{\hat{f}_0(\bar{X}_{p_0 p_1})} + \hat{\lambda}\hat{\sigma}_{i1}(\bar{X}_{p_0 p_1})\frac{\hat{e}(\bar{X}_{p_0 p_1})}{\hat{f}_1(\bar{X}_{p_0 p_1})}}. \tag{3.10}$$

The test statistic (3.10) using the estimate of optimal weights (3.9) results into the form with $(\hat{m}_{i0}(x) - \hat{m}_{i1}(x))^2$ in the numerator. This changes original interpretation of the test statistic, which was the average (weighted) difference of the ICCs, to the average (weighted) square differences of the ICCs. Similarly as for the optimal weights (3.8), this allows to detect differences between the ICCs even in case when the curves cross. If the weight function was assumed to be non-negative, which is common but not assumed in Srihera and Stute (2010), this would not be possible as the differences between the curves may sum up to zero.

Wild bootstrap. To evaluate the properties of the test statistic using the estimate of optimal weights (3.10), the wild bootstrap technique is considered (Wu, 1986; Mammen, 1993). This method is suitable when data exhibits heteroskedasticity (see, for example, Hardle & Mammen, 1993), which is in line with nature of binary responses discussed here. Wild bootstrap can be described as follows (see also Figure 3.3):

(1) First, at the initial step, estimates of the ICCs are calculated using (3.1). Then, the estimate of optimal weights (3.9) is computed and DIF detection procedure is evaluated using the test statistic (3.10) (left part of Figure 3.3).

(2) Further, under the null hypothesis (i.e., no DIF), one ICC for both groups is estimated and corresponding fitted values $\left\{\hat{y}_{pi}\right\}_{p=1}^{n}$ and residuals $\{\hat{e}_{pi}\}_{p=1}^{n}$ are calculated. This is done in two sub-steps.

(2a) For each $b \in \{1, \ldots, B\}$, B being a number of bootstrap samples, a random variable v_{pib} is generated. Here we consider two options: First, we use Mammen's two-point distribution (Mammen, 1993) which is the most common choice:

$$v_{pib} = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{with probability } \frac{\sqrt{5}+1}{2\sqrt{5}}, \\ \frac{\sqrt{5}+1}{2} & \text{with probability } \frac{\sqrt{5}-1}{2\sqrt{5}}. \end{cases}$$

The bootstrap samples y^*_{pib} are then created by combining fitted values and residuals multiplied by the random variables v_{pib}:

$$y^*_{pib} = \hat{y}_{pi} + v_{pib}\hat{e}_{pi}.$$

Second, we generate y^*_{pib} directly from Bernoulli distribution using $\hat{y}_{pi}$, i.e.,

$$y^*_{pib} \sim \text{Bernoulli}(\hat{y}_{pi}),$$

to account for binary nature of the data.

(2b) For each bootstrap sample, the DIF detection procedure is applied as for the original sample at the initial step, resulting in a set of test statistics $\left\{\hat{T}_{ib}\right\}_{b=1}^{B}$.

(3) Finally, in the last step, the set of test statistics $\left\{\hat{T}_{ib}\right\}_{b=1}^{B}$ is compared to the test statistic of original sample and conclusion on DIF is made based on two sided p-value:

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\hat{T}_i < \hat{T}_{ib}},$$

and predefined level of significance (see also the right part of Figure 3.3).

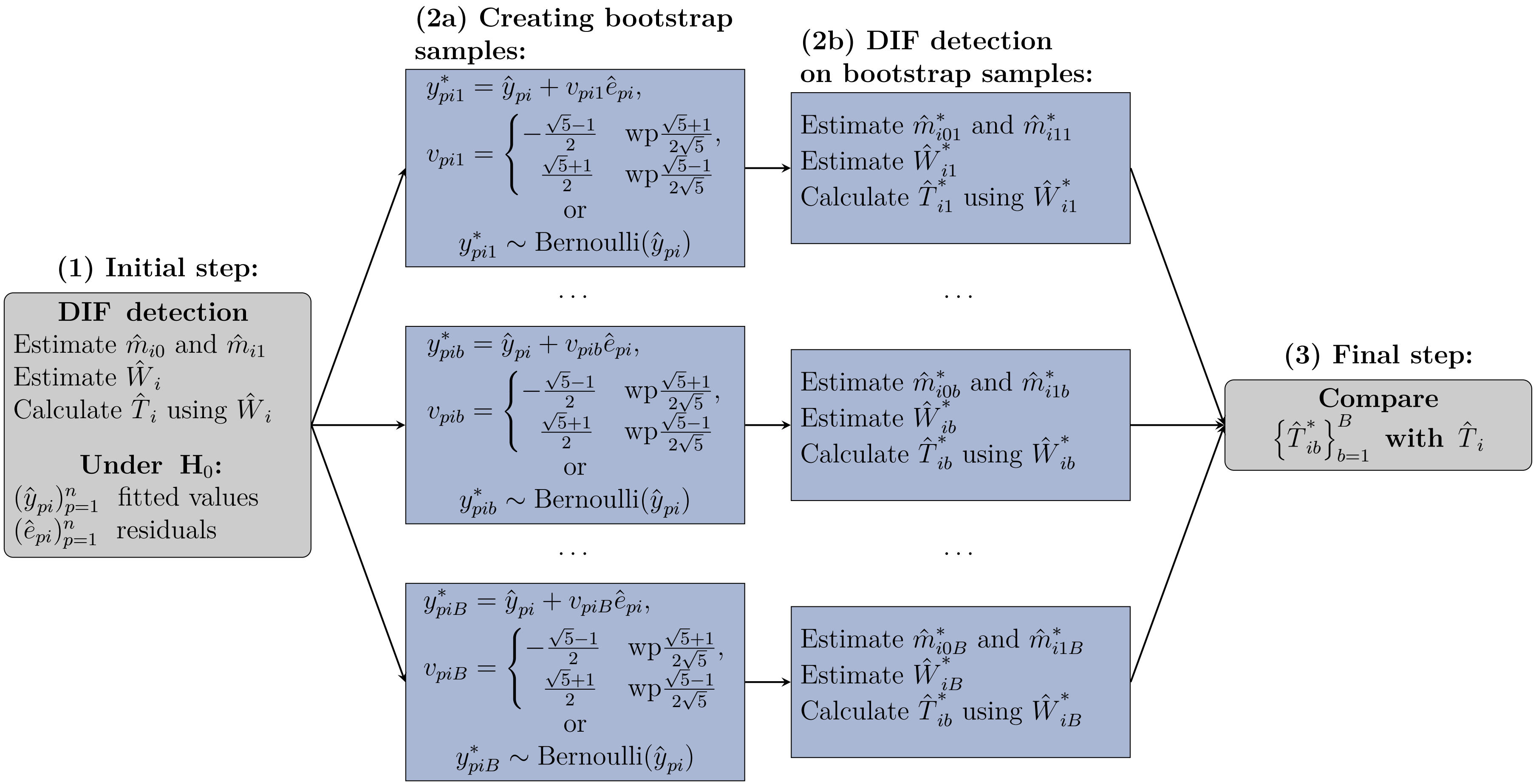

(1) Initial step:
DIF detection
Estimate $\hat{m}_{i0}$ and $\hat{m}_{i1}$
Estimate $\hat{W}_i$
Calculate $\hat{T}_i$ using $\hat{W}_i$
Under $\mathbf{H}_0$:
$(\hat{y}_{pi})_{p=1}^n$ fitted values
$(\hat{e}_{pi})_{p=1}^n$ residuals
(2a) Creating bootstrap samples:
$y^*_{pi1} = \hat{y}_{pi} + v_{pi1}\hat{e}_{pi}$,
$v_{pi1} = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{wp} \frac{\sqrt{5}+1}{2\sqrt{5}}, \\ \frac{\sqrt{5}+1}{2} & \text{wp} \frac{\sqrt{5}-1}{2\sqrt{5}} \end{cases}$
or
$y^*_{pi1} \sim \text{Bernoulli}(\hat{y}_{pi})$
...
$y^*_{pib} = \hat{y}_{pi} + v_{pib}\hat{e}_{pi}$,
$v_{pib} = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{wp} \frac{\sqrt{5}+1}{2\sqrt{5}}, \\ \frac{\sqrt{5}+1}{2} & \text{wp} \frac{\sqrt{5}-1}{2\sqrt{5}} \end{cases}$
or
$y^*_{pib} \sim \text{Bernoulli}(\hat{y}_{pi})$
...
$y^*_{piB} = \hat{y}_{pi} + v_{piB}\hat{e}_{pi}$,
$v_{piB} = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{wp} \frac{\sqrt{5}+1}{2\sqrt{5}}, \\ \frac{\sqrt{5}+1}{2} & \text{wp} \frac{\sqrt{5}-1}{2\sqrt{5}} \end{cases}$
or
$y^*_{piB} \sim \text{Bernoulli}(\hat{y}_{pi})$
(2b) DIF detection on bootstrap samples:
Estimate $\hat{m}^*_{i01}$ and $\hat{m}^*_{i11}$
Estimate $\hat{W}^*_{i1}$
Calculate $\hat{T}^*_{i1}$ using $\hat{W}^*_{i1}$
...
Estimate $\hat{m}^*_{i0b}$ and $\hat{m}^*_{i1b}$
Estimate $\hat{W}^*_{ib}$
Calculate $\hat{T}^*_{ib}$ using $\hat{W}^*_{ib}$
...
Estimate $\hat{m}^*_{i0B}$ and $\hat{m}^*_{i1B}$
Estimate $\hat{W}^*_{iB}$
Calculate $\hat{T}^*_{ib}$ using $\hat{W}^*_{iB}$
(3) Final step:
Compare
$\left\{\hat{T}^*_{ib}\right\}_{b=1}^B$ with $\hat{T}_i$

Figure 3.3: Wild bootstrap scheme.

To compare the two approaches to generate bootstrap samples, i.e., the one based on Mammen's two-point distribution and the one based on Bernoulli distribution, a short simulation study was performed (results not shown). Both approaches yielded similar results in terms of power and rejection rate. As the later approach corresponds to the binary nature of the data, we decided to use it in this work.

3.2 Design of the simulation study

To gain insight into the properties of the nonparametric DIF detection method (3.2) and to compare various options for weight functions, a Monte Carlo simulation study was performed and its design is described in this section.

3.2.1 Data and DIF generation

Dichotomous item responses for the reference and the focal group were generated under a true 4PL IRT model (1) and also by logistic curve with several inflection points:

$$\mathsf{P}(Y_{pi} = 1|\theta_p) = c_i + (d_i - c_i)\frac{e^{-a_i(\theta_p - b_i - e_i\theta_p^2 - f_i\theta_p^3 - g_i\theta_p^5)}}{1 + e^{-a_i(\theta_p - b_i - e_i\theta_p^2 - f_i\theta_p^3 - g_i\theta_p^5)}}, \tag{3.11}$$

where $a_i, b_i, c_i, d_i, e_i, f_i,$ and g_i are item parameters. Ability levels θ_p in both groups were assumed to follow standard normal distribution.

Responses on non-DIF items were generated with the true 4PL IRT model (1) and its parameters were drawn from normal distributions: Discrimination $a_i \sim \mathcal{N}(1.1, 0.3)$, difficulty $b_i \sim \mathcal{N}(0, 1.1)$, guessing $c_i \sim \mathcal{N}(0.2, 0.05)$, and inattention $d_i \sim \mathcal{N}(0.8, 0.05)$, and set to be the same for the reference and focal group.

To generate diferentially functioning items, six different sources of DIF were considered. First, a true 4PL IRT model (1) was used in four scenarios to incorporate DIF caused by difference in particular parameter (i.e., either a_i, b_i, c_i, or d_i). Second, two different settings were considered for generating DIF caused by logistic curve with several inflection points (3.11). In the first setting, parameters were selected so that the ICCs intersected exactly once, while in the second setting they intersected twice. The difference in parameters, either when using model (1) or (3.11), was chosen to approximately match the WAM between the two ICCs (Siebert, 2013, see also Section 1.5) of value 0.196 to obtain DIF of large magnitude. The ICCs and corresponding optimal weight functions of DIF items are illustrated in Figure 3.2 and their parameters can be found in Table 3.1.

Standardized total test score was used as the matching criterion, which is not a continuous random variable as was assumed by Srihera and Stute (2010). We weakened this assumption in the simulation study as the standardized total score is the most common and the simplest estimate of the underlying ability.

Besides the source of DIF, sample sizes were manipulated to generate data. The total sample sizes of 50, 100, 200, 300, and 400 were selected, while groups were equally sized. The test length of 20 items was considered of which only one item functioned differently, i.e., 5% of the total number of items.

Table 3.1: Item parameters used to generate DIF items with models (1) and (3.11).

DIF source	Reference group							Focal group						
	a	b	c	d	e	f	g	a	b	c	d	e	f	g
a	0.42	0.00	0.00	1.00	0.00	0.00	0.00	2.00	0.00	0.00	1.00	0.00	0.00	0.00
b	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00
c	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.39	1.00	0.00	0.00	0.00
d	1.00	0.00	0.00	0.61	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00
mix1	1.90	0.28	0.07	1.00	1.00	−0.70	0.00	0.35	−1.75	0.03	0.98	1.60	−0.90	0.00
mix2	4.20	0.00	0.10	1.00	0.85	0.00	−0.50	0.18	−1.50	0.00	1.00	1.00	−0.90	−0.50

3.2.2 DIF detection

Five methods were used for DIF detection: Four variations of the nonparametric method and logistic regression for DIF detection with the likelihood ratio test (7) (Swaminathan & Rogers, 1990). The following four different weight functions were used in the nonparametric method: Fixed weight function (3.6), optimal weight function (3.8), estimate of optimal weight function (3.9) without bootstrap (i.e., assuming asymptotically normal distribution of the test statistic (3.2)), and estimate of optimal weight function using bootstrap (Figure 3.3). Optimal weight function was applied only for DIF items, while its value was set to zero for non-DIF items, hence no non-DIF items could have been detected as functioning differently for the two groups. Thus, reported rejection rates are zeros in this case. Estimate of optimal weight function using bootstrap was applied with a number of bootstrap samples of $B = 500$.

To estimate the ICCs for nonparametric method, we used Epanechnikov kernel with three different bandwidth parameters $h = n_0^{-\zeta}$, where $\zeta \in \left\{0.26, \frac{7}{24}, 0.32\right\}$, to cover interval of its possible values meeting the assumptions by Srihera and Stute (2010), see also page 90. All tests were performed at 0.05 significance level.

3.2.3 Evaluation of the results

The five different approaches (4 different settings for nonparametric approach and the logistic regression method) were compared in terms of power (i.e., the proportion of true positives) and rejection rate (i.e., the proportion of false positives). Further, we checked accuracy of the estimate of the optimal weights by computing the MSE, i.e., mean square difference between the optimal weights and their estimates. Finally, we also evaluated performance of DIF detection methods in terms of execution time. Each condition was replicated 1,000 times.

3.2.4 Implementation

Estimation of the ICCs, calculation of the test statistics, and the whole simulation study was performed in the statistical software **R** (R Core Team, 2020) and its packages. Empirical density functions were calculated using the `ecdf()` function from the **`stats`** package (R Core Team, 2020). Weights of kernel functions and kernel estimates were calculated by the `locCteWeightsC()` and `locWeightsEval()` functions from the **`locpol`** package (Cabrera, 2018). Estimates of densities of standardized total scores and common support of the test statistic were evaluated by the `bkde()` function from the **`KernSmooth`** package (Wand, 2019). The logistic regression method with the likelihood ratio test was performed using the `difLogistic()` function from the **`difR`** (Magis et al., 2010). Finally, graphical representation of the results was made using the **`ggplot2`** package (Wickham, 2016).

3.3 Results

3.3.1 Rejection rates and power

Estimate of the optimal weights (3.1) without the wild bootstrap was the most powerful approach in all scenarios (mean power rate 0.724), however, this was accompanied by rejection rates considerably exceeding the significance level of 0.05 (mean rejection rate 0.272). Therefore, results of this approach were removed from further analysis and are not shown.

All other approaches were able to control for type I error for all parameters ζ and for all sources of DIF (Figure 3.4). Slight excesses of the significance level were present for the nonparametric approach with fixed weights and the logistic regression method, especially when total sample sizes of 50 a 100 were considered.

All methods gained slightly lower power for small sample sizes, however, with the increasing sample size they were able to yield sufficient power and the differences between the approaches were diminishing.

The nonparametric approach with fixed weights (3.6) gained power rates close to those obtained by the optimal weight function (3.8) for almost all scenarios, however, the main drawback of this approach was the inability to detect DIF caused by difference in parameter a (the first row of Figure 3.5).

Power rates of the approach using the wild bootstrap technique were slightly lower than those obtained by the fixed weights in most scenarios, however, this approach had much greater power in case of the crossing DIF caused by parameter a (the first row of Figure 3.5) and somewhat also in the second case of the ICCs with several inflection points (row mix2 of Figure 3.5).

As expected, the greatest potential of the nonparametric approaches was visible in the second case of the ICCs with several inflection points (row mix2 of Figure 3.5) where the nonparametric approach using optimal weights was superior for all sample sizes. However, the nonparametric methods slightly outperformed the logistic regression also when DIF was caused by parameters b or c and in the first case of the ICCs with several inflection points (rows c, d, and mix1 of Figure 3.5) for optimal and fixed weights especially for smaller sample sizes.

Difference between the nonparametric methods using different bandwidth parameters h were small. With the lower value of ζ, i.e., larger bandwidth parameter $h = n_0^{-\zeta}$, optimal weight functions and their estimates using the wild bootstrap yielded slightly larger mean power while for the fixed weights the largest mean power was gained with $\zeta = \frac{7}{24}$. Differences in mean rejection rates were negligible.

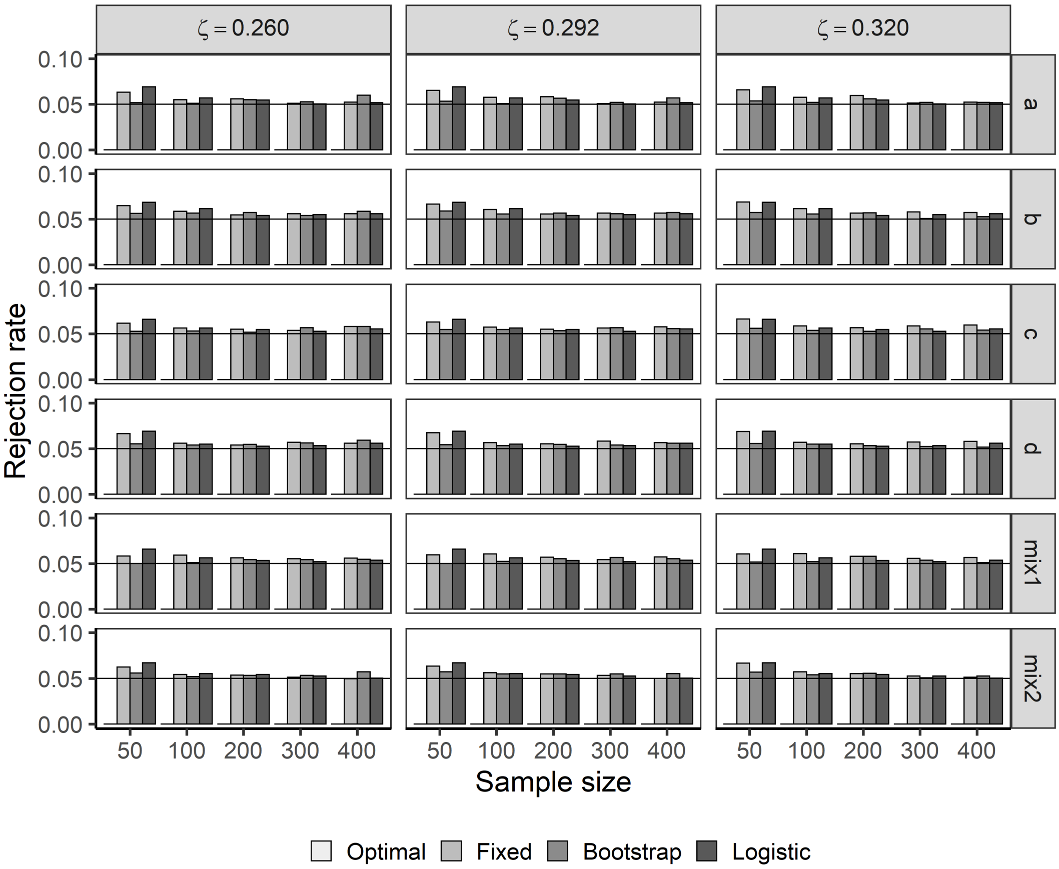


Figure 3.4: Rejection rates by nonparametric approach with the various weight functions and by the logistic regression method with respect to the sample size and the ζ parameter of the bandwidth for different sources of DIF. The horizontal line shows significance level of 0.05.

3.3.2 Estimates of optimal weights

With $\zeta = 0.292$ and $\zeta = 0.320$, MSE of the optimal weights when parameter c was causing DIF had value larger than 10^6. These scenarios were removed from further analysis and are not shown here.

Estimation of the optimal weight was the most precise when parameters b, c, and d were sources of DIF, while it was the least precise when mixture of parameters was causing the DIF (Figure 3.6), especially for the second case of the ICCs with several inflection points (row mix2 of Figure 3.6). The smallest overall MSE of 0.242 was gained for $\zeta = 0.320$ (the smallest bandwidth parameter h) and the largest overall MSE of 0.270 for $\zeta = 0.260$ (the largest h).

All three choices of ζ parameter overall gained more accurate estimates for the large sample sizes than for the smaller ones. While there were no large differences between bandwidth parameters when parameters b, c, or d were sources of DIF, this was not the case when the ICCs crossed. In such a case, $\zeta = 0.320$ (the smallest bandwidth parameters) gained the greatest accuracy while for $\zeta = 0.260$ the largest values of the MSE were observed for all levels of sample size. Moreover, the precision of the estimates was not strictly decreasing with the increasing sample size (rows a and mix1 of Figure 3.6).

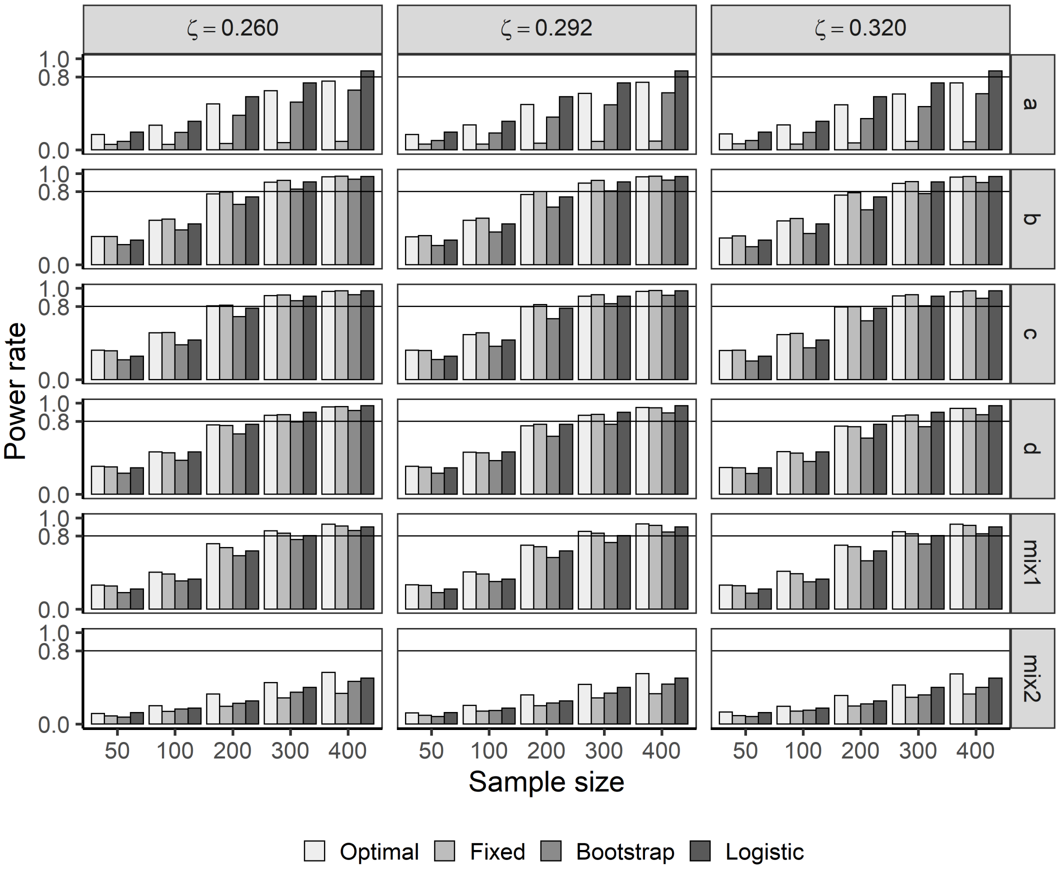


Figure 3.5: Power rates by nonparametric approach with the various weight functions and by the logistic regression method with respect to the sample size and the ζ parameter of the bandwidth for different sources of DIF. The horizontal line shows sufficient power of 0.80.

3.3.3 Execution time

Mean time to perform the DIF detection remained on a user-friendly rate for all the methods except for the estimate of the optimal weights with the wild bootstrap. Execution time of this method seemed to be exponential with the increasing sample size (Table 3.2).

3.4 Summary

In this chapter we dealt with the nonparametric comparison of regression curves for DIF detection among binary items. We adapted general approach for testing differences between the regression curves proposed by Srihera and Stute (2010) to test differences between the ICCs for binary data and to test for DIF. Specifically, we proposed an alternative estimate of the asymptotic variance of the test statistic to account for binary nature of data. Further, we focused on topic of weight functions which may have great impact on power of the test. We derived form of the optimal weights for binary data in sense of maximizing local power of the test and we newly proposed their estimates, considering the fact that the optimal weights are not available in real situations such as in case of DIF detection.

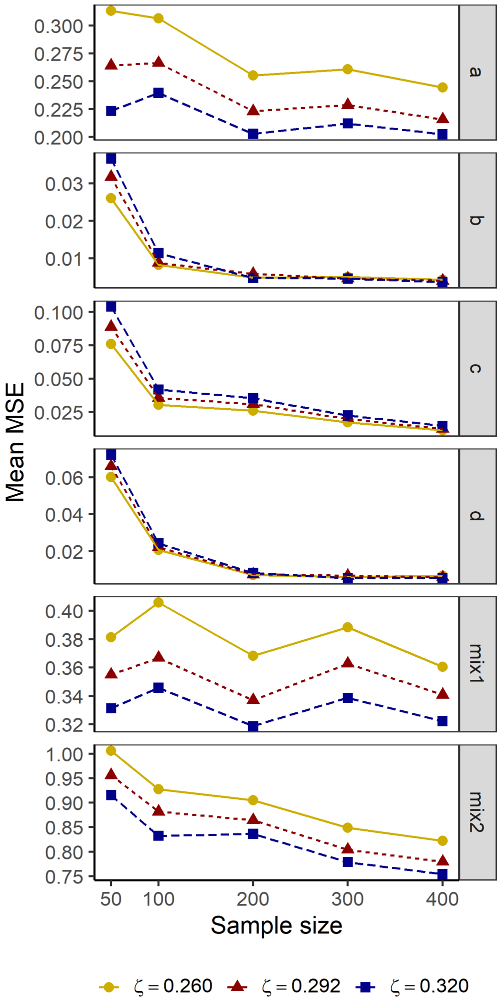


Figure 3.6: MSE of the estimates of optimal weights with respect to the parameter ζ, source of DIF, and sample size.

Unlike in the case of the optimal weights assumed by Srihera and Stute (2010), the underlying test statistic does not have asymptotically normal distribution any longer and, moreover, asymptotic distribution is not known. Therefore, we proposed using the wild bootstrap to evaluate properties of the asymptotic distribution of the test statistic and to test for DIF.

We further performed simulation study to assess properties of the nonparametric approach in terms of power and rejection rates using different weight functions in comparison to the logistic regression method. All methods performed good control of type I error. The nonparametric approach using the optimal weights gained power rates close to those by the logistic regression method and it outperformed it in several scenarios, especially in scenario with the multiple crossings of the ICCs. Comparing different weight functions in the nonparametric approach, the fixed weight function performed similarly to method using the optimal weights in case that the ICCs did not cross and may be recommended when it can be assumed that one group is advantaged over the other group for all levels of the matching criterion. The newly proposed estimate of the optimal weights using the wild bootstrap outperformed fixed weights in case that underlying ICCs

Table 3.2: Mean time to perform DIF detection methods in seconds with respect to the sample size.

Method	Sample size				
	50	100	200	300	400
Optimal weights	0.025	0.042	0.098	0.189	0.316
Fixed weights	0.024	0.040	0.098	0.190	0.320
Est. weights w/o bootstrap	0.074	0.091	0.149	0.243	0.370
Est. weights with bootstrap	32.867	41.473	70.648	116.309	180.416
Logistic regression	0.284	0.311	0.333	0.356	0.378

crossed. With the increasing sample size, powers of all methods were increasing and differences between them diminished.

The power of the nonparametric method with the newly proposed estimate of optimal weights using the wild bootstrap is closely connected with the precision of the weights estimate. This is also related to the choice of the bandwidth parameter h. In case that the ICCs did not cross, all bandwidth choices gained similar precision which was increasing with the increasing sample size. In case that the ICCs crossed, the smallest bandwidth parameter gained the greatest accuracy. However, the precision was not strictly increasing with an increasing sample size, especially when DIF was caused by parameter a or in the first mixed scenario.

The simulation study offered here was limited in terms of number of items and proportion of DIF items. It should be noted that only small or moderate sample sizes were included to keep simulations on computationally-friendly level, which excludes the nonparametric model (1.1) as it requires sufficiently large sample size for both groups. Further, we considered only three levels of bandwidth parameter. The choice of bandwidth parameter h is connected to precision of the estimation of optimal weight functions: If the h is too small, we may obtain an under-smoothed estimate. On the other hand, in the case that the value of h is large, the estimate may be over-smoothing. However, in our simulation study, no large differences were observed in terms of power or rejection rates with respect to the choice of the bandwidth parameter.

In summary, the nonparametric approaches, including the newly proposed estimate of the optimal weights with the wild bootstrap, were able to control significance level and, in most cases, dealt with DIF detection as effectively as the logistic regression method. Moreover, the nonparametric method seems to have the potential to outperform the logistic method for scenarios with several inflection points.

4. Further issues in DIF detection

This chapter deals with some further issues in DIF detection among dichotomously scored items including *item purification* and *multiple comparisons corrections* as presented in Hladká, Martinková, and Magis (2021). It also comprises topic of so called *DIF effect sizes* – measures which classify magnitude and importance of DIF being detected.

4.1 Introduction

Most of the DIF detection methods rely on the basic principle of testing for DIF one item after another, the remaining items being considered as anchor (DIF-free) items. This process is known to have at least two drawbacks. First, when DIF items are truly present in the data, gradual DIF testing implies that DIF items are included in the matching variable (for instance the test score), which is known to be a source of a potentially serious bias and misidentification of DIF and non-DIF items (Jodoin & Gierl, 2001; Kopf, Zeileis, & Strobl, 2013, 2015; Woods, 2009). Second, testing each item after another usually yields inflated type I error rates (i.e., proportion of falsely detected items) because traditional methods do not adjust for multiple comparisons, which is actually what happens with this repeated, item-by-item process.

Each issue was to some extent addressed in the DIF literature in different ways. To reduce the impact of DIF items on the matching variable, the *item purification* process was proposed (Lord, 1980; first suggested by Marco, 1977 and extended and improved by many authors including Candell & Drasgow, 1988, Clauser, Mazor, & Hambleton, 1993, and French & Maller, 2007). The item purification consists of an iterative removal of items flagged as DIF from the set of anchor items (Candell & Drasgow, 1988). Item purification was shown to improve the results of most DIF detection methods (Clauser et al., 1993; French & Maller, 2007; Navas-Ara & Gómez-Benito, 2002; Wang & Su, 2004), with the notable exception of Angoff's delta plot method (Magis & Facon, 2013). The other issue, inflated type I error rates due to multiple comparisons, on the other hand, can be accurately controlled with adequate *multiple comparison adjustment* procedures. Corrections for multiple comparisons are easy to implement, non-iterative, and were also shown to improve the accuracy of DIF identification (i.e., non-inflated type I errors and larger power; see Kim & Oshima, 2013).

4.1.1 Item purification

DIF analysis is based on the principle of comparing item performance of the test takers being matched by the ability. Thus, defining an appropriate matching criterion is mandatory. For non-IRT DIF detection methods such as the Mantel-Haenszel test (4) (P. W. Holland & Thayer, 1988; Mantel & Haenszel, 1959) or logistic regression procedure (7) (Swaminathan & Rogers, 1990), the total test score, i.e., the number of correct responses, is usually used as the matching criterion. For IRT-based techniques such as Lord's test (2) (Lord, 1980), the estimate of latent ability level is used instead.

The danger of computing such matching criterion for the set of administered items is that the inclusion of DIF items could seriously impact the results of the identification process. It is then of primary importance to ensure that anchor (i.e., DIF-free) items are available for proper computation of this matching variable. For non-IRT methods, the matching criterion (observed ability) should be computed by only using anchor items. For IRT-based methods, linking the two scales (one for the reference group and one for the focal group) should be based only on these anchor items.

Because it is often impossible to predict which items will function differently, Candell and Drasgow (1988) proposed an iterative process that is currently referred to as *item purification*. In test-score-based DIF detection methods, it starts with one run of the DIF detection method per item, all other items being considered as anchor items. All items flagged as DIF are then removed from the set of anchor items, and the method is re-run using this reduced anchor set. These two steps (running DIF analysis and removing flagged items from the anchor set) are repeated until two successive runs yield the same set of items identified as DIF (see Figure 4.1).

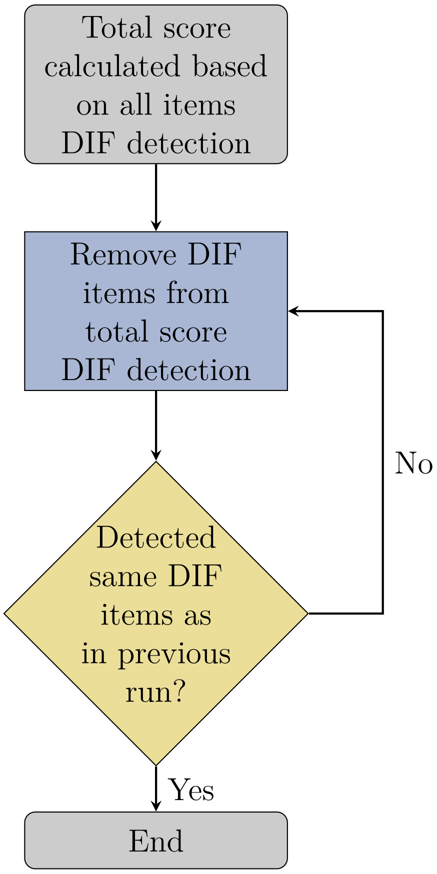

Total score calculated based on all items DIF detection
Remove DIF items from total score DIF detection
No
Detected same DIF items as in previous run?
Yes
End

Figure 4.1: Item purification scheme.

To illustrate item purification algorithm, let's assume an artificial test consisting of 10 items and an arbitrary non-IRT DIF detection method. At the initial step, total test score was calculated based on all 10 items. Using DIF detection method and total test score, items 1, 7, and 8 were detected as DIF items.

In the first step of the item purification, such items were removed from calculation of the total score and DIF detection procedure was then applied using this new matching criterion. In the second step, only items 1 and 8 were detected as functioning differently. The set of DIF items did not respond to the set of the previous iteration and thus the matching criterion was recalculated and DIF detection procedure was run again. In the third step, items 1, 2, and 8 were detected as DIF. Again, current and previous sets of DIF items were not the same and the matching criterion needed to be calculated once more. Finally, in the fourth step, items 1, 2, and 8 were detected as in previous iteration and the algorithm stopped (Table 4.1). Illustration of practical implementation in `R` within the nonlinear model (1.1) and **difNLR** package can be found in Section 1.4 on page 42.

Table 4.1: Illustration of item purification.

Total score	$\sum_{i=1}^{10} Y_i$	$\sum_{i\neq\{1,7,8\}} Y_i$	$\sum_{i\neq\{1,8\}} Y_i$	$\sum_{i\neq\{1,2,8\}} Y_i$
Item	Step 1	Step 2	Step 3	Step 4
1	DIF	DIF	DIF	DIF
2	NON-DIF	NON-DIF	DIF	DIF
3	NON-DIF	NON-DIF	NON-DIF	NON-DIF
4	NON-DIF	NON-DIF	NON-DIF	NON-DIF
5	NON-DIF	NON-DIF	NON-DIF	NON-DIF
6	NON-DIF	NON-DIF	NON-DIF	NON-DIF
7	DIF	NON-DIF	NON-DIF	NON-DIF
8	DIF	DIF	DIF	DIF
9	NON-DIF	NON-DIF	NON-DIF	NON-DIF
10	NON-DIF	NON-DIF	NON-DIF	NON-DIF

Item purification is an approach which is intuitively appealing and simple to implement. Though item purification can be done efficiently in most cases, it can sometimes become time consuming (especially for IRT-based methods), and there is no guarantee that the iterative process will converge (for example, see troubleshooting in Section 1.4).

4.1.2 Multiple comparison corrections

Conceptually different drawback often present in DIF detection is that each item is being tested individually, while all other items are considered free of DIF. This implies that multiple comparisons among all test items arise, which is without adjustment of the significance level known to lead to inflated type I error rates. In the DIF framework, Kim and Oshima (2013) proposed adjusting the results of the item-by-item investigation using methods to control for multiple comparisons. Two such adjustment procedures were shown to be superior in the DIF context: Holm's procedure (Holm, 1979) and Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995). These methods can be schematically described as follows: First, for each tested item i (say from 1 to I), let p_i be the corresponding p-value of the DIF detection method (obtained when all other items are set as anchor items), and let $p_{(1)}, \ldots, p_{(I)}$ be the I values sorted in increasing order. Then, for a given global significance level α, the index k is defined as

(1) the minimal index that satisfies $p_{(k)} > \frac{\alpha}{I+1-k}$ for Holm's procedure,

(2) the maximal index that satisfies $p_{(k)} \leq \frac{k}{I}\,\alpha$ for BH procedure.

Eventually, items with corresponding ordered p-values $p_{(1)}$ to $p_{(k-1)}$ (for Holm's procedure) or to $p_{(k)}$ (for BH procedure) are flagged as DIF, while the remaining items are flagged as non-DIF.

These methods are illustrated using an artificial example of ten items (Table 4.2). Holm's and BH boundaries were calculated by formulas in (1) and (2), and then compared with ordered p-values. With Holm's procedure, index k was equal to three; thus only the first two listed items (i.e., items 5 and 10) were eventually flagged as DIF. This is an important reduction compared to the original classification (without Holm's correction) that led to flagging seven out of the ten items as DIF. With BH procedure, k index equalled to five; therefore the first five items (according to their classification in increased order of p-values) were flagged as DIF, compared to the seven items when no adjustment was considered.

Table 4.2: Impact of Holm's and Benjamini-Hochberg corrections for multiple comparisons on DIF detection.

Item	Order	p-value	Decision	Holm's boundary	Holm's decision	BH boundary	BH decision
5	1	0.0014	DIF	0.0050	DIF	0.0050	DIF
10	2	0.0039	DIF	0.0056	DIF	0.0100	DIF
9	3	0.0111	DIF	0.0062	NON-DIF	0.0150	DIF
8	4	0.0182	DIF	0.0071	NON-DIF	0.0200	DIF
3	5	0.0209	DIF	0.0083	NON-DIF	0.0250	DIF
6	6	0.0306	DIF	0.0100	NON-DIF	0.0300	NON-DIF
2	7	0.0388	DIF	0.0125	NON-DIF	0.0350	NON-DIF
4	8	0.2430	NON-DIF	0.0167	NON-DIF	0.0400	NON-DIF
7	9	0.3623	NON-DIF	0.0250	NON-DIF	0.0450	NON-DIF
1	10	0.7826	NON-DIF	0.0500	NON-DIF	0.0500	NON-DIF

This example highlights how correction methods for multiple comparison have a straightforward impact on detection of DIF items. Holm's procedure (Holm, 1979) is an improvement of Bonferroni's procedure that is more powerful (B. S. Holland & Copenhaver, 1988). It is intended to control family-wise error, that is, the probability of making one or more type I errors. BH procedure controls a false discovery rate, that is, the expected proportion of type I errors (Benjamini & Hochberg, 1995). Procedures to control false discovery rate have greater power at the cost of increased type I error rates (Shaffer, 1995). Kim and Oshima (2013) discuss these approaches in detail in the DIF context. Again, illustration of practical implementation in **R** within the nonlinear model (1.1) and the **difNLR** package can be found in Section 1.4 on page 43. Syntax when using the **difR** package, which offers several DIF detection methods among binary data such as the Mantel-Haenszel test (4) or the logistic regression method (7), is analogous.

4.1.3 Aims of the study

Though conceptually different and with different purposes, both approaches, item purification and corrections for multiple comparisons, share the same objective, that is, the improvement of the classification of items into DIF and non-DIF groups. While both methods are still being studied intensively (Chen & Hwu, 2018; Fikis & Oshima, 2017; Khalid & Glas, 2014; Kim & Oshima, 2013), surprisingly, and to our best knowledge, performance of these approaches has not yet been jointly evaluated in a comprehensive study. Moreover, combinations of these techniques, to our best knowledge, have not yet been explored. This represents a potential gap in the DIF literature, as both approaches were shown to improve DIF detection to a certain extent.

In this work we introduce two different settings for applying both methods, item purification and multiple comparison corrections, together. First, we consider *simple* combination of both approaches. Simple combination contains full item purification process being followed by multiple comparison adjustment of the final purification results. Second, we propose another mixture of the approaches, here referenced as *combined*. Combined approach performs item purification followed by correction for multiple comparison in the each iteration of item purification.

The aim of this work is to perform a simulation which would allow us to improve the knowledge about these techniques and for practical purposes (i.e., in order to formulate tractable recommendations for better DIF practices in real data analyses). In a complex DIF simulation study, we aim to evaluate properties of both approaches (purification and adjustments) and their combinations under selected DIF detection methods and under various scenarios for three selected DIF detection methods: The Mantel-Haenszel test (4) (Mantel & Haenszel, 1959), the logistic regression method (7) (Swaminathan & Rogers, 1990), and the SIBTEST (6) (Shealy & Stout, 1993),

4.2 Methods

This part comprises design of the simulation study to assess effects of correction methods on DIF detection procedures, the summary statistics considered for output analysis, and practical implementation details.

4.2.1 Data and DIF generation

Six design factors were manipulated to generate the data: (a) sample size, (b) test length, (c) amount of DIF items, (d) type of DIF, (e) size of DIF effect, and (f) distribution of ability for focal group. The total sample sizes 250 (125 per group), 500 (250 per group), 1,000 (500 per groups), and 2,000 (1,000 per group) were selected, while test lengths of 20, 40, and 80 items were considered. Four different proportions of DIF items (0%, 5%, 15%, and 30%) were considered. Parameters of DIF items were chosen to incorporate both types of DIF (uniform and non-uniform) in two different sizes of DIF effect (0.4 and 0.8) quantified by the AM (Raju, 1988, see also Section 1.5). DIF effect sizes correspond to small and large

DIF magnitudes and were selected following Swaminathan and Rogers (1990) and Narayanan and Swaminathan (1996).

The item responses were generated under a true 3PL IRT model (1.30). In all scenarios, the ability of a reference group was drawn from the standard normal distribution. For a focal group we considered three options for ability levels. First, ability levels were the same as for the reference group – drawn from a standard normal distribution. Second, ability levels for the focal group were drawn from a normal distribution with a mean equal to 1 but with the same standard deviation as for the reference group. Third, moreover standard deviation for normal distribution for focal group was manipulated and set to 1.5.

Parameters of non-DIF items were selected from problem solving of GMAT (Kingston et al., 1985, see Table at p. 47 for all 80 non-DIF items) to reflect realistic values. When tests of 20 or 40 items were considered, only the set of parameters of the first 20 or 40 items were used.

The DIF item parameters creation was inspired by Narayanan and Swaminathan (1996). The c-parameter was fixed at a value of 0.2 for all DIF items. The choice of discrimination and difficulty parameter values depends on the type of DIF effect generated. For a uniform DIF, discrimination parameter a was fixed for both groups, and difficulty parameter b varied to gain the desired DIF effect size (either small 0.4, or large 0.8). 12 uniform DIF items were simulated with a varying values of b parameter – low ($b = -1$ or $b = -0.5$ for reference group, $b = 0$ for focal group) and high ($b = 0$ fo reference group, $b = 0.5$ or $b = 1$ for focal group); and varying values of a parameter – low ($a = 0.5$), medium ($a = 1$), and high ($a = 1.5$). Table 4.3 summarizes all these options and highlights which non-DIF item(s) were replaced by those parameter values.

For a non-uniform DIF items, difficulty parameter b was fixed for both groups, and discrimination parameter a varied to gain desired DIF effect size. 12 non-uniform DIF items were simulated with varying values of common b parameter – low ($b = -1$), medium ($b = 0$), and high ($b = 1$) and varying values of a parameter – low ($a = 0.43$ or $a = 0.50$ for reference group, $a = 0.72$ or $a = 0.91$ for focal group) and high ($a = 0.56$ or $a = 0.90$ for reference group and $a = 1.79$ or $a = 2.01$ for focal group). These combinations are also listed in Table 4.3.

This simulation design yields 36 settings in absence of DIF (four sample sizes, three test lengths, and three ability distributions) and 324 settings in presence of DIF (in addition, three proportions of DIF, two DIF sizes, and two types of DIF effect), thus 468 design settings in total. For each such setting, 1,000 data sets were generated. Note that given the fact that some DIF detection methods may yield convergence issues, no results are then obtained for items that failed to converge and thus no conclusion about DIF detection could be drawn. To overcome this problem, simulation runs with convergence issues were excluded, and simulations were re-run until 1,000 replications without convergence failures were obtained.

4.2.2 DIF identification

Three methods to detect DIF were selected: the Mantel-Haenszel chi-squared statistic (4) (P. W. Holland & Thayer, 1988; Mantel & Haenszel, 1959), the logistic regression procedure (7) (Swaminathan & Rogers, 1990) with the likelihood ratio test accounting for both types of DIF (i.e., uniform and non-uniform), and the SIBTEST (5) (Shealy & Stout, 1993).

Table 4.3: Parameters of DIF items for the reference and the focal group.

	Item number				Item parameters			
DIF proportion	5%		15%		Reference		Focal	
Test length	20	40	20	40	*a*	*b*	*a*	*b*
Uniform DIF size = 0.4								
	20	39	18	35	1.00	0.00	1.00	0.50
		40	19	36	1.00	−0.50	1.00	0.00
			20	37	0.50	0.00	0.50	0.50
				38	0.50	−0.50	0.50	0.00
				39	1.50	0.00	1.50	0.50
				40	1.50	−0.50	1.50	0.00
Uniform DIF size = 0.8								
	20	39	18	35	1.00	0.00	1.00	1.00
		40	19	36	1.00	−1.00	1.00	0.00
			20	37	0.50	0.00	0.50	1.00
				38	0.50	−1.00	0.50	0.00
				39	1.50	0.00	1.50	1.00
				40	1.50	−1.00	1.50	0.00
Non-uniform DIF size = 0.4								
	20	39	18	35	0.90	0.00	2.01	0.00
		40	19	36	0.50	0.00	0.72	0.00
			20	37	0.90	−1.00	2.01	−1.00
				38	0.50	−1.00	0.72	−1.00
				39	0.90	1.00	2.01	1.00
				40	0.50	1.00	0.72	1.00
Non-uniform DIF size = 0.8								
	20	39	18	35	0.56	0.00	1.79	0.00
		40	19	36	0.43	0.00	0.91	0.00
			20	37	0.56	−1.00	1.79	−1.00
				38	0.43	−1.00	0.91	−1.00
				39	0.56	1.00	1.79	1.00
				40	0.43	1.00	0.91	1.00

Note. Item number = number of item to be replaced in table of non-DIF items (see Kingston et al., 1985, p. 47). Parameter c is 0.2 for all items.

All three DIF detection methods were employed for each generated data set, together with eight possible procedures to control type I error: (a) item purification, (b) Holm's adjustment method, (c) BH method, (d) simple combination of purification with Holm's method, (e) simple combination of purification with BH method, (f) item purification followed by Holm's method after each iteration, (g) item purification followed by BH method after each iteration, and (h) no correction procedure (for bench-marking purposes). Thus, altogether 24 combinations of DIF detection method and type I error controlling procedures were applied to each data set. In case of using item purification, either alone or in combination with adjustment method, maximal number of iterations was set to 50. The significance value was set to 5%.

DIF effect size. In DIF analysis, a question of interest is often not only whether items significantly function differently but also whether detected DIF is of practical significance (Suh, 2016). Thus, besides testing statistically for a presence of items which function differently, DIF analysis is often accompanied by calculation of so called *DIF effect sizes* and their classification (see, e.g., Jodoin & Gierl, 2001; Potenza & Dorans, 1995). Classification of DIF effect sizes helps to assess practical importance and interpretation of DIF being detected, as for example power and often also type I error are increasing with increasing sample size (see, e.g. Swaminathan & Rogers, 1990), while no practical importance is present. DIF effect size is typically classified into three categories: A – negligible, B – moderate, and C – large.

Thus, besides testing which items are significantly functioning differently, DIF effect sizes were calculated to evaluate the magnitude of DIF. For the Mantel-Haenszel test, we used classification based on log-transformation of the common odds ratio $\Delta_{\mathrm{MH}i} = \log(\alpha_{\mathrm{MH}i})$ as proposed by P. W. Holland and Thayer (1985). For the logistic regression method, we considered classification based on Nagalkerke's R^2 (Nagelkerke, 1991) with bounds proposed by Jodoin and Gierl (2001). The effect size is classified based on difference ΔR_i^2 between R^2 coefficients of the two nested models. Finally, for the SIBTEST, we applied classification based on value of $\hat{\omega}_i$ (Roussos & Stout, 1996). Bounds for negligible (A), moderate (B), and large (C) DIF effect sizes are summarized in Table 4.4.

Table 4.4: Bounds for DIF effect size classification.

DIF method	Mantel-Haenszel test	Logistic regression method	SIBTEST
Measure	$\lvert\Delta_{\mathrm{MH}i}\rvert$	ΔR_i^2	$\lvert\hat{\omega}_i\rvert$
A – negligible	0.000	0.000	0.000
B – moderate	1.000	0.035	0.059
C – large	1.500	0.070	0.088

4.2.3 Summary statistics and simulation evaluation

Three summary statistics (type I error rate, rejection rate, and power rate) together with DIF effect sizes were computed across the 1,000 generated data sets per study design, and separately for each of the 24 combinations of DIF detection method and controlling procedures. Type I error was estimated as the proportion of falsely detected items when none of the items were considered as DIF. Rejection rate was calculated as the proportion of falsely detected items among all non-DIF items (in the cases when DIF items were present in the simulation scenario). Finally, power rate was calculated as the proportion of correctly detected DIF items among all truly DIF items. DIF effect sizes were computed for truly identified DIF items as well as those which were falsely detected.

The results were interpreted with respect to the following research questions:

1. Are the DIF detection methods (the Mantel-Haenszel test, the logistic regression method, or the SIBTEST) able to control for type I error (i.e., type I error and rejection rates close to the 5% significant level) with sufficient power (i.e, over 80%) even without any controlling procedure?

2. How do the studied controlling procedures (item purification, Holm's adjustment, BH adjustment) and their combinations (simple combination of purification and Holm's adjustment, simple combination of purification and BH adjustment, item purification followed by Holm's method after each iteration, item purification followed by BH method after each iteration) compare in different scenarios in terms of power?

3. Which design factors have significant impact on type I error rate, rejection rates, and power rates?

The first question was investigated by many authors (see, e.g., van de Water, 2014). In the context of this simulation study, the answer to the first question will help to set the bench-marking values to which other methods will be compared.

To get initial idea, summarizing figures with observed type I error, rejection rates, and power rates were produced. For simplicity, presented values were averaged by scenarios with the same level of a given factor. Type I error and rejection rates were considered as suitable if they were close to the 5% significance level. Power rates was considered as satisfactory if it achieved a value of at least 80%.

To test for significance of the differences between controlling procedures and other study factors, beta regression models for type I error, rejection rates, and power rates were fitted with logit link. All possible double interactions between factors were included into models. Note that since the beta regression model cannot handle extreme values (i.e., 0 or 1), such type I error rates, rejection rates, and power rates values were replaced by values 10^{-6} higher or lower. To simplify the interpretation of the results with respect to the sample size, 250 was subtracted from the sample size variable and then it was divided by 100. Interpretation of the parameter effects in the beta regression model is the same as in logistic regression (e.g., Agresti, 2010). It should be noted that interpretation of the results was made primarily with focus on controlling procedures and their possible interactions with other factors (see research questions above). Since, we were not interested in differences between DIF detection methods (the Mantel-Haenszel test, the logistic regression method, and the SIBTEST) in this study, three separate models were fitted, one for each method.

Besides three summary statistics and their analysis using beta regression model, we also evaluated DIF effect size measures for truly identified DIF items as well as those which were false positives (non-DIF items which were falsely detected as functioning differently). Proportions of negligible, moderate, and large DIF magnitudes among truly and falsely detected DIF items were calculated separately. We then explored whether classification based on DIF effect size measures correspond to true underlying DIF magnitude which was used for generation of DIF items. In the case of false DIF items, large proportion of negligible effects would be desirable. In the case of truly DIF items, DIF effect size classification should correspond to the true underlying DIF effect size which was used for generating data, i.e., for true small DIF effect size, it can be expected that classification mostly varies between negligible effect A and moderate effect B, while for true large DIF effect size, it should vary between moderate effect B and large effect C.

4.3 Results

4.3.1 Mantel-Haenszel test

Empirical rates. For small sample sizes, all correction methods were able to control type I error and rejection rates in almost all scenarios. It is a common phenomenon that with increasing sample size the rejection rates increase which can be also observed for the Mantel-Haenszel test here. When using item purification, proportion of cases when rejection rate exceeded significance level of 0.05 is lower than when using no correction method and, moreover, mean rejection rate remained near the significance level even for large sample sizes (Figure 4.2, right panel). Multiple comparison corrections and their combinations with item purification yielded rejection rates under the significance level in most of the scenarios. However, when using only multiple comparison corrections (without item purification), there was an increase in proportion of scenarios with rejection rates exceeding 0.15, more often than in scenarios with item purification only. In such a case, the BH adjustment yielded even larger mean rejection rate than item purification (Figure 4.2).

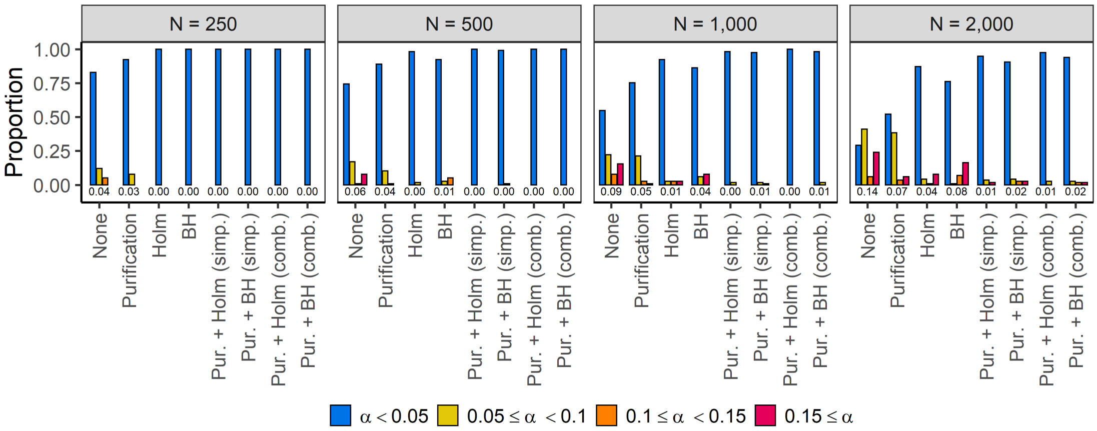


Figure 4.2: Empirical type I error and rejection rates α for the Mantel-Haenszel test. Plot shows proportions of 4 levels of rejection rates within given correction method and sample size. Values below the bars indicate mean rejection rates.

For small sample sizes, there was only small proportion of scenarios when power was sufficient and mean power of all correction methods remained on low level (Figure 4.3, left panel). However, power rates were generally increasing with the increasing sample size. While multiple comparison adjustments and their combinations with item purification gained lower power rates than item purification alone or when using no correction method, this difference was somehow softened when sample size was large. Item purification seemed to gain the largest power, followed by scenario when using no correction method and then by simple combination of BH correction and item purification (Figure 4.3).

Using item purification alone, mean number of iterations of item purification was decreasing with larger sample size and increasing with larger proportion of DIF items. Mean number of iterations varied from 5.95 to 18.94. Mixtures of item purification and adjustments for multiple comparison yielded generally lower mean number of iterations (varied from 0.21 to 2.82 for Holm's correction and from 0.39 to 4.33 for BH). While the effect of increasing proportion of DIF items was similar as for item purification, number of iterations increased with increasing sample size. The lower mean number of iterations when using combined mixture is not surprising as the mixtures generally identified less items, while this also included cases when no DIF item was identified in the initial run.

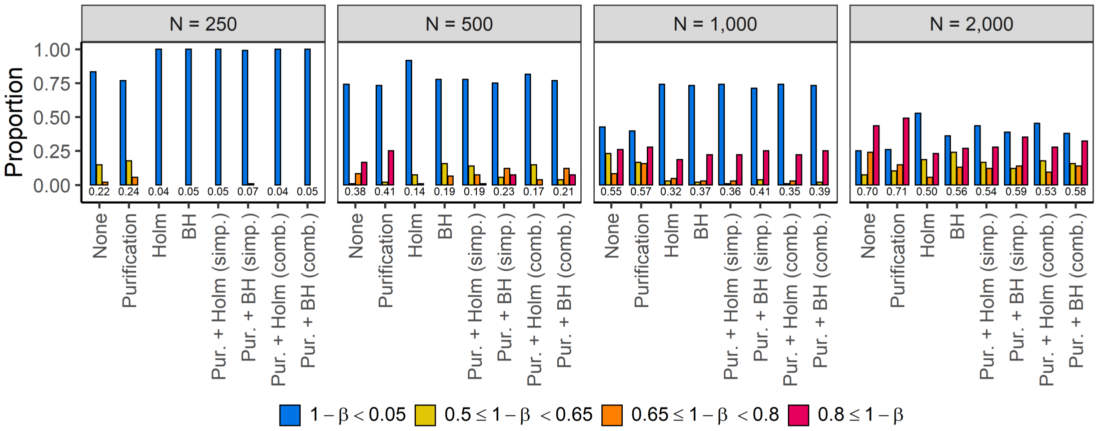


Figure 4.3: Empirical power rates $1 - \beta$ for the Mantel-Haenszel test. Plot shows proportions of 4 levels of power within given correction method and sample size. Values below the bars indicate mean power.

Beta regression model. Beta regression model confirmed increasing rejection and power rates with the increasing sample size. While there was no significant effect of item purification in power with increasing sample size, this method significantly improved control of rejection rates compared to scenario using no correction, the finding suggested also by empirical rates (displayed in Figure 4.2 and discussed above). Further, item purification improved control of rejection rates in case of large amount of DIF items and when underlying DIF magnitude was large. In case of large proportion of DIF items, item purification also significantly but only slightly increased power rate. Generally, using multiple comparison correction led to substantial decrease in power which significantly improved with increased sample size. This was also accompanied by significant decrease of rejection rates, which was somehow softened by increased sample size when using BH multiple comparison alone. All multiple comparison corrections and their combinations with item purification yielded lower values in all three summary statistics (rejection rate, type I error and power) when considering test consisting of 40 or 80 items (Figure 4.4).

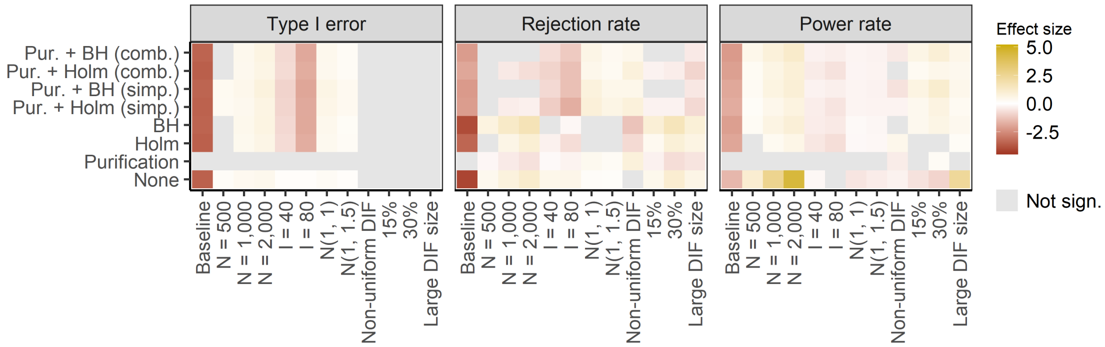


Figure 4.4: Effects of correction methods (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates by beta regression model for the Mantel-Haenszel test.

DIF effect size. For the Mantel-Haenszel test and for the smallest examined sample size of $N = 250$, all correction methods classified DIF effect size as large (category C), no matter whether item was truly functioning differently or only falsely detected and regardless of underlying DIF magnitude. Differences between corrections could be observed only for larger sample sizes. Item purification worked in similar way as when using no correction method, however, falsely detected items were more often classified to smaller DIF effect size. Classification using corrections for multiple comparisons and their combinations with item purification became appropriate with increasing sample size. Among these correction methods, BH adjustment and its mixtures corresponded slightly better to true underlying DIF magnitude, while falsely detected items were more often classified as negligible (Figure 4.5).

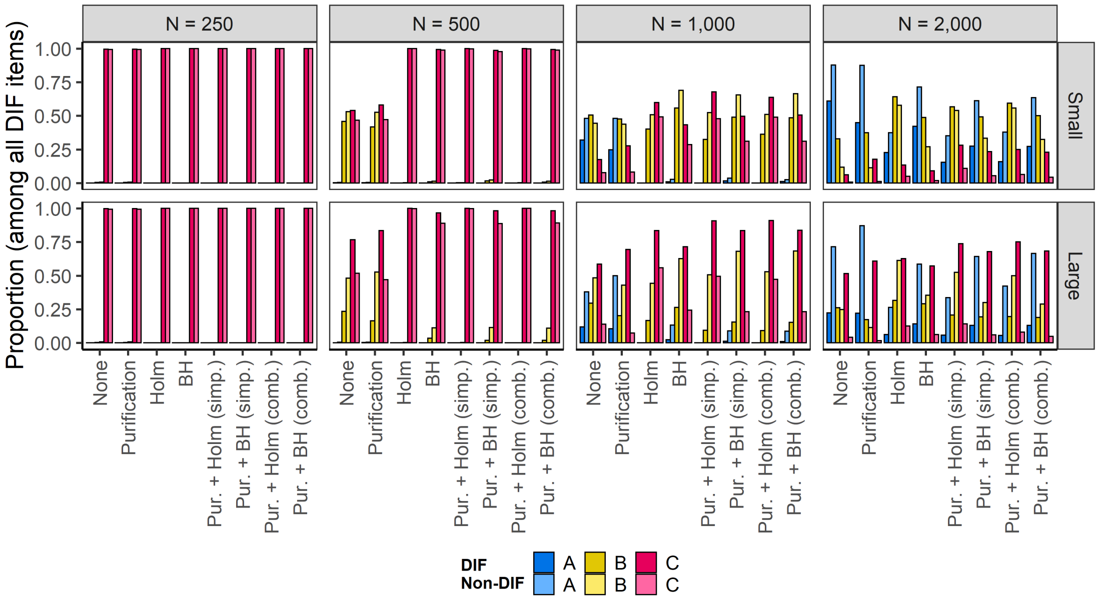


Figure 4.5: DIF effect size classification for the Mantel-Haenszel test among all truly and falsely detected items.

4.3.2 Logistic regression method

Empirical rates. When using no correction method or item purification alone, there was a large proportion of scenarios slightly exceeding significance level of 0.05, i.e., rejection rates varied mostly between 0.05 and 0.1. In both cases, proportion of severe overrun of significance level increased for large sample sizes. Also mean value of type I error and rejection rates exceeded significance level of 0.05 especially for large sample sizes (Figure 4.6, right panel). While item purification yielded larger proportion of scenarios with good control of type I error at the same time, it also gained large proportion of severe overruns resulting in slightly increased rejection rate compared to the case when using no correction method. All multiple comparison corrections and their combined mixtures with item purification were able to control for type I error. However, both simple combinations showed increased proportions of severe overrun for large sample size and thus increased mean rejection rates (Figure 4.6).

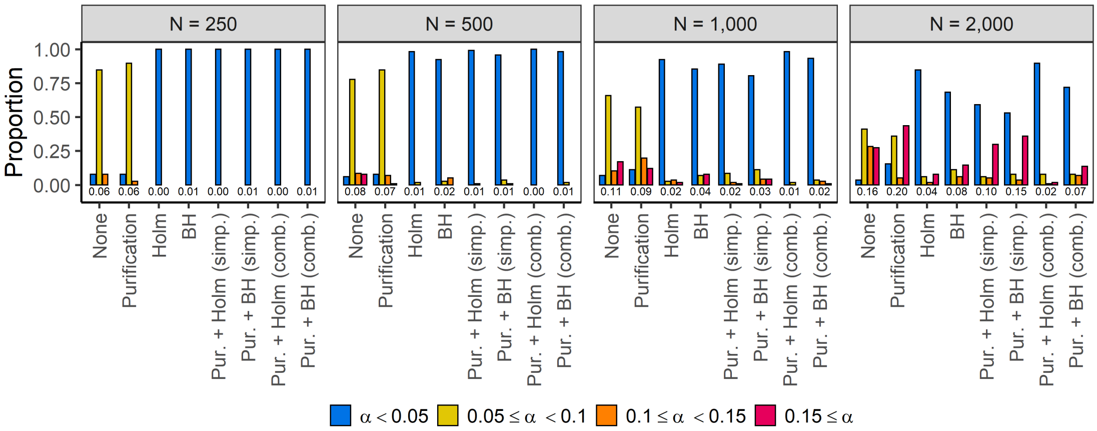


Figure 4.6: Empirical type I error and rejection rates α for the logistic regression method. Plot shows proportions of 4 levels of rejection rates within given correction method and sample size. Values below the bars indicate mean rejection rates.

None of the correction methods was able to gain sufficient power for small sample sizes. However, power rates were increasing with the increasing sample size, while item purification yielded the largest proportion of scenarios with the sufficient power (i.e., at least 80%), followed by setting with no correction method and combinations of BH correction and item purification (Figure 4.7).

Mean number of iterations of item purification was increasing with the increasing sample size and with the increasing proportion of DIF items in all methods. However, item purification alone (and its simple combinations) yielded larger mean number of iterations (varied from 2.46 to 9.41) than when used in combined setting with multiple comparison corrections (varied from 0.23 to 2.77 for Holm's correction and from 0.36 to 5.72 for BH).

Beta regression model. Similarly to the Mantel-Haenszel test, item purification in the logistic regression method improved rejection rate control in case of large DIF effect size and large proportion of DIF items. However, unlike in the Mantel-Haenszel test, the power of the logistic regression DIF detection

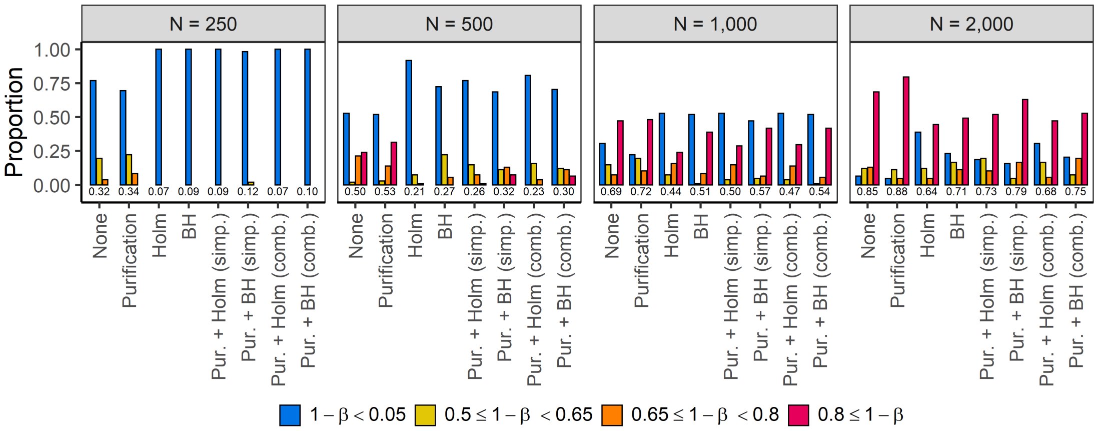


Figure 4.7: Empirical power rates $1 - \beta$ for the logistic regression method. Plot shows proportions of 4 levels of power within given correction method and sample size. Values below the bars indicate mean power.

method increased when sample size increased. On the other hand, the power slightly decreased when DIF was non-uniform and control of rejection rates worsened when latent trait of focal groups was drawn from normal distribution with different mean and variance. Multiple comparison corrections and their combinations with item purification indicated generally lower rejection and power rates. While there were no crucial differences between the correction methods in terms of power and their interactions with other factors, their control for rejection rates differed. Especially, purification followed by multiple comparison correction in each step, either BH or Holm's, performed better control when sample size increased (Figure 4.8).

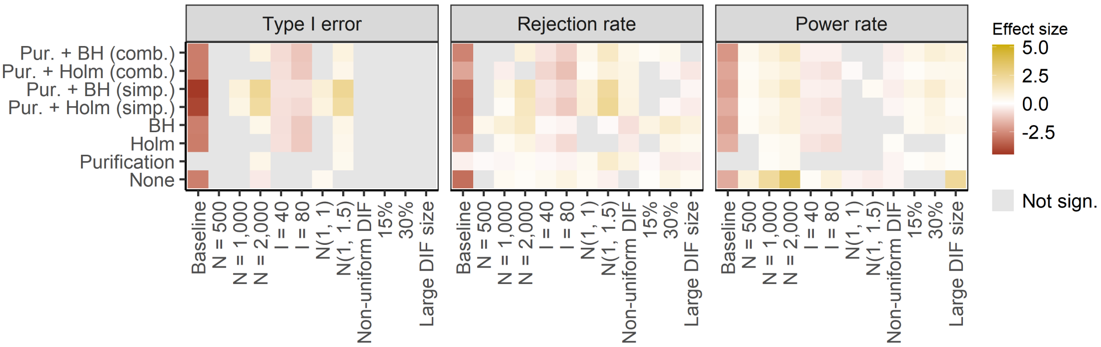


Figure 4.8: Effects of correction methods (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates by beta regression model for the logistic regression method.

DIF effect size. Classification of the logistic regression method using ΔR_i^2 seemed to reflect true underlying DIF magnitudes only for small sample sizes. With the increasing sample size, almost all DIF items were classified as of negligible effect (category A), which is undesired for truly DIF items. Moreover, the classification did not distinguish between the true DIF magnitudes at all, whether they were small or large. Purification and especially its combinations with multiple comparison corrections at least in some cases classified DIF effect size into the category B when underlying DIF magnitude was large, however even this effect became negligible with the increasing sample size. Generally, multiple comparison corrections and their combinations with item purification overvalued small DIF magnitude (Figure 4.9).

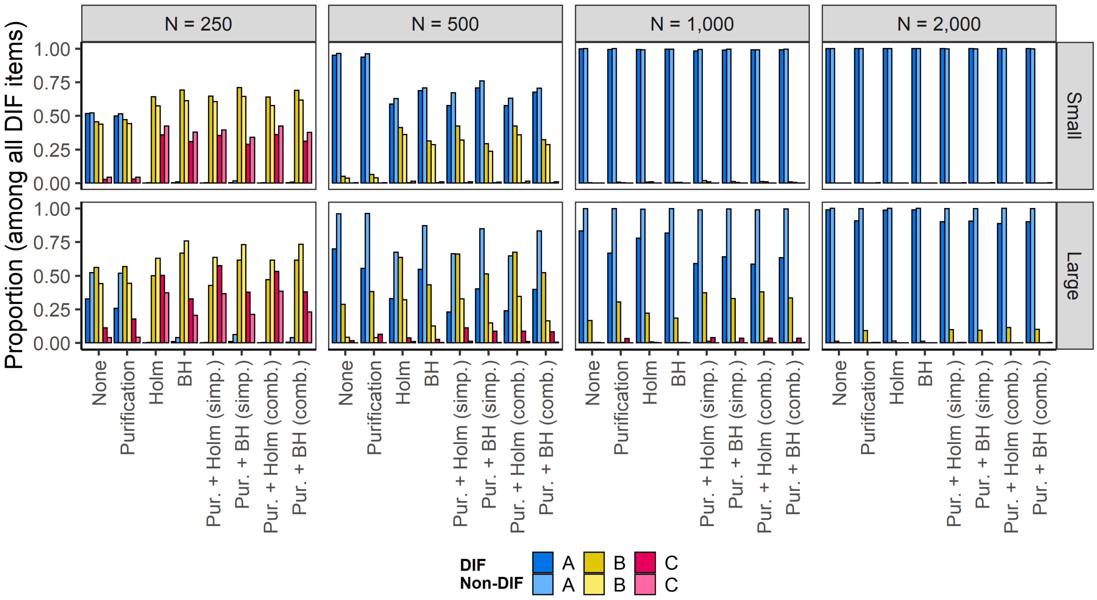


Figure 4.9: DIF effect size classification for the logistic regression method among all truly and falsely detected items.

4.3.3 SIBTEST

Empirical rates. All correction methods improved control of type I error even in case of large sample size, when increased rejection rates were observed using no correction method. While item purification and BH adjustment itself slightly overrun a significance level, all four combinations of multiple comparison corrections and item purification kept rejection rates under the significance level for almost all scenarios (Figure 4.10).

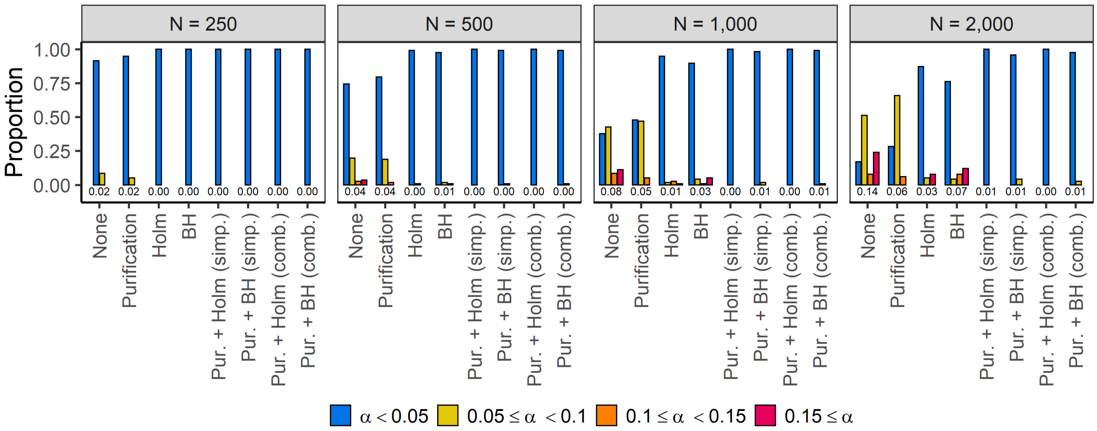


Figure 4.10: Empirical type I error and rejection rates α for the SIBTEST. Plot shows proportions of 4 levels of rejection rates within given correction method and sample size. Values below the bars indicate mean rejection rates.

Sufficient power was gained only for larger sample sizes regardless of used correction method. In such a case, item purification performed slightly better than other procedures, followed by scenario when using no correction method and then by both combinations of BH multiple comparison correction and item purification (Figure 4.11).

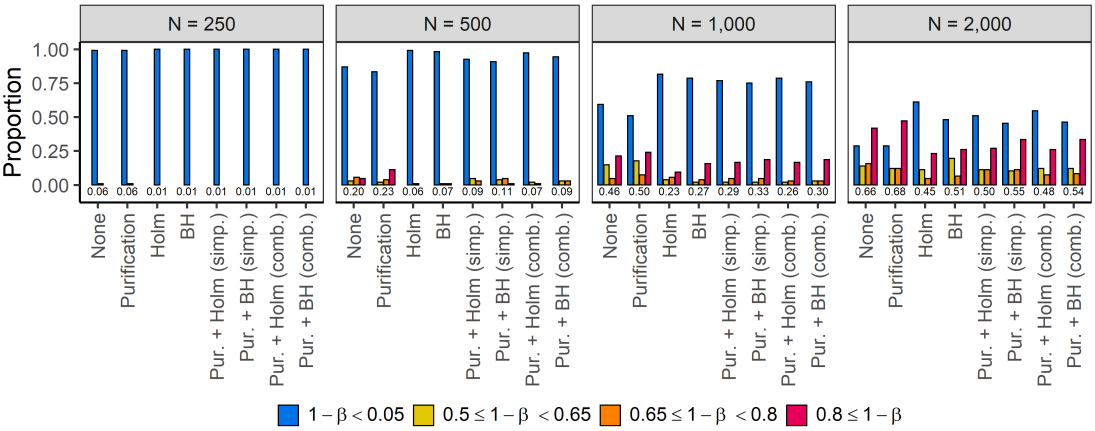


Figure 4.11: Empirical power rates $1-\beta$ for the SIBTEST. Plot shows proportions of 4 levels of power within given correction method and sample size. Values below the bars indicate mean power.

Similarly as for the logistic regression method, mean number of iterations of item purification in the SIBTEST method was increasing with the increasing sample size and with the increasing proportion of DIF items in all methods. Item purification alone (and its simple combinations) yielded again larger mean number of iterations (varied from 8.67 to 25.42) than when used in combined setting with the multiple comparison corrections (varied between 0.24 and 6.67 for Holm's correction and between 0.38 and 9.91 for BH).

Beta regression model. Similarly to previous DIF detection methods, the SIBTEST also showed increasing power with the increasing sample size. However, it seemed that the SIBTEST struggled when number of items increased, as the power and rejection rates decreased rapidly. Moreover, it was somehow more difficult to identify DIF when larger proportion of DIF items was present. This was slightly better when item purification with the BH multiple comparison correction (either simple or combined) was applied. Item purification again performed better control of rejection rate in case of large proportion of DIF items, large DIF effect size, and also increased sample size. Its effect on power was, however, limited (Figure 4.12).

DIF effect size. Classification of the SIBTEST using $|\hat{\omega}_i|$ measure was somehow non-informative for small sample sizes, as almost all DIF items were classified into the category C regardless of the correction method or the true underlying DIF magnitude. Differences between the methods were observed only for larger sample sizes (i.e., $n > 500$, Figure 4.13, right panels). In that case using no correction method or item purification yielded similar classification, which with the increasing sample size corresponded to the underlying DIF effect magnitude. Item purification seemed to more precisely classify falsely detected items as those

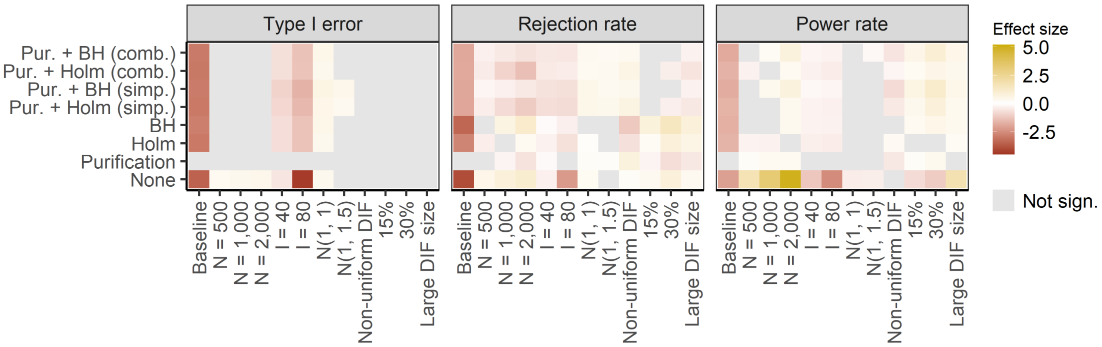


Figure 4.12: Effects of correction methods (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates by beta regression model for the SIBTEST.

with negligible effect. When using multiple comparison corrections and their combinations with item purification, classification assigned higher category for falsely detected items and also for those truly detected (Figure 4.13).

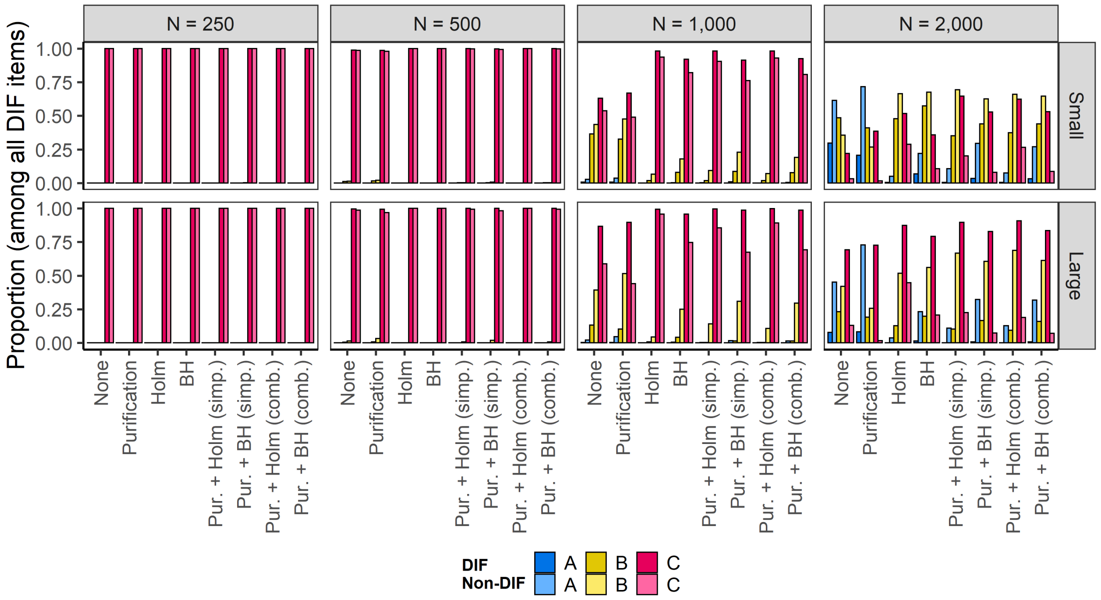


Figure 4.13: DIF effect size classification for the SIBTEST among all truly and falsely detected items.

4.4 Discussion

In the simulation study we investigated the impact of item purification and adjustments for multiple comparison and their combination on the properties of DIF detection procedures, specifically on their type I error, rejection, and power rates. We considered two combination settings – item purification followed by adjustment in a final run and item purification followed by adjustment after each iteration. To evaluate results of the simulation study, we used empirical values of summary statistics, beta regression model, and also DIF effect sizes.

In general, the results suggest that all three DIF detection methods when applied item by item and without any correction for multiple comparisons lead to rejection rates somehow exceeding the nominal significance level in some scenarios, especially with large sample sizes, which has already been noted by many authors including DeMars (2009), Güler and Penfield (2009), and Herrera and Gómez (2008). Adjustments reduced both rejection rates and type I error, however, this reduction was also accompanied by a decrease of the power. Kim and Oshima (2013) already noted that adjustments caused a decrease in power to some extent, however in this study we demonstrate how in some scenarios power rates are no longer sufficient. That means, in general, fewer items are detected as DIF, and some potentially unfair items may remain undetected.

The effect of purification has been researched earlier by many authors including Candell and Drasgow (1988), Navas-Ara and Gómez-Benito (2002), and Wang and Su (2004). In our study, we confirmed some improvements in DIF detection when using item purification primarily in the Mantel-Haenszel test and the SIBTEST method. Generally, we observed improvement of DIF detection when larger proportion of DIF items was present which was also showed for example by French and Maller (2007). However, item purification yielded increased type I error and rejection rates when applying within logistic regression especially when large sample size was considered, meaning that more items need to be assessed by content experts which may give an impression of suspicious test.

Both settings of item purification combined with adjustment for multiple comparisons improved control of type I error and rejection rates in almost all scenarios in the Mantel-Haenszel and the SIBTEST method. Applying BH correction resulted in slightly larger power rates than in case of combinations with Holm's adjustment. However, item purification alone performed better in both above mentioned DIF detection methods. In contrast, the logistic regression benefited more from combined mixture of item purification and BH adjustment which was followed by BH applied alone.

Weaker performance for small sample sizes is not surprising as all considered DIF detection methods are asymptotic. Although the Mantel-Haenszel test generally works well even for small sample sizes, its inability to detect non-uniform DIF was already stressed in previous studies (e.g., Swaminathan & Rogers, 1990). This is also the case for the SIBTEST (see, e.g., Li & Stout, 1996; Chalmers, 2018). It should be noted that collapsed summary statistics presented here also include a non-uniform DIF effect and small DIF effect sizes. As mentioned above, these factors may influence summary statistics significantly (i.e., reduction of power), which may give an impression that the performance of DIF detection procedures and correction methods is somewhat lower than reported in literature.

In summary, good control of type I error and rejection rate together with decent power rates and adequate classification of DIF effect size suggest item purification alone and its simple combination with BH correction to be promising controlling procedures when applying the Mantel-Haenszel test to identify DIF. For large sample sizes, DIF detection based on logistic regression model may benefit from using combined item purification with multiple comparison corrections as they reported decent power with a good control of rejection rates. Finally, detecting DIF using the SIBTEST method may profit from using item purification as it improved control of type I error while it also gained larger power. While our simulation study contained various factors in a complex setting, general conclusions and recommendations cannot be made and further studies, preferably including meta analysis, need to be performed.

Besides empirical summary statistics and beta regression models, DIF effect sizes were computed to assess magnitude of DIF among truly and falsely detected items. For small sample sizes, DIF effect sizes were uninformative in the Mantel-Haenszel test and the SIBTEST. Together with the low power of both methods in such scenarios, this suggests that if the item was detected as DIF at all, it was classified into category C – large DIF magnitude. Moreover, classification became appropriate when using no correction method or item purification alone with smaller sample sizes than in case of multiple comparison adjustments or their mixtures with item purification. DIF effect size classification based on ΔR_i^2 worked in a different way for the logistic regression. While it was sufficient for small sample sizes, with increasing sample size it became senseless as all items were categorized as those with negligible effect.

There are some limitations of this simulation study which need to be considered. First, only a limited number of DIF detection methods were used, excluding, for example differential functioning of items and tests framework (Raju, van der Linden, & Fleer, 1995) or methods based on IRT models such as Lord's or Raju's tests (Lord, 1980; Raju, 1988, 1990). Simulation study showed that different correction techniques have different effect on DIF detection methods, thus conclusions need to be made with respect to only those used in this work. Further studies are needed to explore effect of correction methods in above mentioned and other DIF detection approaches. Second, in our simulation study we determined the significance of simulation factors via beta regression models, where only double interactions were considered. However, increasing complexity of the study design goes in hand with increased complexity of the results. Thus any further extension to the study design may complicate interpretability and thus lower the readability of the results.

While some of the simulation settings were inspired by previous studies to allow comparing the results, our study is more complex and its design goes beyond previous studies including Kim and Oshima (2013) by also incorporating non-uniform DIF, a larger variety of sample sizes, and various distributions of ability levels for the reference group. Our study covers the current gap in the DIF literature as it allows for joint evaluation of properties of different correction types – purification and correction for multiple comparisons. Moreover, we considered two settings of their combinations, which to our best knowledge, have not yet been explored in literature. Despite its limitations, this study offers a detailed assessment of controlling procedures in DIF detection and a deeper insight which may be helpful to researchers and practitioners when testing for DIF.

Conclusion

The thesis dealt with the topic of DIF, a phenomenon that can arise in various contexts of educational, psychological, or health-related measurements. We focused on non-IRT statistical models and methods which can be used for DIF detection.

Chapter 1 introduced generalized logistic regression models for DIF detection among binary items which allow for possibility of guessing and/or inattention when answering. We described several methods and algorithms to estimate item parameters, namely the nonlinear least squares, the maximum likelihood, the EM algorithm, and the newly proposed algorithm based on parametric link function. We offered two simulation studies. The first simulation study, already published, evaluated the properties of the newly proposed procedure based on the nonlinear models and compared it to the commonly used DIF detection methods (Drabinová & Martinková, 2017). The second simulation study, which is planned for publication in Hladká, Brabec, and Martinková (2021), compared presented methods to estimate item parameters in the nonlinear models. Future work might contain an extension of the second simulation study including an improvement of the specification of starting values. Finally, we presented the implementation of these methods within the statistical software **R** and its package **difNLR** (Hladká & Martinková, 2020).

Chapter 2 focused on generalized logistic regression models for DIF and DDF detection among ordinal and nominal models, namely the cumulative logit model, the adjacent category logit model, and the multinomial model. Besides providing detailed model specifications, the maximum likelihood method to estimate item parameters was described and the implementation of these models into the **difNLR** package (Hladká & Martinková, 2020) was presented.

Chapter 3 proposed nonparametric comparison of ICCs for DIF detection. We built on work by Srihera and Stute (2010) and we adapted and improved their approach in some scenarios by proposing the estimate of the optimal weights and by evaluating asymptotic properties of the underlying test statistic using wild bootstrap. Future work may include theoretical derivation of the asymptotic distribution when the matching criterion is discrete as was the case in offered simulation study. Further extension of the simulation study may provide more convincing examples of situations when the logistic regression is not capable to detect differences between ICCs while the newly proposed method is, including simulated as well as real data. These improvements are planned for publication in Hladká and Martinková (2021).

Chapter 4 discussed further issues in DIF detection including item purification and multiple comparison corrections. We newly proposed the combination of both approaches in the two settings – item purification followed by correction in final step and item purification followed by correction after each iteration. The presented complex simulation study is planned for publication in Hladká, Martinková, and Magis (2021).

In summary, parametric models discussed in the thesis can be seen as computationally less demanding proxies to more complex IRT models, while they account for possibility of guessing and/or inattention, or for polytomous items.

The newly proposed nonparametric approach, on the other hand, does not require a parametric form of the mean function and it seems to have a potential to outperform parametric approaches when several inflection points are present. In this thesis, we focused on proper statistical specification, interpretation, estimation procedures, and asymptotic properties of the newly proposed methods, but we also discussed further issues in DIF detection. Finally, we offered empirical proofs of appropriateness of the methods by simulation studies and, importantly, also their practical implementation in the statistical software. As such, the thesis extends the existing methods for DIF detection and shows how the newly proposed methods may be used in practice.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (Second ed.). John Wiley & Sons. doi: 10.1002/9780470594001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–105. doi: 10.1111/j.1745-3984.1973.tb00787.x

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), 1–8. doi: 10.1007/s13398-014-0173-7.2

Basu, A., & Rathouz, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, *6*(1), 93–109. doi: 10.1093/biostatistics/kxh020

Battauz, M. (2019). On Wald tests for differential item functioning detection. *Statistical Methods & Applications*, *28*(1), 103–118. doi: 10.1007/s10260-018-00442-w

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 (Statistical Methodology)*, *57*(1), 289–300. doi: 10.2307/2346101

Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. *Journal of Educational and Behavioral Statistics*, *41*(6), 559–592. doi: 10.3102/1076998616659371

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, *22*(1), 134–167. doi: 10.1111/j.1744-7348.1935.tb07713.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/BF02293801

Bock, R. D., & Moustaki, I. (2006). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 469–513). Elsevier. doi: 10.1016/S0169-7161(06)26005-X

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313–333. doi: 10.1111/j.1745-3984.2006.00019.x

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. doi: 10.1137/0916069

Cabrera, J. L. O. (2018). locpol: Kernel local polynomial regression [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=locpol` (R package version 0.7-0)

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253–260. doi: 10.1177/014662168801200304

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2018). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika*, *83*(2), 376–386. doi: 10.1007/s11336-017-9583-8

Chen, C.-T., & Hwu, B.-S. (2018). Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: The PISA example. *Applied Psychological Measurement*, *42*(3), 206–220. doi: 10.1177/0146621617726786

Cho, S.-J., Suh, Y., & Lee, W.-Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48–61. doi: 10.1111/emip.12093

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269–279. doi: 10.1207/s15324818ame0604_2

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34*(2), 149–170. doi: 10.3102/1076998607313923

Dennis, J. E. J., Gay, D. M., & Welsch, R. E. (1981). An adaptive nonlinear least-squares algorithm. *Transactions on Mathematical Software*, *7*(3), 348–368. doi: 10.1145/355958.355965

Dette, H., & Neumeyer, N. (2001). Nonparametric analysis of covariance. *The Annals of Statistics*, *29*(5), 1361–1400. doi: 10.1214/aos/1013203458

Dinse, G. E. (2011). An EM algorithm for fitting a four-parameter logistic model to binary dose-response data. *Journal of Agricultural, Biological, and Environmental Statistics*, *16*(2), 221–232. doi: 10.1007/s13253-010-0045-3

Doob, J. L. (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, *6*(3), 160–169. doi: 10.1214/aoms/1177732594

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics*, *21*(4), 333–363. doi: 10.3102/10769986021004333

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. doi: 10.1111/jedm.12158

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158. doi: 10.1137/1114019

Fikis, D. R., & Oshima, T. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement*, *77*(3), 415–428. doi: 10.1177/0013164416645844

Flach, N. (2014). *Generalized linear models with parametric link families in R* (Unpublished master's thesis). Technische Universität München, Department of Mathematics, München.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373–393. doi: 10.1177/0013164406294781

Gay, D. (n.d.). *Port library documentation.* Retrieved from `http://www.netlib.org/port/` (Accessed: 2020-06-15)

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314–329. doi: 10.1111/j.1745-3984.2009.00083.x

Hall, P., & Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, *85*(412), 1039–1049. doi: 10.1080/01621459.1990.10474974

Härdle, W. (1990). *Applied nonparametric regression* (No. 19). Cambridge University Press.

Hardle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, *21*(4), 1926–1947. doi: 10.1214/aos/1176349403

Herrera, A.-N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739. doi: 10.1007/s11135-006-9065-z

Hladká, A., Brabec, M., & Martinková, P. (2021). *Estimation in generalized logistic regression.* (In preparation for submission)

Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, *12*(1), 300–323. doi: 10.32614/RJ-2020-014

Hladká, A., & Martinková, P. (2021). *Nonparametric comparison of regression curves for DIF detection.* (In preparation for submission)

Hladká, A., Martinková, P., & Magis, D. (2021). *Issues and practice in detection of differential item functioning: Applying item purification, correction for multiple comparisons, or combination of both?* (In preparation for resubmission)

Hogg, R. V., McKean, J., & Craig, A. T. (2018). *Introduction to mathematical statistics* (Eighth ed.). Pearson Education.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, *104*(1), 145. doi: 10.1037/0033-2909.104.1.145

Holland, P. W. (1985). On the study of differential item performance without IRT. In *Proceedings of the 27th Annual Conference of the Military Testing Association* (pp. 282–287). San Diego, CA.

Holland, P. W., & Thayer, D. T. (1985). An alternate definition of the ETS delta scale of item difficulty. *ETS Research Report Series*, *1985*(2), i–10. doi: 10.1002/j.2330-8516.1985.tb00128.x

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. doi: 10.1207/S15324818AME1404_2

Kaiser, M. S. (1997). Maximum likelihood estimation of link function parameters. *Computational Statistics & Data Analysis*, *24*(1), 79–87. doi: 10.1016/S0167-9473(96)00055-2

Khalid, M. N., & Glas, C. A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. doi: 10.1016/j.measurement.2013.12.019

Kim, J., & Oshima, T. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470. doi: 10.1177/0013164412467033

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test. *ETS Research Report Series*, *1985*(2), 1–64. doi: 10.1002/j.2330-8516.1985.tb00119.x

Kopf, J., Zeileis, A., & Strobl, C. (2013). *Anchor methods for DIF detection: A comparison of the iterative forward, backward, constant and all-other anchor class* (Tech. Rep.). Munich, Germany: Department of Statistics, LMU Munich.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. doi: 10.1177/0013164414529792

Lautenschiager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, *12*(4), 365–376. doi: 10.1177/014662168801200404

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677. doi: 10.1007/BF02294041

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (First ed.). New York, NY: Routledge.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302–321. doi: 10.1111/j.2044-8317.2011.02025.x

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A counter-example with Angoff's delta plot. *Educational and Psychological Measurement*, *73*(2), 293–311. doi: 10.1177/0013164412451903

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. doi: 10.1080/

15305058.2011.602810

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 255–285. doi: 10.1214/aos/1176349025

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute*, *22*(4), 719–748. doi: 10.1093/jnci/22.4.719

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. doi: 10.32614/RJ-2018-074

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. doi: 10.1187/cbe.16-10-0307

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286. doi: 10.1016/j.learninstruc.2019.101286

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second ed.). Chapman & Hall.

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., . . . Wright, A. (2017). Development and validation of the homeostasis concept inventory. *CBE-Life Sciences Education*, *16*(2), ar35. doi: 10.1187/cbe.16-10-0305

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research.* Walter de Gruyter.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i–30. doi: 10.1002/j.2333-8504.1992.tb01436.x

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. doi: 10.2307/2337038

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274. doi: 10.1177/014662169602000306

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, *18*(1), 9. doi: 10.1027//1015-5759.18.1.9

Neumeyer, N., & Dette, H. (2003). Nonparametric comparison of regression curves: An empirical process approach. *The Annals of Statistics*, *31*(3), 880–920. doi: 10.1214/aos/1056562466

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Second ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 125–167). Elsevier. doi: 10.1016/S0169-7161(06)26005-X

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38–49. doi: 10.1111/j.1745-3992.2009.01135.x

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*(1), 23–37. doi: 10.1177/014662169501900104

Pratt, J. W. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, *76*(373), 103–106. doi: 10.1080/01621459.1981.10477613

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(1), 15–24. doi: 10.2307/2346405

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. doi: 10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. doi: 10.1177/014662169001400208

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. doi: 10.1177/014662169501900405

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630. doi: 10.1007/BF02294494

Ramsay, J. O. (2000). *Testgraf: A program for the graphical analysis of multiple choice test and questionnaire data.*

Richards, F. S. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *23*(2), 469–475. doi: 10.1111/j.2517-6161.1961.tb00430.x

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185. doi: 10.1037/1082-989X.8.2.185

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R.* Springer, New York, NY. doi: 10.1007/978-0-387-09616-2

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*(2), 215–230. doi: 10.1111/j.1745-3984.1996.tb00490.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(Suppl 1). doi: 10.1007/BF03372160

Scallan, A., Gilchrist, R., & Green, M. (1984). Fitting parametric link functions in generalised linear models. *Computational Statistics & Data Analysis*, *2*(1), 37–49. doi: 10.1016/0167-9473(84)90031-8

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression.* John Wiley & Sons. doi: 10.1002/0471725315

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584. doi: 10.1146/annurev.ps.46.020195.003021

Shealy, R., & Stout, W. (1993). 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. doi: 10.1007/BF02294572

Siebert, C. F. (2013). *Differential item functioning identification strategy for items with dichotomous responses using the item information curve: A weighted area method (WAM)* (Unpublished doctoral dissertation). The Florida State University.

Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, *101*(9), 2039–2059. doi: 10.1016/j.jmva.2010.05.001

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement*, *53*(4), 403–430. doi: 10.1111/jedm.12123

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (p. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge University Press.

van de Water, E. (2014). *A meta-analysis of type I error rates for detecting differential item functioning with logistic regression and Mantel-Haenszel in Monte Carlo studies* (Unpublished doctoral dissertation). Georgia State University.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer. doi: 10.1007/978-0-387-21706-2

Wand, M. (2019). KernSmooth: Functions for kernel smoothing supporting Wand & Jones (1995) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=KernSmooth` (R package version 2.23-16)

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113–144. doi: 10.1207/s15324818ame1702_2

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. doi: 10.1177/0146621603259902

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Second ed.). Springer-Verlag New York. Retrieved from `https://ggplot2.tidyverse.org`

Wickham, H., Hester, J., & Chang, W. (2020). devtools: Tools to make developing R packages easier [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=devtools` (R package version

2.3.0)

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi: 10.1177/0146621607314044

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295. doi: 10.1214/aos/1176350142

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34. doi: 10.18637/jss.v032.i10

Zumbo, B. D. (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.

List of figures

List of tables

Glossaries

Abbreviations

AIC Akaike's Information Criterion.

AM Area Measure.

BH Benjamini-Hochberg.

BIC Bayesian Information Criterion.

CTT Classical Test Theory.

DDF Differential Distractor Functioning.

DIF Differential Item Functioning.

EM Expectation-Maximization.

GMAT Graduation Management Admission Test.

HCI Homeostasis Concept Inventory.

ICC Item Characteristic Curve.

IRT Item Response Theory.

MSD Mean Squared Deviation.

MSE Mean Squared Error.

PL Parameter Logistic.

RSS Residual Sum of Squares.

SA Signed Area.

SIBTEST Simultaneous Item Bias Test.

UA Unsigned Area.

WAM Weighted Area Measure.

Nomenclature

i Index related to item.

I The number of items in multi-item test.

p Index related to person or respondent.

n The number of respondents.

∂ Partial derivative.

$\xrightarrow[n\to\infty]{\mathcal{D}}$ Convergence in distribution.

$\xrightarrow[n\to\infty]{P}$ Convergence in probability.

$a \overset{!}{=} b$ a shall be equal to b.

List of publications

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. doi: 10.1111/jedm.12158

Hladká, A., Brabec, M., & Martinková, P. (2021). *Estimation in generalized logistic regression.* (In preparation for submission)

Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, *12*(1), 300–323. doi: 10.32614/RJ-2020-014

Hladká, A., & Martinková, P. (2021). *Nonparametric comparison of regression curves for DIF detection.* (In preparation for submission)

Hladká, A., Martinková, P., & Magis, D. (2021). *Issues and practice in detection of differential item functioning: Applying item purification, correction for multiple comparisons, or combination of both?* (In preparation for resubmission)

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. doi: 10.32614/RJ-2018-074

Martinková, P., Drabinová, A., & Houdek, J. (2017). ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů. *TESTFÓRUM*, *6*(9), 16–35. doi: 10.5817/TF2017-9-129

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. doi: 10.1187/cbe.16-10-0307

Martinková, P., & Hladká, A. (2021). *Computational aspects of psychometric methods in education, psychology, and health: With examples in R.* CRC Press. (In preparation)

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286. doi: 10.1016/j.learninstruc.2019.101286

A. Appendices

A.1 Integrable dominating functions for the nonlinear least squares

To verify the regularity condition [R3] for the nonlinear least squares method described in Section 1.2.1, we first need to calculate the second partial derivatives of the $\psi_{ik}(y, x, g; \boldsymbol{\gamma}_i)$. For simplicity, we will now consider only the reference group, i.e. $g = 0$. Calculation would be analogous for the focal group $g = 1$.

$$\begin{aligned}\left|\frac{\partial^2 \psi_{ik}(y, x, g = 0; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij} \partial \gamma_{il}}\right| = 2 \Bigg| &\frac{\partial \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ik}} \frac{\partial^2 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij} \partial \gamma_{il}} + \frac{\partial \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij}} \frac{\partial^2 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \partial \gamma_{il}} \\ &+ \frac{\partial \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{il}} \frac{\partial^2 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \partial \gamma_{ij}} \\ &- (y - \pi(x; \boldsymbol{\gamma}_i)) \frac{\partial^3 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \partial \gamma_{ij} \partial \gamma_{il}} \Bigg|,\end{aligned} \tag{A.1}$$

where

$$\pi_i(x, \boldsymbol{\gamma}_i) = c_i + (d_i - c_i) \frac{e^{a_i(x - b_i)}}{1 + e^{a_i(x - b_i)}} = c_i + (d_i - c_i)\phi(x; a_i, b_i).$$

Using the fact that $\phi(x; a_i, b_i) \in (0, 1)$, it is easy to see that

$$\phi(x; a_i, b_i)(1 - \phi(x; a_i, b_i)) \leq \phi(x; a_i, b_i) < 1$$

and, moreover,

$$|x - b_i|\phi(x; a_i, b_i)(1 - \phi(x; a_i, b_i)) \leq |x - b_i| \leq x^2 + A_1,$$

for some $A_1 \in \mathbb{R}$. Thus all first partial derivatives are dominated by

$$\left|\frac{\partial \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ik}}\right| \leq x^2 + A,$$

where $A \in \mathbb{R}$ is sufficiently large, $k = 1, \dots, 8$.

Similarly, for the second partial derivatives (1.16), using in addition the fact that $|1 - 2\phi(x; a_i, b_i))| < 1$, we have

$$\begin{aligned}|x - b_i|\phi(x; a_i, b_i)(1 - \phi(x; a_i, b_i))|1 - 2\phi(x; a_i, b_i))| &\leq |x - b_i| \leq x^2 + B_1, \\ (x - b_i)^2\phi(x; a_i, b_i)(1 - \phi(x; a_i, b_i))|1 - 2\phi(x; a_i, b_i))| &\leq (x - b_i)^2 \leq x^4 + B_2,\end{aligned}$$

for some $B_1, B_2 \in \mathbb{R}$, and, therefore,

$$\left|\frac{\partial^2 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij} \partial \gamma_{il}}\right| \leq x^4 + B,$$

for $B \in \mathbb{R}$ sufficiently large, $j, l = 1, \dots, 8$, and thus

$$\left|\frac{\partial \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij}}\right| \left|\frac{\partial^2 \pi(x; \boldsymbol{\gamma}_i)}{\partial \gamma_{ij} \partial \gamma_{il}}\right| \leq (x^2 + A)(x^4 + B) \tag{A.2}$$

are also dominated.

Finally, the third partial derivatives of the $\pi(x, \boldsymbol{\gamma}_i)$ with respect to parameters c_i and d_i are equal to zeros or contain terms analogous to those in the first and the second partial derivatives which have been already shown to be bounded. Thus, all we need to show is that the third partial derivatives with respect to parameters a_i and b_i are dominated by some integrable function. We use the fact that polynomial of $\phi(x; a_i, b_i)$ is a bounded function:

$$\begin{aligned}
\left|\frac{\partial^3\pi(x;\boldsymbol{\gamma}_i)}{\partial a_i^3}\right| &= \Big|(d_i-c_i)(x-b_i)^3\Big[(1-\phi(x;a_i,b_i)) - 7(1-\phi(x;a_i,b_i))^2 \\
&\quad + 12(1-\phi(x;a_i,b_i))^3 - 6(1-\phi(x;a_i,b_i))^4\Big]\Big| \le C_1x^4 + C_2, \\
\left|\frac{\partial^3\pi(x;\boldsymbol{\gamma}_i)}{\partial a_i^2\partial b_i}\right| &= \Big|(d_i-c_i)a_i(x-b_i)^3\phi(x;a_i,b_i)(1-\phi(x;a_i,b_i)) \\
&\quad \Big[1-\ 14(1-\phi(x;a_i,b_i))^2 + 36(1-\phi(x;a_i,b_i))^3 \\
&\quad -\ 24(1-\phi(x;a_i,b_i))^4\Big] + (d_i-c_i)(x-b_i)^2(1-\phi(x;a_i,b_i)) \\
&\quad \Big[-\ 3 + 21(1-\phi(x;a_i,b_i)) - 36(1-\phi(x;a_i,b_i))^2 \\
&\quad + 18(1-\phi(x;a_i,b_i))^3\Big]\Big| \le C_3x^4 + C_4, \\
\left|\frac{\partial^3\pi(x;\boldsymbol{\gamma}_i)}{\partial a_i\partial b_i^2}\right| &= \Big|(d_i-c_i)(x-b_i)a_i^2\phi(x;a_i,b_i)(1-\phi(x;a_i,b_i))\,(1-6\phi(x;a_i,b_i) \\
&\quad + 6\phi^2(x;a_i,b_i)\Big) + 2a_i\phi(x;a_i,b_i)(1-\phi(x;a_i,b_i))(1-\phi(x;a_i,b_i) \\
&\quad +\ \ \phi^2(x;a_i,b_i))\Big| \le C_5x^2 + C_6, \\
\left|\frac{\partial^3\pi(x;\boldsymbol{\gamma}_i)}{\partial b_i^3}\right| &= \Big|(d_i-c_i)a_i^3\Big[(1-\phi(x;a_i,b_i)) - 7(1-\phi(x;a_i,b_i))^2 \\
&\quad + 12(1-\phi(x;a_i,b_i))^3 - 6(1-\phi(x;a_i,b_i))^4\Big]\Big| \le C_7 \in \mathbb{R}.
\end{aligned}$$

In summary, for the third partial derivatives of $\pi(x, \boldsymbol{\gamma}_i)$ we have

$$\left|\frac{\partial^3\pi(x;\boldsymbol{\gamma}_i)}{\partial\gamma_{ik}\partial\gamma_{ij}\partial\gamma_{il}}\right| \le Cx^4 + D, \tag{A.3}$$

where $C, D \in \mathbb{R}$ are sufficiently large, $k, j, l = 1, \ldots, 8$.

In summary, combining (A.2) and (A.3), the triangle inequality for the (A.1) and the fact that $y - \pi(x; \boldsymbol{\gamma}_i) \in (0, 1)$, the second partial derivatives of $\psi_{ik}(y, x, g = 0, \boldsymbol{\gamma}_i)$ are dominated by

$$\left|\frac{\partial^2\psi_{ik}(y,x,g=0;\boldsymbol{\gamma}_i)}{\partial\gamma_{ij}\partial\gamma_{il}}\right| \le K_1x^6 + K_2x^4 + K_3x^2 + K_4 =: \ddot{\boldsymbol{\psi}}(y,x,g),$$

where $K_1, K_2, K_3, K_4 \in \mathbb{R}$ are sufficiently large, $k, j = 1, \ldots, 8$, and $\ddot{\boldsymbol{\psi}}(y, x, g)$ is an integrable function. This completes the proof that the condition [R3] holds.

A.2 Integrable dominating functions for the maximum likelihood

To verify that the regularity condition [R4*] holds for the maximum likelihood method described in Section 1.2.2, we start with the first partial derivatives of $f(y|x, g, \boldsymbol{\gamma}_i)$.

$$\begin{aligned}
\left|\frac{\partial f(y|x, g, \boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right| &= \left|\frac{y - \pi(x, g; \boldsymbol{\gamma}_i)}{\pi(x, g; \boldsymbol{\gamma}_i)(1 - \pi(x, g; \boldsymbol{\gamma}_i))} \frac{\partial \pi(x, g; \boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right| \\
&= \begin{cases} \left|\frac{1}{\pi(x,g;\boldsymbol{\gamma}_i)} \frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right|, & y = 1, \\ \left|\frac{1}{1-\pi(x,g;\boldsymbol{\gamma}_i)} \frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right|, & y = 0. \end{cases}
\end{aligned}$$

Analogously as in Appendix A.1, for simplicity, we will now consider only the reference group, i.e. $g = 0$:

$$\begin{aligned}
\frac{\partial \pi(x, g; \boldsymbol{\gamma}_i)}{\partial a_i} &= (d_i - c_i)\phi(x)(1 - \phi(x))(x - b_i), \\
\frac{\partial \pi(x, g; \boldsymbol{\gamma}_i)}{\partial b_i} &= (d_i - c_i)\phi(x)(1 - \phi(x))a_i, \\
\frac{\partial \pi(x, g; \boldsymbol{\gamma}_i)}{\partial c_i} &= 1 - \phi(x), \\
\frac{\partial \pi(x, g; \boldsymbol{\gamma}_i)}{\partial d_i} &= \phi(x).
\end{aligned}$$

Thus we need to show that $\frac{\phi(x)}{\pi(x,g=0;\boldsymbol{\gamma}_i)}$, $\frac{1-\phi(x)}{\pi(x,g=0;\boldsymbol{\gamma}_i)}$, $\frac{\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)}$, and $\frac{1-\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)}$ can be dominated by some integrable functions.

The function

$$\frac{\phi(x)}{\pi(x, g = 0; \boldsymbol{\gamma}_i)} = \frac{e^{a_i(x-b_i)}}{c_i + d_i e^{a_i(x-b_i)}} > 0$$

is increasing when $a_i > 0$ with the upper asymptote given by

$$\lim_{x\to\infty} \frac{e^{a_i(x-b_i)}}{c_i + d_i e^{a_i(x-b_i)}} = \frac{1}{d_i},$$

and thus

$$\frac{\phi(x)}{\pi(x, g = 0; \boldsymbol{\gamma}_i)} \leq K_1 \in \mathbb{R}.$$

Similarly, the function

$$\frac{1 - \phi(x)}{\pi(x, g = 0; \boldsymbol{\gamma}_i)} = \frac{1}{c_i + d_i e^{a_i(x-b_i)}} > 0$$

is decreasing when $a_i > 0$ with the upper asymptote given by

$$\lim_{x\to-\infty} \frac{1}{c_i + d_i e^{a_i(x-b_i)}} = \frac{1}{c_i},$$

and thus

$$\frac{1 - \phi(x)}{\pi(x, g = 0; \boldsymbol{\gamma}_i)} \leq K_2 \in \mathbb{R}.$$

Analogously, it can be shown that $\frac{\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)} > 0$ is an increasing function with an upper asymptote of $\frac{1}{1-d_i}$ and that $\frac{1-\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)} > 0$ is an decreasing function with an upper asymptote of $\frac{1}{1-c_i}$.

Further, as $\phi(x) \in (0,1)$, $1-\phi(x) \in (0,1)$, $d_i - c_i \in (0,1)$, and $|x-b_i| \leq x^2 + K_3$, for some $K_3 \in \mathbb{R}$, it is clear to see that

$$\left|\frac{\partial f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \gamma_{ik}}\right| \leq K_1^* + K_2^* x^2,$$

for $k = 1, \dots, 8$, where $K_1^*, K_2^* \in \mathbb{R}$ are sufficiently large, which is an integrable function.

$$\begin{aligned}
\left|\frac{\partial^2 f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right| &= \Bigg|\frac{y-\pi(x,g;\boldsymbol{\gamma}_i)}{\pi(x,g;\boldsymbol{\gamma}_i)\left(1-\pi(x,g;\boldsymbol{\gamma}_i)\right)}\Bigg[\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top} \\
&\qquad - \frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top \frac{y-\pi(x,g;\boldsymbol{\gamma}_i)}{\pi(x,g;\boldsymbol{\gamma}_i)\left(1-\pi(x,g;\boldsymbol{\gamma}_i)\right)}\Bigg]\Bigg| \\
&= \begin{cases}
\left|\frac{1}{\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top} - \frac{1}{\pi^2(x,g;\boldsymbol{\gamma}_i)}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top\right|, & y=1 \\
\Big|\frac{1}{(1-\pi(x,g;\boldsymbol{\gamma}_i))^2}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top & \\
\quad - \frac{1}{1-\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\Big|, & y=0
\end{cases} \\
&\leq \begin{cases}
\left|\frac{1}{\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right| + \left|\frac{1}{\pi^2(x,g;\boldsymbol{\gamma}_i)}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top\right|, & y=1, \\
\left|\frac{1}{1-\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right| & \\
\quad + \left|\frac{1}{(1-\pi(x,g;\boldsymbol{\gamma}_i))^2}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top\right|, & y=0.
\end{cases}
\end{aligned}$$

We have already shown that the elements of vectors $\left|\frac{1}{\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right|$ and $\left|\frac{1}{1-\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right|$ are all dominated by an integrable function $K_1^* + K_2^* x^2$ for K_1^*, $K_2^* \in \mathbb{R}$ sufficiently large. Therefore, also elements of the vectors outer products, i.e. matrices $\left|\frac{1}{\pi^2(x,g;\boldsymbol{\gamma}_i)}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top\right|$ and $\left|\frac{1}{(1-\pi(x,g;\boldsymbol{\gamma}_i))^2}\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\left(\frac{\partial \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i}\right)^\top\right|$, are all dominated by an integrable function $(K_1^* + K_2^* x^2)^2$.

Further, we have also shown that all terms $\frac{\phi(x)}{\pi(x,g=0;\boldsymbol{\gamma}_i)}$, $\frac{1-\phi(x)}{\pi(x,g=0;\boldsymbol{\gamma}_i)}$, $\frac{\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)}$, and $\frac{1-\phi(x)}{1-\pi(x,g=0;\boldsymbol{\gamma}_i)}$ can be bounded by a constant. Moreover, using also the fact that $\left|\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \partial \gamma_{il}}\right| \leq x^4 + K_3$, $k,l = 1,\dots,8$, already shown in Appendix A.1, it is clear that also all elements of $\left|\frac{1}{\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right|$ and $\left|\frac{1}{1-\pi(x,g;\boldsymbol{\gamma}_i)}\frac{\partial^2 \pi(x,g;\boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_i^\top}\right|$ are all bounded by and integrable function $x^4 + K_3^*$ for $K_3^* \in \mathbb{R}$ sufficiently large.

Finally we get

$$\left|\frac{\partial^2 f(y|x,g,\boldsymbol{\gamma}_i)}{\partial \gamma_{ik} \partial \gamma_{il}}\right| \leq K_4 x^4 + K_5 x^2 + K_6,$$

for $k,l = 1,\dots,8$, where $K_4, K_5, K_6 \in \mathbb{R}$ are sufficiently large which completes the proof that the condition [R4*] holds.

A.3 R script for the EM algorithm

```
# data generation
set.seed(42)

x <- rnorm(1000)
g <- rep(c(0, 1), each = 500)
p <- 0.2 + 0.1 * g + (1 - 0.1 * g - 0.2 - 0.1 * g) /
  (1 + exp(0 - x + 1 * g - 0.5 * x * g))
y <- rbinom(1000, 1, p)

# initial values
b0_new <- 0.1
b1_new <- 0.85
b2_new <- -1.1
b3_new <- 0.6
c_new <- 0.15
cDif_new <- 0.15
d_new <- 0.95
dDif_new <- -0.05

par <- list()
k <- 1
dev_new <- 0

# EM algorithm
repeat({
  par[[k]] <- c(
    b0_new, b1_new, b2_new, b3_new,
    c_new, cDif_new, d_new, dDif_new
  )
  # checking maximal number of iterations
  # actual number of iterations is k - 1,
  # the first is for the initial run
  if (k == 2001) {
    break
  }

  # E-step
  Z <- expectation(
    y, x, g,
    b0_new, b1_new, b2_new, b3_new,
    c_new, cDif_new, d_new, dDif_new
  )
  # M-step
  fit1 <- glm(cbind(Z$z2, Z$z3) ~ x + g + x:g,
    family = binomial(),
    start = c(b0_new, b1_new, b2_new, b3_new)
  )

  b0_old <- b0_new
  b1_old <- b1_new
  b2_old <- b2_new
  b3_old <- b3_new
  b0_new <- coef(fit1)[1]
  b1_new <- coef(fit1)[2]
  b2_new <- coef(fit1)[3]
  b3_new <- coef(fit1)[4]

  fit2 <- multinom(cbind(Z$z2 + Z$z3, Z$z1, Z$z4) ~ g, trace = FALSE)
  par_asympt <- as.data.frame(unique(cbind(g, fitted(fit2))))
  # calculating upper asymptotes for the two groups
  par_asympt$V4 <- 1 - par_asympt$V4
  # differences in parameters between focal and reference group
  par_asympt[3, ] <- par_asympt[par_asympt$g == 1, ] -
    par_asympt[par_asympt$g == 0, ]
  pars <- par_asympt[c(1, 3), c(1, 3, 4)]

  c_old <- c_new
  c_new <- pars[pars$g == 0, "V3"]
  cDif_old <- cDif_new
  cDif_new <- pars[pars$g == 1, "V3"]
  d_old <- d_new
  d_new <- pars[pars$g == 0, "V4"]
  dDif_old <- dDif_new
  dDif_new <- pars[pars$g == 1, "V4"]

  # deviance
  dev_old <- dev_new
  dev_new <- deviance(fit1) + deviance(fit2)

  # checking stopping criterion
  if (abs(dev_old - dev_new) / (0.1 + dev_new) < 1e-6) {
    k <- k + 1
    par[[k]] <- c(
      b0_new, b1_new, b2_new, b3_new,
      c_new, cDif_new, d_new, dDif_new
    )
    break
  }
  k <- k + 1
})
par <- as.data.frame(do.call(rbind, par))
colnames(par) <- c("b0", "b1", "b2", "b3", "c", "cDif", "d", "dDif")

# final parameter estimates
par[nrow(par), ]
      b0      b1       b2      b3       c     cDif       d     dDif
-0.17602 0.98603 -0.87166 0.85062 0.21989 0.096286 0.99945 -0.14934
# standard errors of the estimates
sqrt(diag(covariance.matrix(x, y, g, par[nrow(par), ])))
0.79370 0.44492 1.00681 0.91258 0.20527 0.21341 0.16233 0.19429
```

A.4 R script for the algorithm based on parametric link function

```
# data generation
set.seed(42)

x <- rnorm(1000)
g <- rep(c(0, 1), each = 500)
p <- 0.2 + 0.1 * g + (1 - 0.1 * g - 0.2 - 0.1 * g) /
  (1 + exp(0 - x + 1 * g - 0.5 * x* g))
y <- rbinom(1000, 1, p)

# model matrix
X <- cbind(1, x, g, x * g)

# initial values
b0_new <- 0.1
b1_new <- 0.85
b2_new <- -1.1
b3_new <- 0.6
c_new <- 0.15
cDif_new <- 0.15
d_new <- 0.95
dDif_new <- -0.05

k <- 1
par <- list()
ll_new <- 0

# Algorithm based on parametric link function
repeat({
  par[[k]] <- c(
    b0_new, b1_new, b2_new, b3_new,
    c_new, cDif_new, d_new, dDif_new
  )
  # checking maximal number of iterations
  # actual number of iterations is k - 1,
  # the first is for the initial run
  if (k == 2001) {
    break
  }

  # Step 1: fitting GLM with parametric link function
  fit_glm <- glm(y ~ x + g + x:g,
    family = binomial(
      link = plogit(c_new, cDif_new, d_new, dDif_new, g)
    ),
    start = c(b0_new, b1_new, b2_new, b3_new)
  )

  b0_old <- b0_new
```

```
b1_old <- b1_new
b2_old <- b2_new
b3_old <- b3_new

b0_new <- coef(fit_glm)[1]
b1_new <- coef(fit_glm)[2]
b2_new <- coef(fit_glm)[3]
b3_new <- coef(fit_glm)[4]

# bound for asymptotes
c0_max <- max(min(fitted(fit_glm)[g == 0], na.rm = TRUE), 0)
c1_max <- max(min(fitted(fit_glm)[g == 1], na.rm = TRUE), 0)
d0_min <- min(max(fitted(fit_glm)[g == 0], na.rm = TRUE), 1)
d1_min <- min(max(fitted(fit_glm)[g == 1], na.rm = TRUE), 1)

# Step 2: estimating asymptotes parameters
fit_cd <- optim(
  fn = param.likel.cd,
  par = setNames(
    c(
      (c_new + c0_max) / 2,
      (c_new + cDif_new + c1_max) / 2,
      (d0_min + d_new) / 2,
      (d_new + dDif_new + d1_min) / 2
    ),
    c("c0", "c1", "d0", "d1")
  ),
  method = "L-BFGS-B",
  lower = c(0, 0, d0_min, d1_min),
  upper = c(c0_max, c1_max, 1, 1)
)

c_old <- c_new
cDif_old <- cDif_new
d_old <- d_new
dDif_old <- dDif_new

c_new <- fit_cd$par[1]
cDif_new <- fit_cd$par[2] - fit_cd$par[1]
d_new <- fit_cd$par[3]
dDif_new <- fit_cd$par[4] - fit_cd$par[3]

# log-likelihood
ll_old <- ll_new
ll_new <- logLik(fit_glm) - fit_cd$value

par[[k]] <- c(
  b0_new, b1_new, b2_new, b3_new,
  c_new, cDif_new, d_new, dDif_new
)
# checking stopping criterion
```

```
  if (abs(abs(ll_old - ll_new) / (0.1 + ll_new)) < 1e-6) {
    break
  }
  k <- k + 1
})
par <- as.data.frame(do.call(rbind, par))
colnames(par) <- c("b0", "b1", "b2", "b3", "c", "cDif", "d", "dDif")

# final parameter estimates
par[nrow(par), ]
      b0      b1       b2      b3       c     cDif       d     dDif
-0.15238 0.97252 -0.89734 0.86781 0.21307 0.10343 1.00000 -0.15020

# standard errors of the estimates
sqrt(diag(covariance.matrix(x, y, g, par[nrow(par), ])))
0.78076 0.43197 0.99690 0.90788 0.20486 0.21298 0.16488 0.19642
```

A.5 R script for the calculation of starting values based on CTT

```
startCTT <- function(x, y, num.groups = 3) {
  # split matching criterion x to num.groups
  breaks <- unique(quantile(x, (0:num.groups) / num.groups,
    na.rm = TRUE
  ))
  groups <- cut(x, breaks, include.lowest = TRUE)
  levels(groups) <- LETTERS[1:num.groups]

  # c is average y (empirical probability) for those whose matching
  # criterion x is smaller than average value of x in the first group
  # accounting for the variability of x
  c <- mean(y[x < (mean(x[groups == LETTERS[1]],
    na.rm = TRUE
  ) - sd(x) / 2)])
  # d is average y (empirical probability) for those whose matching
  # criterion x is greater than average value of x in the last group
  # accounting for the variability of x
  d <- mean(y[x > (mean(x[groups == LETTERS[num.groups]],
    na.rm = TRUE
  ) + sd(x) / 2)])

  # ULI index = difference in empirical probabilities in the first and
  # the last group
  uli <- mean(y[groups == LETTERS[num.groups]], na.rm = TRUE) -
    mean(y[groups == LETTERS[1]], na.rm = TRUE)
  # slope
  b1 <- 4 * uli

  # center point between asymptotes, empirical probability Y._i
  dotY <- (d + c) / 2

  tmp <- c()
  # rounded x and corresponding empirical probabilities
  sorted_x <- unique(round(sort(x), 1))
  for (i in sorted_x) {
    tmp <- c(tmp, mean(y[round(x, 1) == i]))
  }

  # b0 (resp. difficulty b) is a value of matching criterion x, which
  # gives probability of dotY
  # looking for the smallest absolute distance of tmp from dotY
  tmp <- abs(tmp - dotY)
  # weighting absolute distances
  w <- prop.table(table(round(x, 1)))
  tmp <- tmp / w

  # smoothing accounting for neighbors
  tmp <- tmp +
    0.1 * c(tmp[-1], tmp[length(tmp)]) +
    0.1 * c(tmp[1], tmp[-length(tmp)])

  # Looking for the smallest distance
  min_sorted_x <- sorted_x[which(tmp == min(tmp))]
  # creating of neighborhood of the point
  min_msm <- min(min_sorted_x) - sd(x) / 2
  max_msm <- max(min_sorted_x) + sd(x) / 2

  # b0 = -b1 * b, where b is difficulty
  b0 <- -b1 * mean(x[x > min_msm & x < max_msm])

  results <- as.data.frame(cbind(b0, b1, c, d))
  return(results)
}
```

A.6 R script for the calculation of starting values based on grid search

```
startGRID <- function(x, y, num.groups = 3) {
  # initial values based on CTT
  init_ctt <- startCTT(x, y, num.groups)

  X <- cbind(1, x)
  # parametric expit
  param.expit <- function(x, c, d) {
    c + (d - c) / (1 + exp(-x))
  }
  # log-likelihood calculation
  log.likel <- function(theta) {
    n <- nrow(X)
    c <- theta[1]
    d <- theta[2]
    b0 <- theta[3]
    b1 <- theta[4]

    h <- param.expit(X %*% c(b0, b1), c, d)
    l <- sum((y * log(h)) + ((1 - y) * log(1 - h)))
    return(l)
  }

  # creating grid
  # values for c, accounting for variance of c
  c_seq <- seq(max(0, init_ctt$c - init_ctt$c * (1 - init_ctt$c)),
    min(0.5, init_ctt$c + init_ctt$c * (1 - init_ctt$c)),
    length.out = 10
  )
  # values for d, accounting for variance of d
  d_seq <- seq(max(0.5, init_ctt$d - init_ctt$d * (1 - init_ctt$d)),
    min(1, init_ctt$d + init_ctt$d * (1 - init_ctt$d)),
    length.out = 10
  )
  # values for b0, accounting for variability of x
  b0_seq <- seq(init_ctt$b0 - sd(x), init_ctt$b0 + sd(x),
    length.out = 10
  )
  # values for b1, accounting for variability of y
  b1_seq <- seq(init_ctt$b1 * sd(y), init_ctt$b1 / sd(y),
    length.out = 10
  )

  grid <- expand.grid(c_seq, d_seq, b0_seq, b1_seq)
  colnames(grid) <- c("c", "d", "b0", "b1")
  grid$loglik <- apply(grid, 1, log.likel)

  # looking for maximum value of log-likelihood and final parameters
  maxll <- which(grid$loglik == max(grid$loglik))
  results <- unlist(
    unique(grid[maxll, c("b0", "b1", "c", "d")])
  )
  return(results)
}
```

A.7 Tables

Table A.1: Arguments of the `difNLR()` function.

Argument	Description
`Data`	`data.frame` or `matrix`: dataset which rows represent scored examinee answers (`"1"` correct, `"0"` incorrect) and columns correspond to the items. In addition, `Data` can hold the vector of group membership.
`group`	`numeric` or `character`: a dichotomous vector of the same length as `nrow(Data)` or a column identifier of `Data`.
`focal.name`	`numeric` or `character`: indicates the level of `group` which corresponds to the focal group.
`model`	`character`: generalized logistic regression model to be fitted. See Table 1.1.
`constraints`	`character`: which parameters should be the same for both groups. Possible values are any combinations of parameters `"a"`, `"b"`, `"c"`, and `"d"`.
`type`	`character`: type of DIF to be tested. Possible values are `"all"` for detecting difference in any parameter (default), `"udif"` for uniform DIF only (i.e., difference in difficulty parameter `"b"`), `"nudif"` for non-uniform DIF only (i.e., difference in discrimination parameter `"a"`), `"both"` for uniform and non-uniform DIF (i.e., difference in parameters `"a"` and `"b"`), or combination of parameters `"a"`, `"b"`, `"c"`, and `"d"`. Can be specified as a single value (for all items) or as an item-specific vector.
`method`	`character`: method used to estimate parameters. Either `"nls"` for non-linear least squares (default), or `"likelihood"` for maximum likelihood method.
`match`	`numeric` or `character`: matching criterion to be used as an estimate of trait. Can be either `"zscore"` (default, standardized total score), `"score"` (total test score), or vector of the same length as number of observations in `Data`.
`anchor`	numeric or character: specification of DIF free items. Either `NULL` (default), or a vector of item names (column names of `Data`), or item identifiers (integers specifying the column number) determining which items are currently considered as anchor (DIF free) items. Argument is ignored if match is not `"zscore"` or `"score"`.
`purify`	`logical`: should the item purification be applied? (default is `FALSE`).
`nrIter`	`numeric`: the maximal number of iterations in the item purification (default is 10).
`test`	`character`: test to be performed for DIF detection. Can be either `"LR"` for likelihood ratio test of a submodel (default), `"W"` for Wald test, or `"F"` for F-test of a submodel.

Argument	Description
`alpha`	`numeric`: significance level (default is 0.05).
`p.adjust.method`	`character`: method for multiple comparison correction. Possible values are `"holm"`, `"hochberg"`, `"hommel"`, `"bonferroni"`, `"BH"`, `"BY"`, `"fdr"`, and `"none"` (default).
`start`	`numeric`: initial values for estimation of parameters. If not specified, starting values are calculated with the `startNLR()` function. Otherwise, list with as many elements as a number of items. Each element is a named numeric vector of length 8 representing initial values for parameter estimation. Specifically, parameters `"a"`, `"b"`, `"c"`, and `"d"` are initial values for discrimination, difficulty, guessing, and inattention for the reference group. Parameters `"aDif"`, `"bDif"`, `"cDif"`, and `"dDif"` are then differences in these parameters between the reference and focal group.
`initboot`	`logical`: in case of convergence issues, should be starting values re-calculated based on bootstraped samples? (default is `TRUE`; newly calculated initial values are applied only to items/models with convergence issues).
`nrBo`	numeric: the maximal number of iterations for calculation of starting values using bootstraped samples (default is 20).
`sandwich`	`logical`: should be sandwich estimator used for covariance matrix of parameters when using method = `"nls"`? Default is `FALSE`.

Table A.2: Arguments of the `difORD()` function.

Argument	Description
`Data`	`data.frame` or `matrix`: dataset which rows represent ordinaly scored examinee answers and columns correspond to the items. In addition, `Data` can hold the vector of group membership.
`group`	`numeric` or `character`: a dichotomous vector of the same length as `nrow(Data)` or a column identifier of `Data`.
`focal.name`	`numeric` or `character`: indicates the level of group which corresponds to the focal group.
`model`	`character`: logistic regression model for ordinal data (either `"adjacent"` (default) or `"cumulative"`).
`type`	`character`: type of DIF to be tested. Either `"both"` for uniform and non-uniform DIF (i.e., difference in parameters `"a"` and `"b"`) (default), or `"udif"` for uniform DIF only (i.e., difference in difficulty parameter `"b"`), or `"nudif"` for non-uniform DIF only (i.e., difference in discrimination parameter `"a"`). Can be specified as a single value (for all items) or as an item-specific vector.
`match`	`numeric` or `character`: matching criterion to be used as an estimate of trait. Can be either `"zscore"` (default, standardized total score), `"score"` (total test score), or vector of the same length as number of observations in `Data`.
`anchor`	`numeric` or `character`: specification of DIF free items. Either `NULL` (default), or a vector of item names (column names of `Data`), or item identifiers (integers specifying the column number) determining which items are currently considered as anchor (DIF free) items. Argument is ignored if match is not `"zscore"` or `"score"`.
`purify`	`logical`: should the item purification be applied? (default is `FALSE`).
`nrIter`	`numeric`: the maximal number of iterations in the item purification (default is 10).
`p.adjust.method`	character: method for multiple comparison correction. Possible values are `"holm"`, `"hochberg"`, `"hommel"`, `"bonferroni"`, `"BH"`, `"BY"`, `"fdr"`, and `"none"` (default).
`parametrization`	`character`: parametrization of regression coefficients. Possible options are `"irt"` for difficulty-discrimination parametrization (default) and `"classic"` for intercept-slope parametrization.
`alpha`	`numeric`: significance level (default is 0.05).

Table A.3: Arguments of the ddfMLR() function.

Argument	Description
Data	data.frame or matrix: dataset which rows represent unscored examinee answers (nominal) and columns correspond to the items. In addition, Data can hold the vector of group membership.
group	numeric or character: a dichotomous vector of the same length as nrow(Data) or a column identifier of Data.
focal.name	numeric or character: indicates the level of group which corresponds to the focal group.
key	character: the answer key. Each element corresponds to the correct answer of one item.
type	character: type of DIF to be tested. Either "both" for uniform and non-uniform DIF (i.e., difference in parameters "a" and "b") (default), or "udif" for uniform DIF only (i.e., difference in difficulty parameter "b"), or "nudif" for non-uniform DIF only (i.e., difference in discrimination parameter "a"). Can be specified as a single value (for all items) or as an item-specific vector.
match	numeric or character: matching criterion to be used as an estimate of trait. Can be either "zscore" (default, standardized total score), "score" (total test score), or vector of the same length as number of observations in Data.
anchor	numeric or character: specification of DIF free items. Either NULL (default), or a vector of item names (column names of Data), or item identifiers (integers specifying the column number) determining which items are currently considered as anchor (DIF free) items. Argument is ignored if match is not "zscore" or "score".
purify	logical: should the item purification be applied? (default is FALSE).
nrIter	numeric: the maximal number of iterations in the item purification (default is 10).
p.adjust.method	character: method for multiple comparison correction. Possible values are "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", and "none" (default).
parametrization	character: parametrization of regression coefficients. Possible options are "irt" for difficulty-discrimination parametrization (default) and "classic" for intercept-slope parametrization.
alpha	numeric: significance level (default is 0.05).

Table A.4: Mean and median parameter estimates with the bias, model based standard errors, and empirical standard deviations for $n = 1,000$.

	NLS	MLE	EM	PLF		NLS	MLE	EM	PLF
b_0					c				
Count	969	1000	1000	981	Count	969	1000	1000	981
Mean	0.047	−3.213	0.002	−0.002	Mean	0.218	0.222	0.226	0.223
Median	0.108	0.032	0.033	0.027	Median	0.239	0.245	0.248	0.248
Bias	0.047	−3.213	0.002	−0.002	Bias	0.018	0.022	0.026	0.023
MBSE	0.724	1.569	0.652	0.624	MBSE	0.180	0.205	0.190	0.197
ESD	1.070	98.688	0.841	0.641	ESD	0.140	0.140	0.134	0.133
b_1					c_{DIF}				
Count	969	1000	1000	981	Count	969	1000	1000	981
Mean	1.636	11.065	1.497	1.422	Mean	0.072	0.071	0.068	0.063
Median	1.257	1.240	1.256	1.243	Median	0.066	0.065	0.065	0.055
Bias	0.636	10.065	0.497	0.422	Bias	−0.028	−0.029	−0.032	−0.037
MBSE	1.134	3.525	0.871	0.808	MBSE	0.218	0.244	0.225	0.235
ESD	2.331	291.513	1.312	1.134	ESD	0.171	0.172	0.164	0.164
b_2					d				
Count	969	1000	1000	981	Count	969	1000	1000	981
Mean	−3.440	−47.063	−107.484	−0.696	Mean	0.939	0.949	0.947	0.949
Median	−1.216	−1.204	−1.200	−1.133	Median	0.963	0.991	0.978	0.983
Bias	−2.440	−46.063	−106.484	0.304	Bias	−0.061	−0.051	−0.053	−0.051
MBSE	1.420	30.865	1.288	1.539	MBSE	0.113	0.146	0.140	0.145
ESD	53.023	1823.992	2338.516	10.733	ESD	0.068	0.065	0.063	0.062
b_3					d_{DIF}				
Count	969	1000	1000	981	Count	969	1000	1000	981
Mean	4.559	228.338	153.963	0.271	Mean	−0.041	−0.045	−0.049	−0.042
Median	0.439	0.457	0.454	0.416	Median	−0.019	−0.004	−0.024	−0.013
Bias	4.059	227.838	153.463	−0.229	Bias	0.059	0.055	0.051	0.058
MBSE	2.376	60.206	1.933	1.909	MBSE	0.208	0.255	0.240	0.255
ESD	89.025	4183.318	2826.562	6.610	ESD	0.118	0.118	0.115	0.112

Note. NLS = nonlinear least squares, MLE = maximum likelihood estimation, EM = expectation-maximization algorithm, PLF = method based on parametric link function, Count = number of parameter estimates excluding crashed simulation runs, MBSE = model based standard error, ESD = empirical standard deviation.