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 Based on Non-Linear Regression

Adéla Drabinová, Patrícia Martinková

Technical report No. V1229

May 2016

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 053 400, fax: +420 286 585 789,
e-mail:martinkova@cs.cas.cz

Institute of Computer Science
The Czech Academy of Sciences

Detection of Differential Item Functioning Based on Non-Linear Regression

Adéla Drabinová[12], Patrícia Martinková[3]

Technical report No. V1229

May 2016

Abstract:

In this article, we present a new method for estimation of Item Response Function and for detection of uniform and non-uniform Differential Item Functioning (DIF) in dichotomous items based on Non-Linear Regression (NLR). Proposed method extends Logistic Regression (LR) procedure by including pseudo-guessing parameter. NLR technique is compared to LR procedure and Lord's and Raju's statistics for three-parameter Item Response Theory (IRT) models in simulation study based on Graduate Management Admission Test. NLR shows superiority in power at low rejection rate over IRT methods and outperforms LR procedure in power for case of uniform DIF detection. Our research suggests that the newly proposed non-IRT procedure is an attractive and user friendly approach to DIF detection.

Keywords:
Differential Item Functioning, Non-Linear Regression, Logistic Regression, Item Response Theory

[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

Introduction

Detection of DIF has been considered one of the most important topics in measurement. 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 potentionaly unfair. Recently, the topic of item and test fairness has been recognized as one of the three most important properties of any educational assessment in Standards for Educational and Psychological Testing (AERA et al., 2014) and methods for DIF detection are still being studied intensively (Kim et al., 2007; Kim and Oshima, 2013; Loken and Rulison, 2010; Magis et al., 2010a; Magis, 2013; Magis et al., 2014).

Two often used DIF detection methods are logistic regression (LR) procedure (Swaminathan and Rogers, 1990) and procedures based on IRT models (Lord, 1980; Raju, 1988, 1990). LR can be seen as a bridging method between IRT and non-IRT methods (Camilli and Shepard, 1994). LR procedure is a straightforward method which is easier to explain to audience and easier to apply in standard statistical packages than IRT methods, nevertheless it does not account for possibility of guessing, unlike methods based on three parameter (3PL) IRT model. The LR procedure has been extensively discussed in literature by many authors, including Choi et al. (2011); French and Maller (2007); Hidalgo and López-Pina (2004); Holland and Wainer (2012) and Zumbo et al. (1999).

In this work, we present general non-IRT approach to detect both uniform and non-uniform DIF in dichotomous items with presence of guessing. We use an extension of LR model to estimate item response function, which counts for the probability of guessing 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 newly proposed method fills the logical gap in DIF detection methodology.

The paper proceeds as follows: Methodology for DIF detection with NLR, simulation study and practical implementation within ... R software package are described in section . Results of simulation study are presented in section in which proposed NLR procedure is compared to LR model and to Lord's and Raju's procedures based on 3PL IRT model in simulations based on Graduate Management Admission Test (GMAT) data set (Kingston et al., 1985). Discussion and conclusion is offered in section .

Methodology

Non-Linear Regression for Description of Item Properties

To provide more proper analysis of item properties but stay within non-IRT framework, we propose an extension of LR model by accounting for probability of guessing. We assume that correct answer may be guessed with certain probability c without the necessary knowledge, and if it is not guessed, the probability is modeled by LR model. Using logistic parameterization, the probability of correct answer to an item i by j-th examinee is given by equation

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

where Y_{ij} is response of j-th examinee to an item i (1 for correct, 0 for incorrect) and X_j stands for his/her observed knowledge (the standardized total test score). When considering zero probabily of guessing ($c_i = 0$) then (1) reduces into LR model. The interpretation of regression coefficients is the same as for LR (Agresti and Kateri, 2003, Chapter 5).

Using IRT parameterization model (1) can be rewritten as

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

where variables Y_{ij} and X_j are as above. Regression parameter a_i is discrimination parameter of i-th

item and b_i is difficulty parameter of i-th item[4]. The interpretation of guessing parameter c_i is the same for both parameterizations, that is probability that correct answer of item i is guessed without necessary knowledge. From now on we will call model given by equation (2) NLR model. NLR model is proxy for 3PL IRT model (Lord et al., 1968) as it is by definition score-based method and thus non-IRT.

Non-Linear Regression DIF Detection Procedure

Usage of LR model for detection of DIF was introduced by Swaminathan and Rogers (1990) and is one of the most widely used methods in uniform and non-uniform DIF detection. This procedure became very popular in the study field mainly due to its easy interpretation and straightforward estimation of parameters and performance of tests. Formally, the detection is based on introducing group membership variable into LR model.

We extend the LR procedure by including pseudo-guessing parameter. We assume that probability of guessing is the same for both groups. Using LR parameterization, the probability of correct answer to an item is then given by equation

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

where variables Y_{ij} and X_j are as above and variable G_j represents group membership of j-th examinee (1 for reference group, 0 for focal group). For case without guessing ($c_i = 0$), the model is formally equivalent to model proposed by Swaminathan and Rogers (1990) .

Using IRT parameterization, equation (3) can be rewritten as

$$\mathrm{P}(Y_{ij}=1|X_j,G_j)=c_i+(1-c_i)\frac{e^{(a_i+a_{\mathrm{DIF}i}G_j)(X_j-(b_i+b_{\mathrm{DIF}i}G_j))}}{1+e^{(a_i+a_{\mathrm{DIF}i}G_j)(X_j-(b_i+b_{\mathrm{DIF}i}G_j))}}, \tag{4}$$

where variables Y_{ij}, X_j and G_j and regression parameters a_i, b_i and c_i are as above. Parameter $a_{\mathrm{DIF}i}$, respectively $b_{\mathrm{DIF}i}$, represents the difference in discrimination, respectively in difficulty, of reference and focal group.[5] Henceforward we will call model given by (4) DIF NLR model.

In what follows, we stick with DIF NLR model. For logistic parameterization (3), parameter estimation procedures and DIF detection methods would be analogous.

Estimation and DIF Detection

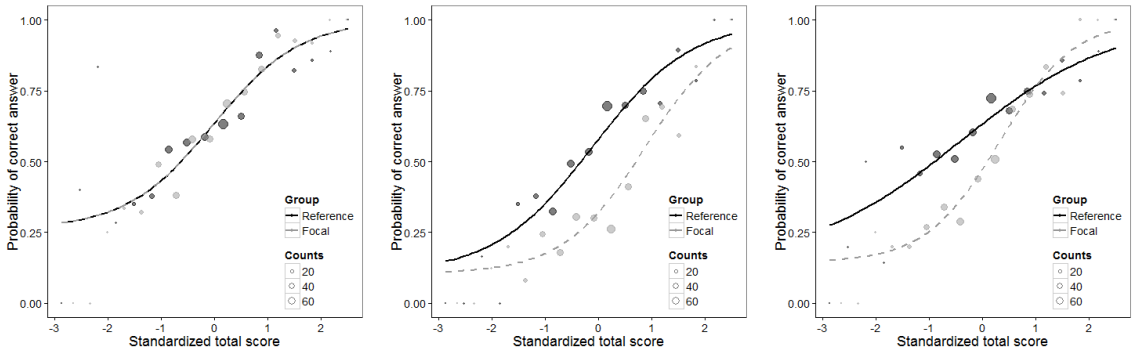
The parameter estimates of model are determined by non-linear least square estimation, that is by minimization of the residual sums of squares (RSS) with respect to $(a_i, b_i, a_{\mathrm{DIF}i}, b_{\mathrm{DIF}i}, c_i)$:

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

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 (4) can be utilized to detect DIF in a simple way. If value of $a_{\mathrm{DIF}i}$ is zero and value of $b_{\mathrm{DIF}i}$ differs from zero, this suggests presence of uniform DIF. If value of $a_{\mathrm{DIF}i}$ 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:

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



(a) (b) (c)

Figure 1: Examples of fitting characteristic curves by NLR DIF procedure for three different items with respect to DIF nature. Figure 1a represents item with no DIF, Figure 1b uniform DIF item and Figure 1c non-uniform DIF item. Plotted points represent probability of correct answer for particular values of standardized total score. Their size is defined by number of examinees with the same value of standardized total score.

These scenarios can be presented by graphical representation of characteristic curves, see Figure 1.

To compare two nested NLR models, and thus test for DIF presence in item i, the F-test or likelihood ratio test can be used with similar results (Dennis et al., 1981; Ritz and Streibig, 2008). 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 is held only approximately (Ritz and Streibig, 2008). In simulation study and data analysis, we stick with F-test for simplicity. It should be noted that, in contrast with IRT methods, tests for NLR DIF and LR procedures are performed separately for each item. Thus Benjamini-Hochberg (BH) adjustment procedures for multiple testing is applied, as suggested by Kim and Oshima (2013).

Simulation Study

We compare the proposed NLR method with Lord's and Raju's statistics (Lord, 1980; Raju, 1988, 1990) in terms of power and I. type error rate. In order to illustrate possible benefit by including guessing parameter, also detection procedure based on LR (Swaminathan and Rogers, 1990) is considered into simulation study.

The probabilities of correct answers are calculated based on 3PL IRT model. 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 parameters and to be in line with previous simulation studies (Swaminathan and Rogers, 1990; Narayanan and Swaminathan, 1996; Jodoin and Gierl, 2001; Güler and Penfield, 2009; Kim and Oshima, 2013), simulation study is based on parameters according to dataset from the 1985 problem solving of the GMAT (Kingston et al., 1985, p. 47). Responses of examinees are generated from Bernoulli distribution with calculated probabilities, which come from the true values of item and person parameters.

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

[5]The effect of group membership β_{2i} in parameterization (3) is equal to $-a_i b_{\mathrm{DIF}i} - a_{\mathrm{DIF}i} b_i - a_{\mathrm{DIF}i} b_{\mathrm{DIF}i}$ in parameterization (4). The effect of interaction of total score and group membership β_{3i} is equal to $a_{\mathrm{DIF}i}$.

In total, the simulated test consists of 20 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). 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 nonuniform 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. One, or three items are considered as uniform, or non-uniform DIF. When one DIF item is considered, the large size of DIF is chosen. Mixture of DIF sizes are 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.

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					

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 the estimation procedures in NLR and 3PL IRT models, convergence issues can be observed. In such cases, estimation is carried out without problematic items and no results for these items are obtained. All tests are performed on $\alpha = 0.05$ significance level. Based on items without convergence issues, type I error rate and power of procedure are calculated. 1,000 iterations are considered.

Practical Implementation

For all analyses, software R, Version 3.22 is used. The LR procedure is implemented by function glm from stats package. To detect DIF, likelihood ratio test was performed (Agresti and Kateri, 2003) and BH multiple comparison correction was applied (Benjamini and Hochberg, 1995; Kim and Oshima, 2013). R package difR (Magis et al., 2010a) is used to perform IRT analysis including fitting 3PL IRT model with function itemParEst, Lord's test with function difLord and Raju's test with function difRaju. The NLR procedure is implemented in new function ... within ... R package which uses nls function from stats package with constraints on guessing parameter (Dennis et al., 1981; Ritz and Streibig, 2008). To test for DIF presence, F-test is used.

To detect global minimum by nonlinear least-square estimation, it is necessary 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 inflex 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 $a = \frac{4k}{1-c}$.

Results

Convergence Issues

Due to numerical estimation procedures in NLR DIF and IRT based methods, convergence issues can be noticed. It should be noted that large proportions of convergence failures can have significant impact on power and rejection rates. For convergence problematic items no results are obtained and no conclusion about DIF detection can be drawn. Thus rejection and power rate analysis is based only on converged items. Especially, when there is a large proportion of these items, the results should be interpreted with caution.

Lord's and Raju's procedures perform 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% (0.07 - 0.54%) for all scenarios in contrast with Lord's and Raju's methods where proportions reach up over 10% (0 - 12.29%).

Rejection Rates

For almost all scenarios rejection rates of NLR DIF and LR procedures maintain below the 5% nominal level. Empirical Type I error rates range from 0.56% to 11.66% for NLR DIF and from 0.45% to 9.74% for LR method. The nominal value is exceeded only when 3 uniform DIF items and sample size 5.000 are considered (see Part B of Table 3).

High rejection rates of Lord's and Raju's procedures are apparent in all studied scenarios and they exceed the nominal level of 0.05. The minimal rejection rates were 7.7% for Lord's procedure and 8.3% for Raju's procedure (Table 3). Considerable rates occur primarily for smaller sample size, where proportions of convergence issues are large.

[6] Considering only positive values of parameter k.

Power Rates

The power analysis shows superiority of proposed NLR procedure through all sample sizes when uniform DIF is considered (Parts A and B of Table 3). Except the case of three uniform DIF items and sample size 1,000, NLR performs very high power rates (83.07% - 100%).

For non-uniform DIF and sample size 1,000 no method achieves satisfactory power rates regardless of DIF items proportion (Parts C and D of Table 3). However with increasing sample size power rates increase. LR procedure outperforms other methods in terms of power at low reject rate in all scenarios with power rate ranging from 37.97% to 100%.

With sample size 5,000 all procedures are able to detect presence of DIF almost certainly regardless of the nature or proportion of DIF items, however, only rejection rates of NLR and LR based methods remain below nominal value of 5% even in cases when other procedures do not.

Table 3: Rejection rates (RR), power rates (PR) and proportion of convergence failures (CF) for NLR, LR, LORD and RAJU procedures.

	Sample size = 1,000			Sample size = 2,000			Sample size = 5,000		
	RR	PR	CF	RR	PR	CF	RR	PR	CF
A. One Uniform DIF Item									
NLR	0.79	97.40°	0.45	1.07	100.00°	0.23	1.54	100.00°	0.07
LR	0.65	96.90	0.00	0.87	100.00°	0.00	1.23	100.00°	0.00
LORD	15.71*	95.10	12.29	10.08*	100.00	0.85	9.60*	100.00	0.11
RAJU	14.32*	90.99	12.29	9.91*	99.70	0.85	9.28*	100.00	0.11
B. Three Uniform DIF Items									
NLR	1.92	59.46°	0.56	3.98	83.07°	0.20	11.66*	98.63	0.10
LR	1.56	58.27	0.00	3.17	82.37	0.00	9.74*	98.60	0.00
LORD	17.78*	70.07	10.24	11.27*	87.98	0.92	16.49*	98.80	0.00
RAJU	16.14*	67.95	10.24	10.21*	86.24	0.92	13.31*	98.40	0.00
C. One Non-Uniform DIF Item									
NLR	0.56	36.84	0.53	0.73	81.90	0.21	0.83	100.00°	0.14
LR	0.45	47.00°	0.00	0.57	88.90°	0.00	0.73	100.00°	0.00
LORD	15.78*	68.12	12.03	9.18*	95.87	0.76	7.70*	100.00	0.30
RAJU	14.61*	62.59	12.03	8.96*	96.07	0.76	8.38*	100.00	0.30
D. Three Non-Uniform DIF Items									
NLR	0.78	28.76	0.54	1.77	69.73	0.24	3.43	98.23	0.08
LR	0.74	37.97°	0.00	1.44	78.57°	0.00	2.96	99.37°	0.00
LORD	15.78*	64.72	10.78	9.03*	92.30	0.84	7.37*	99.97	0.10
RAJU	15.19*	61.07	10.78	9.47*	93.88	0.84	8.34*	99.97	0.10

NLR = non-linear regression, LR = logistic regression, LORD = Lord's statistics, RAJU = Raju's statistics.
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%.

Discussion and Conclusion

In this work we suggest using NLR procedure for DIF detection which is natural generalization of LR (Swaminathan and Rogers, 1990) by allowing nonzero probability for guessing c. In a simulation study, we compare newly proposed NLR DIF detection method to LR procedure and Lord's and Raju's methods based on 3PL IRT models.

Results of our simulation study (Table 3) show that NLR keeps pleasant properties of LR, especially low rates of convergence issues. Despite our assumption that sample size of 1,000 each group would provide a sufficient sample size for item calibration (Kim and Oshima, 2013), IRT methods show large proportion of convergence failures (see Table 3). Note that high number of convergence issues can distort rejection and power rates. [7] In practical implementation this may mean that less time and effort is needed to fit NLR DIF procedure and test for DIF than with IRT models.

NLR and LR procedures' rejection rates (Type I error), calculated from converged simulation runs, maintain below 5% nominal value for almost all scenarios, in contrast with both IRT methods (see Table 3). Poor control of rejection rates for IRT methods when considering multiple DIF items is consistent with the finding of Wang and Yeh (2003). However inability of Type I error control in scenarios with low proportion of DIF items is surprising. One explanation can be large proportion of convergence issues. When considering all convergence failures to be not DIF detected (results not shown), rejection rates for 3PL IRT methods decrease, but they maintain over nominal value of 5% for IRT methods. [8] This may indicate that IRT methods are less robust than LR and NLR DIF procedures considering more variable items' parameters such as are present in GMAT (Table 2).

Analysis of power shows that in *uniform DIF* detection the newly proposed NLR DIF procedure is superior to all other methods in terms of power at low rejection rate (parts A and B of Table 3). This may suggest that NLR DIF detection method profits from more precise model, comparing to LR procedure. A comparison of power rate in *non-uniform DIF* detection indicates superiority of LR procedure at low I Type error rate. One explanation may be, that we consider only non-uniform DIF items with the same difficulty parameter for both groups. 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. For sample size of 5,000 power of all procedures is almost 100%.

For NLR and LR methods, BH multiple comparison correction is applied. Negative effect of using such procedures can be decrease of power, as noted by Kim and Oshima (2013). That can be an explanation of unsatisfactory low power rates in non-uniform DIF detection for NLR and LR methods in small sample sizes. Even though IRT procedures achieve very good power rates, their rejection rates exceed nominal value of 5% over all sample size levels. Previous simulation studies suggested that IRT based methods do not benefit from multiple comparison correction (Kim and Oshima, 2013). To be on the safe side, simulations with BH corrections were also conducted, however in agreement with Kim and Oshima (2013) did not yield any benefit (results not shown).

Our NLR procedure fills a logical gap in DIF detection methodology. While in IRT models three parameters are 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 guessing parameter. The main difference between 3PL IRT and NLR DIF models is that in IRT knowledge of examinees is modeled as sample from standard normal distribution in contrast with NLR model where knowledge is represented by standardized total score of the test (Rao and Sinharay, 2006). Although NLR DIF method can be viewed as less precise, it is easier to implement and interpret and thus NLR may be important for educational purposes.

Important feature of our approach is that group membership is considered as independent variable

[7] Analysis of convergence issues shows that some items tend to diverge more often than others. Especially, items with low difficulty (items 2 and 3 in Table 2), items with large difficulty and large probability of guessing (items 16 and 19 in Table 2) and items with low discrimination (item 8 in Table 2) show larger proportion of convergence issues in both methods. Unlike non-DIF items, DIF items tend to converge more often for all scenarios, which is predictable when considering selected parameters (Table 1).

[8] We also notice that some of non-DIF items tend to be detected as DIF more often than others. Especially items with average values of parameters are falsely detected as DIF by IRT methods, suggesting hypersensitivity in such items. On the other hand, convergence problematic items, once they converge, are more likely to be detected correctly as non-DIF. These two phenomenons may cause large rejection rates in IRT methods and also their decreasing tendency with increasing sample size (see Table 3).

and hence model is fitted for both groups, reference and focal, simultaneously. The common way of applying 3PL IRT model in DIF detection, when considering the same probability of guessing for both groups, is to fit model on all data and estimate guessing parameter. Fixed estimate of guessing parameter is then applied into two separate models for focal and reference group (Magis et al., 2010a). Estimated parameters are rescaled (Candell and Drasgow, 1988; Lautenschiager and Park, 1988) and then Lord's and Raju's statistics are calculated. However this approach can lead to underestimation of standard errors. Possible improvement for IRT methods thus may be taking approach similar to ours and fitting models for both groups together.

We believe that also the currently proposed NLR procedure may benefit from further improvements. Better specification of initial values can lead to smaller proportion of convergence issues. Also another 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 parameter for focal and reference group. Another generalization may be to allow upper asymptote to be smaller than one and thus introduce an alternative to four-parameter IRT models (Barton and Lord, 1981). Similarly to IRT models (Reckase, 1985; Reckase and McKinley, 1991; Oshima et al., 1997), also NLR and NLR DIF models can be extended by considering multidimensional latent trait. More than two groups can be taken into account (Magis et al., 2010b). NLR DIF detection method can be refined by implementing iterative purification similarly as for LR (Zumbo et al., 1999) or IRT methods (Candell and Drasgow, 1988; Wang and 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 large sample sizes are considered with equal group size, which is not 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 that NLR appears to be an attractive and user friendly alternative to other procedures used in DIF detection. NLR allows for incorporation of probability of guessing into the model while it keeps the simplicity of LR model.

Acknowledgements

This research was supported by grant funded by Czech Science Foundation under number GJ15-15856Y. We wish to thank Karel Zvára for initial ideas, consultations and help with R code and to David Magis and Marek Brabec for helpful comments on earlier draft.

References

AERA, APA, and NCME (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Agresti, A. and Kateri, M. (2003). *Categorial Data Analysis*.

Barton, M. A. and Lord, F. M. (1981). An Upper Asymptote for the Three-Parameter Logistic Item-Response Model. *ETS Research Report Series*, 1981(1):1–8.

Benjamini, Y. and 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.

Camilli, G. and Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*.

Candell, G. L. and Drasgow, F. (1988). An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory. *Applied Psychological Measurement*, 12(3):253–260.

Choi, S. W., Gibbons, L. E., and 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., and Welsch, R. E. (1981). An Adaptive Nonlinear Least-Squares Algorithm. *Transactions on Mathematical Software (TOMS)*, 7(3):348–368.

French, B. F. and 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.

Güler, N. and 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.

Hidalgo, M. D. and 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.

Holland, P. W. and Wainer, H. (2012). *Differential Item Functioning*.

Jodoin, M. G. and 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. and Oshima, T. C. (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., and Kim, S. (2007). DIF Detection and Effect Size Measures for Polytomously Scored Items. *Journal of Educational Measurement*, 44(2):93–116.

Kingston, N., Leary, L., and 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.

Lautenschiager, G. J. and Park, D.-G. (1988). IRT Item Bias Detection Procedures: Issues of Model Misspecification, Robustness, and Parameter Linking. *Applied Psychological Measurement*, 12(4):365–376.

Loken, E. and Rulison, K. L. (2010). Estimation of a Four-Parameter Item Response Theory Model. *British Journal of Mathematical and Statistical Psychology*, 63(3):509–525.

Lord, F., Novick, M., and Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*.

Magis, D. (2013). A Note on the Item Information Function of the Four-Parameter Logistic Model. *Applied Psychological Measurement*, 37(4):304–315.

Magis, D., Béland, S., Tuerlinckx, F., and De Boeck, P. (2010a). A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning. *Behavior research methods*, 42(3):847–862.

Magis, D., Raiche, G., Beland, S., and Gerard, P. (2010b). A Generalized Logistic Regression Procedure to Detect Differential Item Functioning among Multiple Groups. *International Journal of Testing*, 11(4):365–386.

Magis, D., Tuerlinckx, F., and De Boeck, P. (2014). Detection of Differential Item Functioning Using the Lasso Approach. *Journal of Educational and Behavioral Statistics*, 40(2):111–135.

Narayanan, P. and Swaminathan, H. (1996). Identification of Items that Show Nonuniform DIF. *Applied Psychological Measurement*, 20(3):257–274.

Oshima, T. C., Raju, N. S., and Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests. *Journal of Educational Measurement*, 34(3):253–272.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4):495–502.

Raju, N. S. (1990). Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions.pdf. *Applied Psychological Measurement*, 14(2):197–207.

Rao, C. R. and Sinharay, S. (2006). *Handbook of Statistics: Psychometrics*.

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement*, 9(4):401–412.

Reckase, M. D. and McKinley, R. L. (1991). The Discriminating Power of Items That Measure More Than One Dimension. *Applied Psychological Measurement*, 15(4):361–373.

Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R*. Springer Science \& Business Media.

Swaminathan, H. and Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, 27(4):361–370.

Wang, W.-C. and 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.

Zumbo, B. D., Ph, D., and Wild, W. R. (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*.