Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing

Drabinová, Adéla
2016

Institute of Computer Science
The Czech Academy of Sciences

Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing

Adéla Drabinová, Patrícia Martinková

Technical report No. V1229

September 2016

Institute of Computer Science
The Czech Academy of Sciences

Detection of Differential Item Functioning with Non-Linear Regression: Non-IRT Approach Accounting for Guessing

Adéla Drabinová[1][2], Patrícia Martinková[3]

Technical report No. V1229

September 2016

Abstract:

[1] Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic

[2] Faculty of Mathematics and Physics, Charles University in Prague, Ke Karlovu 3, 121 16, Prague 2, Czech Republic, E-mail: adela.drabinova@gmail.com

[3] Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic, E-mail: martinkova@cs.cas.cz

In this article, we present an extension of two-parameter Logistic Regression (LR) to detect both uniform and non-uniform DIF in dichotomous items. Contrary to other non Item Response Theory (non-IRT) methods, the newly proposed non-linear regression (NLR) procedure accounts for guessing. As a non-IRT approach, NLR procedure can be seen as proxy of three-parameter (3PL) IRT model for DIF detection which is a standard tool in study field. Hence NLR fills the logical gap in DIF detection methodology and as such is important for educational purposes. The advantages of the NLR procedure as well as comparison to other commonly used methods are demonstrated in a simulation study based on Graduate Management Admission Test in which we show pleasant properties of proposed approach including low convergence failure rate and in most cases sufficient power and low rejection rate. Besides simulation study, also a real data analysis is offered to demonstrate practical use of NLR method. The newly proposed NLR method is accompanied by an R package and is implemented in an online Shiny application.

Keywords:
differential item functioning; non-linear regression; logistic regression; item response theory; Mantel-Haenszel procedure

Introduction

Detection of Differential Item Functioning (*DIF*) has been considered one of the most important topics in measurement (AERA, APA, & NCME, 2014). An item is said to function differently when subjects from different groups but with the same level of knowledge (or other latent trait) have different probabilities of answering the item correctly. In such a case, some aspect of the item, unrelated to the tested knowledge, could be unfairly causing the difference, thus DIF items are potentially unfair. A variety of methods for DIF detection has been proposed and they are still being studied intensively (Berger & Tutz, 2016; S.-H. Kim, Cohen, Alagoz, & Kim, 2007; J. Kim & Oshima, 2013; Loken & Rulison, 2010; Magis & De Boeck, 2011; Magis, 2013; Magis, Tuerlinckx, & De Boeck, 2014). Generally, DIF detection approaches can be divided into Item Response Theory (*IRT*) model based procedures and techniques based on test score (here referenced as *non-IRT*).

Procedures based on Mantel-Haenszel (*MH*) test (Mantel & Haenszel, 1959; Holland, 1985; Holland & Thayer, 1988) and two-parameter logistic regression (*LR*) (Swaminathan & Rogers, 1990) are some of the most widely used non-IRT methods in identifying DIF items. Both MH and LR are straightforward methods which are easy to explain to audience, easy to apply in standard statistical packages and also have been used in DIF detection for more than two decades. Both procedures have been extensively discussed in literature by many authors, including Choi, Gibbons, and Crane (2011); French and Maller (2007); Hidalgo and López-Pina (2004); Holland and Wainer (2012); Penfield (2001) and Zumbo (1999). However, neither of these methods accounts for possibility of guessing.

Alternatively, within IRT model framework, the three-parameter (*3PL*) IRT model accounting for guessing (Birnbaum, 1968) is widely used for item calibration and also for DIF detection. Two popular statistics for testing DIF within IRT models are Lord's χ^2-test which compares item parameter estimates (Lord, 1980) and Raju's statistics based on area between item characteristic curves (Raju, 1988, 1990). However, the underlying non-linear 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. Moreover, 3PL IRT model is difficult to fit and large sample size (500 per group) is required (J. Kim & Oshima, 2013).

In this work, we present a new general non-IRT approach to detect DIF in dichotomous items with presence of guessing. We use non-linear regression model - an extension of LR model for DIF detection, which accounts for the probability of guessing the correct answer. To our best knowledge, the possibility of extension of LR to account for guessing has yet not been explored in detection of DIF and thus the proposed method fills the logical gap in DIF detection methodology. As such, the non-IRT method proposed in this paper is important for educational purposes and to increase the use of the DIF methodology.

To investigate properties of proposed technique, method is compared to MH, LR and IRT model based approaches in simulation study based on items parameters from Graduate Management Admission Test (GMAT) data set (Kingston, Leary, & Wightman, 1985) and also in illustrative analysis of admission test to medical school (Vlčková, 2014).

The paper proceeds as follows: Methodology for DIF detection with NLR, simulation study design and practical implementation within new difNLR R software package (Drabinová, Martinková, & Zvára, 2016) are described in section *Methodology.* Results of simulation study and real data set analysis are performed in section *Results.* Discussion and conclusion is offered in the last section.

Methodology

Non-Linear Regression DIF Detection Procedure

To provide more precise analysis of item properties but stay within non-IRT framework, we propose an extension of LR procedure by including the probability of guessing c. For simplicity and to correspond with related IRT-based approaches (Raju, 1988, 1990), we assume that probability of guessing is the same for both groups in one item, however, we allow guessing parameter to vary across items, e.g. to account for the fact that in multiple-choice tests the number of possible answers or attractiveness of distractors can differ across items. Using logistic parameterization, the probability of correct answer to an item i by j-th examinee is given by equation

$$\mathrm{P}(Y_{ij}=1|X_j,G_j)=c_i+(1-c_i)\frac{e^{\beta_{0i}+\beta_{1i}X_j+\beta_{2i}G_j+\beta_{3i}X_jG_j}}{1+e^{\beta_{0i}+\beta_{1i}X_j+\beta_{2i}G_j+\beta_{3i}X_jG_j}}, \tag{1}$$

where Y_{ij} is response of j-th examinee to an item i (1 for correct, 0 for incorrect), X_j stands for his/her observed knowledge (the standardized total test score) and variable G_j represents his/her group membership (1 for reference group, 0 for focal group). For case without guessing (i.e. $c_i = 0$), the model is formally equivalent to model proposed by Swaminathan and Rogers (1990). The interpretation of regression coefficients is the same as for LR (Agresti & Kateri, 2003, Chapter 5).

To allow later comparison of items parameters extimates with IRT-based approaches, model (1) can be reparameterized as follows:

$$\mathrm{P}(Y_{ij}=1|X_j,G_j)=c_i+(1-c_i)\frac{e^{(a_i+a_{\mathrm{DIF}i}G_j)(X_j-(b_i+b_{\mathrm{DIF}i}G_j))}}{1+e^{(a_i+a_{\mathrm{DIF}i}G_j)(X_j-(b_i+b_{\mathrm{DIF}i}G_j))}}. \tag{2}$$

In this *NLR* model 2, the variables Y_{ij}, X_j and G_j are as above. Regression parameter a_i is now the discrimination parameter of the i-th item and b_i is the difficulty parameter of the i-th item[4]. Parameter $a_{\mathrm{DIF}i}$, respectively $b_{\mathrm{DIF}i}$, represents the difference in discrimination, respectively in difficulty, of reference and focal group.[5] The interpretation of guessing parameter c_i is the same for both parameterizations: it is probability that correct answer of item i is guessed without necessary knowledge. NLR model is proxy for 3PL IRT model for DIF detection (Raju, 1990) as it is by definition score-based method and thus non-IRT.

Extension of logistic regression model allowing lower asymptote to differ from zero is in literature known as three-parameter logistic model (Glas & Falcón, 2003). However, to our best knowledge, the application in DIF identification has not yet been studied.

[4] The relationship between regression coefficients of parameterization (1) and (2) is as follows: intercept β_{0i} is equal to $-a_ib_i$ and effect of total score β_{1i} is equal to a_i

[5] The effect of group membership β_{2i} in parameterization (1) is equal to $-a_ib_{\mathrm{DIF}i} - a_{\mathrm{DIF}i}b_i - a_{\mathrm{DIF}i}b_{\mathrm{DIF}i}$ in parameterization (2). The effect of interaction of total score and group membership β_{3i} in parameterization (1) is equal to $a_{\mathrm{DIF}i}$ in parameterization (2).

While models including guessing parameter c are often called 3PL (Glas & Falcón, 2003; Raju, 1990), it should be noted that the definition of logistic model or generalized linear model (Agresti & Kateri, 2003, Chapters 4 and 5) does not hold in models 1 and 2 and thus we will henceforward call model given by equation (2) NLR model. In what follows, we stick with NLR model to be able to compare parameter estimates to those obtained by IRT model. For logistic parameterization (1), parameter estimation procedures and DIF detection methods would be analogous.

Estimation and DIF Detection

The parameter estimates of model are determined by non-linear least square estimation, that is by minimization of the residual sums of squares (RSS) with respect to $(a_i, b_i, a_{\text{DIF}i}, b_{\text{DIF}i}, c_i)$:

$$\text{RSS}(a_i, b_i, a_{\text{DIF}i}, b_{\text{DIF}i}, c_i) = \sum_{j=1}^{n} \left[y_{ij} - c_i + (1 - c_i) \frac{e^{(a_i + a_{\text{DIF}i} g_j)(x_j - (b_i + b_{\text{DIF}i} g_j))}}{1 + e^{(a_i + a_{\text{DIF}i} g_j)(x_j - (b_i + b_{\text{DIF}i} g_j))}} \right]^2 ,$$

where n is number of examinees, y_{ij} is response of j-th examinee to item i, x_j is his/her standardized total score and g_j his/her group membership. Since the minimization in our case is nonlinear problem, a numerical optimization methods need to be applied.

The NLR model (2) can be utilized to detect DIF in a simple way. If value of $a_{\text{DIF}i}$ is zero and value of $b_{\text{DIF}i}$ differs from zero, this suggests presence of uniform DIF. If value of $a_{\text{DIF}i}$ differs from zero, this suggests presence of non-uniform DIF. In short, possible DIF scenarios for item i are characterized by the following null and alternative hypotheses:

Any DIF	$\text{H}_0 : a_{\text{DIF}i} = 0 \ \& \ b_{\text{DIF}i} = 0$	$\text{H}_1 : a_{\text{DIF}i} \neq 0$ or $b_{\text{DIF}i} \neq 0$
Uniform DIF	$\text{H}_0 : b_{\text{DIF}i} = 0 \mid a_{\text{DIF}i} = 0$	$\text{H}_1 : b_{\text{DIF}i} \neq 0 \mid a_{\text{DIF}i} = 0$
Non-uniform DIF	$\text{H}_0 : a_{\text{DIF}i} = 0$	$\text{H}_1 : a_{\text{DIF}i} \neq 0$

To compare two nested NLR models (where one model is defined by alternative hypothesis

H_1 and its submodel by null hypothesis H_0), and thus test for DIF presence in item i, the F-test or likelihood ratio test can be used with similar results (Dennis, Gay, & Welsch, 1981; Ritz & Streibig, 2008). In simulation study and data analysis, we stick with F-test for simplicity.

The F-test statistic measures the distance between model M_0 and its submodel M_1 as the difference between RSS relative to RSS of model M_0. The formula is the same as for linear models

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(df_0 - df_1)}{\mathrm{RSS}_0/df_0},$$

however in non-linear models the F-distribution holds only approximately (Ritz & Streibig, 2008). As the test is performed for each item separately, Benjamini-Hochberg adjustment procedure for multiple testing is applied (Benjamini & Hochberg, 1995), as suggested by J. Kim and Oshima (2013).

Simulation Study

Simulation study is performed to investigate properties of newly proposed NLR procedure and to make comparison to other commonly used DIF detection approaches including MH, LR, Lord's and Raju's methods. The simulation study focuses on convergence behavior, power rate (i.e. proportion of true positives) and rejection rate (i.e. proportion of false positives; type I error).

The dichotomously scored data are generated with a 3PL IRT model as follows: examinees' knowledge is assumed to follow the standard normal distribution. 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 difficulty or discrimination parameter of the focal group is manipulated (see below). To reflect realistic values of items parameters and to be in line with previous simulation studies (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1996; Jodoin & Gierl, 2001; Güler & Penfield, 2009; J. Kim & Oshima, 2013), simulation study is based on item parameters according to 20-item data set from the 1985 problem solving of the GMAT (Kingston et al., 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 be uniform, or non-uniform DIF items. The thresholds for DIF effect size of DIF items, represented by area between characteristic curves, are determined by values 0.4 (low), 0.6 (moderate) and 0.8 (large) following Swaminathan and Rogers (1990) and Narayanan and Swaminathan (1996). When one DIF item is considered, the large size of DIF is chosen. Mixture of DIF sizes is considered for larger proportion of DIF items. The parameters of remaining (19 or 17) items are selected from the problem solving 1985 of the GMAT as reported in Kingston et al. (1985), see Table 2. For all DIF items guessing parameter c is set to 0.2. When uniform DIF is considered, the discrimination parameters for focal and reference group are kept the same and fixed at value 1. The differences in difficulty between reference and focal group are set to 0.5 (low), 0.75 (moderate) and 1 (large). 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. To evaluate rejection rates of procedures also simulations without DIF items are considered.

The above described scenarios are investigated on various levels of the total sample size. Larger sizes of samples are determined to yield satisfactory convergence levels especially for IRT models. Specifically, three levels of sample size are considered: 1,000 (500 per group), 2,000 (1,000 per group), and 5,000 (2,500 per group).

Due to numerical estimation procedures in NLR and IRT-based methods, convergence issues can be observed. It should be noted that large proportions of convergence failures can 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 convergence issue are excluded and the proportion of these events is scored. Convergence failure rate is calculated as ratio of items with 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. As suggested by (J. Kim & Oshima, 2013), Benjamini-Hochberg multiple comparison correction is applied to all methods.

Real Data Analysis

To demonstrate practical use of the newly proposed NLR method and how it compares with commonly used DIF detection approaches including MH, LR, Lord's and Raju's procedures, we offer an illustrative analysis of medical school admission test (AT).

Original AT data set (Vlčková, 2014) consists of responses of 1,407 students (484 males and 923 females) to 80 items. All items have 4 possible answers and in some items there is more than one correct choice. For these items all correct answers and no incorrect answers must have been marked for the item to be recognized as correct. Previous study detected 6 items to be favoring males and 4 favoring females (Vlčková, 2014).

In this study we examine a subset of AT data set including one item (item 49) shown in the past to display DIF and 19 randomly selected items which were not previously detected as DIF (items 1, 2, 7, 9, 10, 17, 24, 25, 27, 28, 38, 41, 45, 47, 61, 64, 68, 75, 76). Item 49 is related to childhood disease, concretely what disease can be caused by deficiency of vitamin D in childhood. Possible answers were A. rickets (correct), B. scurvy, C. dwarfism and D. intellectual disability. An explanation of better performance of females, can be that women tend to be more experienced in looking after children and to know more about childhood diseases. While the item was detected to function differently for males and females, the item was not considered to be unfair, because the knowledge needed for correct answer is related to underlying latent concept being tested.

Practical Implementation

For all analyses, software R, Version 3.22 is used (R Core Team, 2015). The NLR procedure is implemented in new function difNLR within R package difNLR (Drabinová et al., 2016) which uses nls function from stats package for non-linear least squares estimation with constraints on guessing parameter (Dennis et al., 1981; Ritz & Streibig, 2008). To specify suitable initial values, we consider approach based on linear approximation. Mean values of standardized total score of first and third tertiles are spaced by line $\tilde{p}(x) = kx + q$, where x stands for standardized total score. Guessing parameter c stands for asymptotic minimum $p(-\infty)$ but taking into account linear approximation $\tilde{p}$, this value would be $-\infty$.[6] Initial value of 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. Guessing parameter influences 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 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, F-test is used.

The LR procedure is implemented by function glm from stats package (R Core Team, 2015). To detect DIF, likelihood ratio test is performed (Agresti & Kateri, 2003). R package difR (Magis, Béland, Tuerlinckx, & De Boeck, 2010) is used to perform MH test by function difMH and IRT-based DIF detection methods as follows: the 3PL model for all data is fitted with function itemParEst and guessing parameters are estimated. Then 3PL models for both groups are fitted with fixed estimated guessing parameter. Estimated coefficients are then rescaled and Lord's and Raju's statistics are calculated with functions difLord and difRaju. P-values are then calculated based on χ^2 distribution with 2 degrees of freedom for Lord's statistic and based on standard normal distribution for Raju's statistic. For all methods Benjamini-Hochberg multiple comparison correction is applied (Benjamini & Hochberg, 1995; J. Kim & Oshima, 2013).

[6] Considering only positive values of parameter k.

Results

Simulation Study

Convergence Issues

Due to numerical estimation procedures in NLR and IRT-based methods, convergence issues occur. The number of generated data sets is decreasing with increasing sample size in all scenarios. 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 2,000 it is 1,064 (1,045 - 1,095) and for sample size of 5,000 it is 1,027 (1,017 - 1,033).

Lord's and Raju's procedures result in a large proportion of convergence problematic items (see Table 3), however with increasing number of examinees, proportion of convergence failures declines rapidly. Similar tendency can be observed in NLR procedure, however proportion of convergence failure items is less than 1% (0.08 - 0.68%) for all scenarios in contrast with Lord's and Raju's methods where proportions reach up over 10% (0 - 10.35%).

Rejection Rates

For almost all scenarios rejection rates (type I error, i.e. false positives) of the newly proposed NLR and also for MH and LR procedures maintain below the 5% nominal level. The nominal value is exceeded only when 3 uniform DIF items and sample size 5,000 are considered (see Part C of Table 3).

High rejection rates exceeding nominal level of 5% are apparent in IRT-based procedures in all studied scenarios with small sample sizes (1,000). Nevertheless with higher sample size, the rejection rates of Raju's method are below nominal value even in scenarios where it is not the case for non-IRT methods. Rejection rate of Lord's procedure is mildly exceeded also for sample size of 2,000 in case of no DIF (see Part A of Table 3) and in case of three uniform DIF items (see Part C of Table 3). In the case of three uniform DIF items and sample size of 5,000, similarly to non-IRT methods, the rejection rate of Lord's procedure is also exceeded.

The situation is similar in the case where no DIF item is present in data set. The nominal value of 5% is exceeded by IRT-based methods in case of sample size 1,000 and in case of sample size 2,000 it is exceeded only by Lord's procedure (see Part A of Table 3).

Power Rates

When uniform DIF is considered, NLR, LR and HM procedures yield satisfactory high power rate (over 80%) in almost all scenarios. Although the power analysis shows superiority of MH procedure, the differences between non-IRT methods are negligible, especially for smaller proportion of DIF items. While IRT-based procedures yield lower power in all uniform DIF scenarios, they gain satisfactory power on low rejection rate for sample size of 5,000 (see Parts B and C of Table 3).

For non-uniform DIF and sample size 1,000 no method achieves satisfactory power rates regardless of DIF items proportion (Parts D and E of Table 3). However, with increasing sample size power rates increase rapidly. LR 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, NLR procedure outmatches IRT-based methods. For larger proportion of DIF items, the power rates of NLR method and IRT procedures are comparable.

Real Data Analysis

As expected, item 49 is detected as DIF by almost all procedures (see Table 4). Only Lord's statistic does not show significant presence of DIF (χ^2-value = 11.130, p-value 0.077), even though this non-significant finding is not convincing. Confirming results from previous study (Vlčková, 2014), females performed better on this item than males as shown by LR, NLR and 3PL IRT model (see Figure 1).

The non-zero probability of guessing is apparent in both NLR and 3PL IRT models (see Table 5) suggesting that LR model may not have to be sufficient in this case. The estimated parameters are similar for both NLR and IRT models for most non-DIF items, larger differences can be observed in items 1 and 25.

Discussion and Conclusion

In this work we suggest using NLR procedure for DIF detection as a natural generalization of LR (Swaminathan & Rogers, 1990) by allowing nonzero probability for guessing c. In real data analysis, we demonstrate practical use of the newly proposed method and in a simulation study, we show its pleasant properties 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, NLR 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 newly proposed NLR method over IRT-model-based approaches is 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 (J. Kim & Oshima, 2013), IRT methods show large proportion of convergence failures (see Table 3). In practical implementation this may mean that with NLR procedure less time and effort is needed to fit models and to test for DIF than with IRT-based methods. Also, for low sample sizes (1,000) rejection rate (Type I error) exceeds 5% nominal value for IRT methods (see Table 3) and the power rate maintains on low level. Poor control of rejection rates for IRT methods when considering multiple DIF items is consistent with the finding of Wang and Yeh (2003). As expected, strong and consistent increasing trend in power rates with increasing sample size is obvious in all DIF detection procedures and all studied scenarios; except for MH method in non-uniform DIF detection (see further). For sample size of 5,000 power of all other procedures is almost 100%. With increasing sample size, the differences between methods decreases and IRT-based methods become easy to fit. IRT-based approaches then bring more precise model and added value in terms of estimates of latent trait while NLR and LR are only proxies to 3PL and 2PL IRT models.

Looking closer at non-IRT approaches, although MH 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 the newly proposed NLR method, together with LR procedure, can be seen as useful alternatives to IRT methods, especially in small sample sizes (1,000), where NLR and LR procedures outperform other methods in terms of power.

Moreover, in uniform DIF detection, NLR method achieves slightly better results than LR approach. This may suggest that NLR method profits from more precise model by introducing guessing parameter c into LR procedure. In non-uniform DIF detection the LR 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 J. Kim and Oshima (2013), which may be the case in non-uniform DIF detection in small sample sizes.

Our NLR procedure fills a logical gap in DIF detection methodology. While in IRT-based DIF detection methods the third parameter is often taken into account, LR accounts only for two parameters. NLR extends LR procedure in this way and to our best knowledge it is the only non-IRT method for DIF detection with the third, guessing parameter. The main difference between 3PL IRT-based methods and NLR approach is that in IRT-based procedures, the knowledge of examinees is modeled as unobserved latent variable with standard normal distribution; in contrast with NLR model where knowledge is represented by standardized total score of the test (Rao & Sinharay, 2006). Although NLR method can be viewed as less precise, it is easier to implement and interpret and also smaller sample size is required than for IRT model calibration. Thus NLR may be important not only for educational purposes but also can be seen as handy tool in identification of DIF items.

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, Béland, et al., 2010). Further, the estimated parameters are rescaled (Candell & Drasgow, 1988; Lautenschiager & Park, 1988) and then Lord's and Raju's statistics are calculated. It should be noted that this approach can lead to biased standard errors. Simultaneous estimation of parameters for both groups including guessing parameter is offered e.g. in R package mirt (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 NLR procedure may benefit from further improvements. Better specification of initial values could lead to smaller proportion of convergence issues. Also other estimating procedures can be implemented to provide more accurate estimates, such as weighted non-linear least squares or Bayesian based methods. NLR model can be extended by considering different guessing parameters for focal and reference group. Another generalization may be to allow upper asymptote to be smaller than one and thus introduce an non-IRT alternative to four-parameter IRT model (Barton & Lord, 1981). Similarly to IRT models (Reckase, 1985; Reckase & McKinley, 1991; Oshima, Raju, & Flowers, 1997), also NLR and NLR models can be extended by considering multidimensional tests. Besides, more than two groups can be taken into account (Magis, Raiche, Beland, & Gerard, 2010). NLR DIF detection method can be also refined by implementing iterative purification similarly as for LR (Zumbo, 1999) or IRT-based methods (Candell & Drasgow, 1988; Wang & Yeh, 2003).

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 equal group size is considered, which is not a necessarily realistic condition. Another restriction is that we consider only average difficulty items in non-uniform DIF design, where difficulty is the same for both groups.

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

Acknowledgements

This work was supported by Grant GJ15-15856Y from the Czech Science Foundation. The authors would like thank David Magis, Karel Zvára, Marek Brabec and Julie Ann Lorah for their comments on prior versions of this manuscript.

References

Agresti, A., & Kateri, M. (2003). *Categorial data analysis.* Berlin Heidelberg: Springer. doi: 10.1002/0471249688

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *The standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. doi: 10.2307/2346101

Berger, M., & Tutz, G. (2016). Detection of uniform and non-uniform differential item functioning by item focussed trees.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479).

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 (Vol. 48) (Computer software manual No. 6).

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An r package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30.

Dennis, J. E. J., Gay, D. M., & Welsch, R. E. (1981). An adaptive nonlinear least-squares algorithm. *Transactions on Mathematical Software (TOMS)*, *7*(3), 348–368. doi:

10.1145/355958.355965

Drabinová, A., Martinková, P., & Zvára, K. (2016). difnlr: Detection of dichotomous differential item functioning (dif) by non-linear regression function [Computer software manual]. (R package version 0.1)

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

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

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

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and mantel-haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903–915. doi: 10.1177/0013164403261769

Holland, P. W. (1985). On the study of differential item performance without irt. *Proceedings of the Military Testing Association*, 282-287.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the mantel-haenszel procedure. *Test validity*, 129–145.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

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.

Kim, J., & Oshima, T. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). Dif detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93–116. doi: 10.1111/j.1745-3984.2007.00029.x

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: 10.1007/s13398-014-0173-7.2

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

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. doi: 10.1348/000711009X474502

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*(4), 304–315. doi: 10.1177/0146621613475471

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., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research*, *46*(5), 733–755.

Magis, D., Raiche, G., Beland, S., & Gerard, P. (2010). 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

Magis, D., Tuerlinckx, F., & De Boeck, P. (2014). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135. doi: 10.3102/1076998614559747

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.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform dif. *Applied Psychological Measurement*, *20*(3), 257–274. doi: 10.1177/014662169602000306

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional irt-based internal measures of differential functioning of ltems and tests. *Journal of Educational Measurement*, *34*(3), 253–272. doi: 10.1111/j.1745-3984.1997.tb00518.x

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three mantel-haenszel procedures. *Applied Measurement in Education*, *14*(3), 235–259.

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

Rao, C. R., & Sinharay, S. (2006). *Handbook of statistics: Psychometrics* (Vol. 26). Elsevier.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412. doi: 10.1177/014662168500900409

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361–373. doi: 10.1177/014662169101500407

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological methods*, *8*(2), 185.

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with r*. Springer Science & Business Media. doi: 10.1007/978-0-387-93837-0

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

Vlčková, K. (2014). *Test and item fairness* (Unpublished master's thesis). Department of Probability and Mathematical Statistics, Charles University in Prague.

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

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: National Defense Headquarters*.

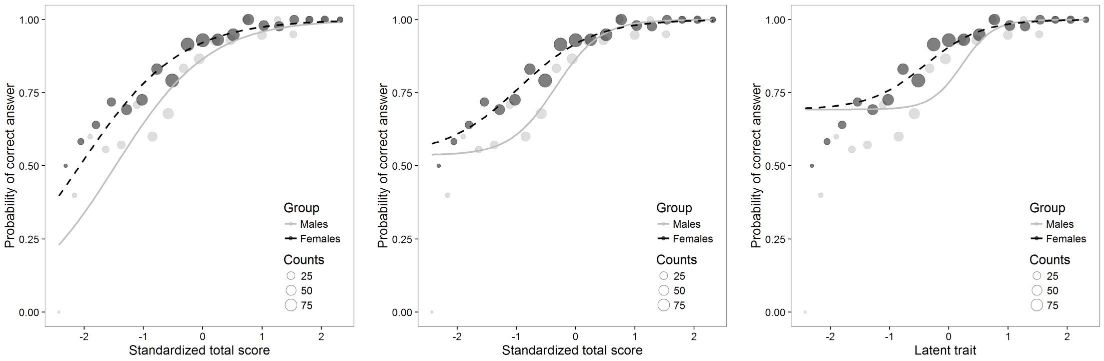

Figure 1: Estimated characteristic curves for females (black) and males (gray) of item 49 by LR (left), NLR (middle) and by 3PL IRT model (right). Plotted points represent proportion of correct answers for particular values of standardized total score. Their size is defined by number of examinees with the same value of standardized total score.

Table 1: Item Parameters Used to Generate DIF Items

DIF Type	Item	DIF Effect Size	Reference Group			Focal Group		
			a	*b*	*c*	*a*	*b*	*c*
Uniform	1	0.8	1	0	0.2	1	1	0.2
Uniform	1	0.4	1	0	0.2	1	0.50	0.2
	2	0.6	1	0	0.2	1	0.75	0.2
	3	0.8	1	0	0.2	1	1	0.2
Non-uniform	1	0.8	0.56	0	0.2	1.79	0	0.2
Non-uniform	1	0.4	0.90	0	0.2	2.01	0	0.2
	2	0.6	0.70	0	0.2	1.97	0	0.2
	3	0.8	0.56	0	0.2	1.79	0	0.2

Table 2: Item Parameters Used to Generate Non DIF Items Based on GMAT data.

Item		Parameters			Item		Parameters		
		a	*b*	*c*			*a*	*b*	*c*
2	4	0.29	−2.95	0.07	12	14	0.52	−1.96	0.07
3	5	0.41	−2.93	0.07	13	15	1.02	1.28	0.22
4	6	0.94	−1.21	0.33	14	16	0.65	0.49	0.16
5	7	0.88	−0.24	0.18	15	17	0.82	0.61	0.07
6	8	0.42	−1.15	0.07	16	18	1.04	2.11	0.37
7	9	0.74	0.60	0.36	17	19	0.95	0.81	0.09
8	10	0.35	−0.35	0.07	18	20	1.01	0.81	0.19
9	11	0.44	−0.30	0.07	19		0.98	1.67	0.28
10	12	0.55	−1.06	0.07	20		0.92	0.42	0.09
11	13	0.82	1.02	0.36	21		0.65	1.68	0.02

Table 3: Rejection rates (RR), power rates (PR) and proportion of convergence failures (CF) for Mantel-Haenszel (MH), Logistic Regression (LR), Non-linear Regression (NLR), Lord's (LORD) and Raju's (RAJU) procedures.

	Sample size = 1,000			Sample size = 2,000			Sample size = 5,000		
	RR	PR	CF	RR	PR	CF	RR	PR	CF
A. None DIF Item									
MH	0.195		0.000	0.175		0.000	0.280		0.000
LR	0.235		0.000	0.230		0.000	0.330		0.000
NLR	0.380		0.518	0.325		0.163	0.450		0.097
LORD	10.925*		5.768	5.095*		0.507	2.035		0.224
RAJU	9.315*		5.913	4.525		0.507	1.505		0.224
B. One Uniform DIF Item									
MH	0.658	97.100°	0.000	0.884	100.000°	0.000	1.516	100.000°	0.000
LR	0.642	96.000	0.000	0.868	100.000°	0.000	1.221	100.000°	0.000
NLR	0.821	96.500	0.682	1.079	100.000°	0.198	1.542	100.000°	0.107
LORD	9.995*	77.100	10.346	4.111	99.800	0.803	2.721	100.000°	0.019
RAJU	8.826*	61.400	10.209	3.626	92.700	0.803	2.184	100.000°	0.019
C. Three Uniform DIF Items									
MH	1.759	63.800°	0.000	4.235	86.067°	0.000	14.365*	99.267	0.000
LR	1.529	58.500	0.000	3.147	82.400	0.000	9.741*	98.633	0.000
NLR	1.953	60.133	0.562	3.988	83.133	0.199	11.618*	98.600	0.084
LORD	11.029*	47.500	8.294	5.229*	75.700	0.971	7.606*	97.200	0.000
RAJU	9.171*	37.767	8.294	3.994	65.900	0.971	4.400	94.867°	0.000
D. One Non-Uniform DIF Item									
MH	0.242	0.400	0.000	0.226	0.200	0.000	0.316	0.200	0.000
LR	0.468	46.400°	0.000	0.579	88.500°	0.000	0.700	100.000°	0.000
NLR	0.600	36.700	0.619	0.721	81.500	0.239	0.800	100.000°	0.126
LORD	10.626*	35.000	10.273	3.584	72.500	0.704	1.968	100.000°	0.393
RAJU	9.353*	14.200	10.133	2.837	45.200	0.704	1.621	100.000°	0.393
E. Three Non-Uniform DIF Items									
MH	0.182	0.133	0.000	0.200	0.167	0.000	0.265	0.300	0.000
LR	0.694	36.633°	0.000	1.382	78.233°	0.000	3.006	99.367	0.000
NLR	0.782	28.200	0.561	1.665	69.167	0.384	3.400	98.367	0.165
LORD	9.771*	35.800	5.690	3.918	77.833	0.589	2.218	99.767	0.000
RAJU	8.688*	17.333	5.539	3.653	68.433	0.589	2.471	99.900°	0.000

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 4: AT real data set analysis results. Calculated statistics of Mantel-Haenszel (MH), Logistic Regression (LR), Non-linear Regression (NLR), Lord's (LORD) and Raju's (RAJU) DIF detection procedures. The p-values were adjusted by Benjamini-Hochberg multiple comparison correction.

	MH		LR		NLR		LORD		RAJU	
	α_{MH}	p-value	Deviation	p-value	F-value	p-value	χ^2-value	p-value	Z-score	p-value
Item 49	12.446	0.008*	−14.760	0.012*	11.135	0.000*	11.130	0.077	−3.201	0.027*
Item 27	0.916	0.484	−1.213	0.727	0.385	0.756	0.398	0.893	−0.383	0.738
Item 41	0.132	0.843	−0.637	0.808	0.424	0.756	0.421	0.893	0.616	0.672
Item 7	1.540	0.429	−1.902	0.607	0.995	0.616	1.960	0.843	−1.193	0.493
Item 38	1.775	0.420	−2.146	0.607	1.895	0.502	1.940	0.843	−1.349	0.493
Item 28	0.387	0.673	−0.696	0.808	0.256	0.815	0.328	0.893	0.493	0.691
Item 9	2.620	0.420	−3.408	0.606	2.087	0.502	3.027	0.734	−1.402	0.493
Item 47	1.251	0.439	−4.323	0.576	3.254	0.279	3.146	0.734	−1.348	0.493
Item 75	0.026	0.951	−0.065	0.968	0.423	0.756	0.428	0.893	−0.680	0.672
Item 17	2.229	0.420	−2.795	0.607	1.555	0.529	1.237	0.893	−1.100	0.493
Item 76	2.192	0.420	−5.383	0.469	2.040	0.502	0.375	0.893	0.550	0.685
Item 10	0.004	0.951	−0.315	0.899	0.111	0.895	0.051	0.975	0.219	0.827
Item 64	0.006	0.951	−2.229	0.607	0.880	0.638	1.154	0.893	−1.039	0.498
Item 45	1.795	0.420	−2.566	0.607	0.690	0.716	3.290	0.734	−1.557	0.493
Item 24	1.397	0.431	−1.810	0.607	1.637	0.529	1.729	0.843	−1.164	0.493
Item 1	1.042	0.473	−1.710	0.607	1.094	0.610	0.770	0.893	0.705	0.672
Item 68	5.087	0.241	−5.307	0.469	3.179	0.279	4.948	0.734	−1.959	0.493
Item 61	1.727	0.420	−3.620	0.606	0.595	0.736	3.298	0.734	1.229	0.493
Item 25	1.863	0.420	−1.932	0.607	1.216	0.593	1.869	0.843	1.140	0.493
Item 2	0.379	0.673	−0.933	0.784	1.262	0.593	0.561	0.893	−0.658	0.672

An asterisk * indicates that p-value is smaller than nominal value of 0.05 and thus item is detected as DIF.

Table 5: AT real data set items' parameters NLR procedure and 3PL IRT model for males (index M) and females (index F). Parameter a represents discrimination, b difficulty and c guessing. Estimates are provided with standard errors (s.e.) in brackets.

	NLR					IRT				
	a_{M} (s.e.)	a_{F} (s.e.)	b_{M} (s.e.)	b_{F} (s.e.)	c (s.e.)	a_{M} (s.e.)	a_{F} (s.e.)	b_{M} (s.e.)	b_{F} (s.e.)	c (s.e.)
Item 49	2.62(0.73)	1.61(0.30)	−0.34(0.14)	−0.95(0.20)	0.54(0.05)	3.78(1.47)	2.11(0.48)	0.22(0.13)	−0.39(0.13)	0.69(0.05)
Item 27	1.52(0.24)	1.51(0.19)	1.30(0.10)	1.21(0.08)	0.04(0.03)	1.35(0.23)	1.33(0.17)	1.38(0.18)	1.29(0.12)	0.05(0.03)
Item 41	2.38(0.35)	2.40(0.27)	0.65(0.07)	0.58(0.05)	0.10(0.02)	2.97(0.67)	3.30(0.55)	0.66(0.08)	0.61(0.06)	0.13(0.02)
Item 7	1.28(0.19)	1.21(0.17)	0.24(0.14)	0.38(0.14)	0.00(0.06)	1.14(0.16)	1.02(0.11)	0.23(0.10)	0.42(0.09)	0.00(0.00)
Item 38	1.60(0.25)	1.78(0.25)	−0.50(0.15)	−0.64(0.13)	0.11(0.08)	2.17(0.42)	2.02(0.26)	−0.20(0.10)	−0.37(0.08)	0.27(0.08)
Item 28	1.88(0.35)	1.98(0.28)	1.72(0.11)	1.63(0.07)	0.08(0.02)	1.76(0.45)	1.90(0.35)	1.82(0.23)	1.68(0.14)	0.09(0.02)
Item 9	1.31(0.23)	1.38(0.21)	−0.79(0.27)	−0.96(0.26)	0.00(0.15)	1.10(0.17)	0.97(0.11)	−0.87(0.14)	−1.19(0.13)	0.00(0.01)
Item 47	3.12(0.96)	1.26(0.22)	−1.56(0.23)	−1.93(0.46)	0.36(0.19)	1.80(0.53)	1.28(0.29)	−1.28(0.27)	−1.15(0.23)	0.67(0.21)
Item 75	1.69(0.31)	1.51(0.22)	0.21(0.12)	0.14(0.12)	0.18(0.05)	2.14(0.48)	1.81(0.26)	0.46(0.10)	0.44(0.08)	0.29(0.05)
Item 17	1.63(0.27)	1.62(0.21)	1.28(0.10)	1.10(0.08)	0.08(0.03)	1.71(0.33)	1.56(0.21)	1.27(0.15)	1.17(0.11)	0.09(0.03)
Item 76	1.22(0.20)	1.61(0.24)	−0.05(0.16)	0.07(0.12)	0.13(0.06)	2.55(0.70)	2.88(0.52)	0.44(0.10)	0.50(0.07)	0.33(0.04)
Item 10	1.57(0.27)	1.69(0.26)	−1.36(0.27)	−1.34(0.25)	0.00(0.19)	1.18(0.20)	1.24(0.15)	−1.63(0.22)	−1.58(0.14)	0.00(0.05)
Item 64	1.25(0.30)	0.95(0.20)	0.66(0.20)	0.77(0.23)	0.20(0.07)	1.75(0.44)	1.32(0.27)	1.08(0.16)	1.31(0.17)	0.35(0.05)
Item 45	1.32(0.23)	1.21(0.19)	−0.80(0.26)	−0.72(0.28)	0.00(0.14)	1.66(0.28)	1.32(0.16)	−0.41(0.11)	−0.28(0.09)	0.22(0.11)
Item 24	1.21(0.18)	1.25(0.17)	0.47(0.13)	0.64(0.12)	0.00(0.05)	1.16(0.16)	1.08(0.11)	0.47(0.11)	0.66(0.09)	0.00(0.00)
Item 1	0.70(0.19)	0.82(0.22)	−0.99(0.89)	−0.68(0.79)	0.09(0.27)	1.59(0.54)	1.94(0.45)	0.74(0.17)	0.87(0.12)	0.54(0.04)
Item 68	1.18(0.22)	1.11(0.20)	−0.94(0.32)	−0.68(0.35)	0.00(0.17)	0.98(0.16)	0.86(0.10)	−1.05(0.17)	−0.82(0.11)	0.00(0.01)
Item 61	1.19(0.22)	1.37(0.23)	−1.15(0.37)	−1.17(0.33)	0.00(0.20)	0.99(0.18)	1.32(0.17)	−0.90(0.19)	−0.91(0.11)	0.20(0.23)
Item 25	1.03(0.21)	1.01(0.20)	−0.46(0.34)	−0.26(0.35)	0.08(0.14)	1.07(0.23)	1.23(0.19)	0.23(0.15)	0.45(0.11)	0.33(0.08)
Item 2	2.45(0.48)	1.99(0.28)	1.41(0.09)	1.32(0.07)	0.13(0.02)	2.75(0.90)	2.26(0.41)	1.39(0.14)	1.35(0.10)	0.16(0.02)

Appendix

Selected R code

```
# standardized total score
score <- c(scale(apply(data, 1, sum)))
# starting values are provided by startNLR() function from difNLR package
starting_values <- startNLR(data, group)

# functions of item characteristic curves
### with DIF
fun_dif   <- deriv3( ~ c + (1 - c) / (1 + exp(-(a + aDif * group) *
                       (score - (b + bDif * group)))),
                     namevec = c("a", "b", "c", "aDif", "bDif"),
                     function.arg = function(score, group,
                                             a, b, c, aDif, bDif){})
### with no DIF
fun_nodif <- deriv3( ~ c + (1 - c) / (1 + exp(-a * (score - b))),
                     namevec = c("a", "b", "c"),
                     function.arg = function(score, group, a, b, c){})

# fitting two models for characteristic curve of item 1
fit_dif <- nls(data[, 1] ~ fun_dif(score, group, a, b, c, aDif, bDif),
                algorithm = "port",
                start = starting_values[1, ],
                lower = c(-Inf, -Inf, 0, -Inf, -Inf),
                upper = c(Inf, Inf, 1, Inf, Inf))
fit_nodif <- nls(data[, 1] ~ fun_nodif(score, group, a, b, c),
                algorithm = "port",
                start = starting_values[1, 1:3],
              lower = c(-Inf, -Inf, 0),
              upper = c(Inf, Inf, 1))

# F-test of submodel
### F-statistic and p-value calculation
F_value <- ((fit_nodif$m$deviance() - fit_dif$m$deviance()) / 2) /
           (fit_dif$m$deviance() / (nrow(data) - 5))
p_value <- 1 - pf(F_value, 2, nrow(data) - 5)
```