A Regularization Method for Detecting Differential Item Functioning under the Framework of Generalized Linear Models

Jing Jiang[1], Zhushan Li[1]
[1]Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467

Abstract

The purpose of the study is to present a regularization method for estimating differential item functioning (DIF) parameters under generalized linear models. DIF occurs when the probabilities of correctly responding to an item are unexpectedly different for individuals from different groups with a same latent ability level. Traditional DIF detection approaches usually require all items except the one under detection to be DIF-free, which is possibly wrong. Otherwise, failing to identify invariant anchors will lead to inflated type I errors. This problem can be solved by simultaneous estimation of DIF parameters in one model by using regularized logistic regression. Simulation studies were conducted to compare this proposed method with other DIF detection techniques such as Mantel-Haenszel method and logistic regression method, and the results indicated the feasibility and applicability of the proposed method.

Key Words: differential item functioning, regularized logistic regression, item response theory model

1. Introduction

Nowadays, educational and psychological tests are widely used to measure individuals' latent traits such as intelligence, attitudes, and other abilities or skills. For this purpose, high-quality items are in demand in order to provide a valid and accurate measure for the latent trait. Differential item functioning (DIF) has received a lot of attention over the past decades, which occurs when the probabilities of correctly answering an item are unexpectedly different for people from group to group with a same latent ability level (Holland & Wainer, 1993). Items with DIF may reflect measurement bias (Millsap & Everson, 1993) and may lead to discrimination against particular groups (Zumbo, 1999).

DIF is a very important indicator for researchers and test developers to confirm that items display the same statistical properties for individuals from different groups within the population. Generally, uniform DIF indicates the item of interest consistently gives one group an advantage across all ability levels, and non-uniform DIF occurs when an item gives an advantage to a reference group at one end of the ability continuum while favors the focal group at the other end (Walker, 2011). In the context of item response theory (IRT), an item showing uniform DIF only varies in the difficulty parameter, while an item displaying non-uniform DIF varies in the discrimination parameter, and possibly the difficulty parameter (Mellenbergh, 1982).

A large number of parametric and non-parametric methods have been developed to detect DIF over the years, such as the Mantel-Haenszel (NH) test (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), SIBTEST (Shealy & Stout, 1993), and Raju's area measures (Raju, 1988). These methods are often conducted for each item and treated as an item-by-item approach (Swanson et al., 2002), which typically focus on analyzing each item individually. There are several problems with these approaches. First, the assumption that all items except the studied item or the anchor items are supposed to be invariant over groups is not guaranteed (Magis, Tuerlinckx, & De Boeck, 2015). Previous studies (e.g., Wang & Yeh, 2003; Wang, 2004; Stark et al., 2006; Woods, 2009) suggest that if a set of anchor items is contaminated, the Type I error rate are often inflated, and the test score may not be a fair measure for the latent traits. In addition, those DIF detection approaches are based on multiple testing since every item is tested at a time, thus adjustment procedures such as Bonferroni correction or Holm's procedure, should be used to evaluate DIF items in order to control Type I error rates (Kim & Oshima, 2012). Moreover, since multidimensionality is commonly recognized as a possible cause of DIF, the individual-item focused approaches may fail to explain such causes of DIF and are unable to guide researchers in reviewing the testing items (O'Neill & McPeek, 1993).

An intuitive solution to address these problems is to detect DIF items on a test or assessment at the same time, and Magis, Tuerlinckx & Boeck (2015) proposed an approach called LR Lasso DIF method which allows simultaneous DIF detection of all items in a single modeling approach. This method is based on logistic regression (LR) and focuses on the identification of uniform DIF in Rasch models. The L1-norm penalty for DIF parameters, that is the Lasso penalty, was added to the log-likelihood in estimation, and a higher value of the penalty term shrinks more model's coefficients for DIF parameters towards zero. However, the authors used examinee's test score as a proxy for ability. From the point of view of IRT, individuals correctly respond to the same number of items may have different levels of ability if those items vary in their difficulties. In order to obtain more accurate results, the DIF analysis model was modified and the estimated person ability from IRT modeling rather than the test score was used to represent examinee's latent ability level in DIF analysis.

2. Research Purpose

This study aims to use the regularization method to detect DIF items. Comprehensive simulation studies were conducted to evaluate the performance of uniform DIF detection under the framework of GLMs. In this case, the regression coefficients representing DIF for each item can be estimated simultaneously by including all item and person characteristics in a single model, and the assumption that the anchor set should be DIF-free is no longer required. This object can be achieved if the Type I error is well controlled below 0.05 and the power is good under different simulated conditions. Also, the proposed method was compared with other commonly used DIF detection techniques including LR and MH test.

3. Method

3.1 DIF Detection Model

The mathematical form of one-parameter logistic model can be written as:

$$P\left(Y_{ij} = 1 \middle| \theta_j\right) = \pi_{ij} = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)}$$

or

$$\text{Logit}\left(\pi_{ij}\right) = \theta_j - b_i, \theta_j \sim N(0, \sigma^2)$$

where $P(Y_{ij} = 1|\theta_j)$ is the probability that an examinee j *(j = 1,..., J)* with ability level of θ_j responds to an item i correctly, $Y_{ij} \sim Bernoulli(\pi_{ij})$, and b_i is the difficulty of item *I (I = 1,..., I)*. This is the standard logistic regression model for predicting a dichotomous outcome variables from given independent variables (Bock, 1975).

The above logistic regression model can be used to model DIF by adding a variable representing group membership (e.g., focal group and reference group):

$$\text{Logit}(\pi_{ijh}) = \theta_j - b_i + \gamma_i G_{jh}$$

where π_{ijh} is the probability that a person j in group *h* responds to item *i* correctly; b_i is the difficulty of item i; θ_j represents the latent ability level of person j; G_{jh} indicates group membership: $G_{jh} = 1$ when person j belongs to group h, otherwise $G_{jh} = 0$; γ_i corresponds to the group difference in terms of item i, which are the parameters of interest for DIF modeling. Once the model is fitted, the DIF effects can be identified by examining γ_i ($\gamma_i \neq 0$).

3.2 Penalized Estimation

Let $\boldsymbol{\omega} = (b_1, \ldots, b_I, \gamma_1, \ldots, \gamma_I)$ which contains all model parameters. Traditional maximum likelihood estimation (MLE) aims to find a set of parameter values that maximizes the log-likelihood $l(\boldsymbol{\omega})$ of making the observations given the parameters. The regularization approach is designed to maximize the penalized log-likelihood function $l_p(\boldsymbol{\omega})$ rather than the log-likelihood function $l(\boldsymbol{\omega})$ when estimating DIF parameters, which can be expressed as follows:

$$l_p(\boldsymbol{\omega}) = l(\boldsymbol{\omega}) - \lambda \boldsymbol{J}(\boldsymbol{\omega})$$

Here, $\boldsymbol{J}(\boldsymbol{\omega})$ is a L1-penalty term that penalizes specific structures in the parameter vector $\boldsymbol{\omega}$, and λ is the penalty parameter.

Two commonly used penalty terms are L1-norm and L2-norm penalty, which are used in Lasso regression and ridge regression correspondingly. The choice of using an L1-penalty shrinks the model's coefficients towards zero and returns a very sparse solution, while using an L2-penalty means the estimated coefficients approach zero but do not equal zero exactly. In terms of our research purpose, that is to detect DIF items, it is more appropriate to employ L1-penalty since we expect the coefficients for non-DIF parameters are exactly zero. Obviously, if $\lambda = 0$, the maximum likelihood estimated is obtained; when λ increases and is greater than zero, more DIF parameters are shrunk to zero; if $\lambda \to \infty$, all parameters are equal to zero.

An important issue in penalized estimation is the choice of the penalty parameter λ which determines the number of items flagged as DIF items. Usually, information criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), as well as cross-validation (CV) are used to find the optimal tuning parameter. According to previous findings (e.g., Magis et al., 2015), AIC and CV have a higher power but also a higher Type I error rate than BIC and flag more non-DIF items as DIF items. Therefore, we only consider BIC in the simulation studies since it is more conservative:

$$\lambda_{BIC} = \arg\min \text{BIC}(\lambda) = \arg\min(-2l(\widehat{\boldsymbol{\omega}}) + K(\widehat{\boldsymbol{\omega}}) \cdot \log n)$$

where $l(\widehat{\boldsymbol{\omega}})$ is the log-likelihood of the current parameter vector $\widehat{\boldsymbol{\omega}}$; $K(\widehat{\boldsymbol{\omega}})$ indicates the number of free parameters to be estimated; and *n* is the number of item responses in the dataset.

3.3 Simulation Design

Simulated datasets were generated to perform DIF analyses with *J* persons and *I* item responses. Three test lengths were examined, consisting of 20 items, 40 items, and 60 items. Three group sizes of 500, 1000, or 2000 subjects were considered and the focal and reference group have the same sample size. Two percentages of DIF with 10% or 20% were taken into account. In terms of direction and magnitude of DIF, only unidirectional drift on the item difficulty parameter with DIF size of 0.4 or 0.8 were considered. Latent ability distribution is set to *N(0,1)* for both reference and focal group. Thus, there are in total 36 conditions; for each condition, 500 replications were generated.

Item parameters of 60 multiple-choice items from a statewide test were used as true item parameters, and the first 20, 40 and all values were used under different conditions. The mean and standard deviation for 20 items, 40 items, and 60 items are 0.70 and 0.76, 0.64 and 0.82, and 0.68 and 0.81correspondingly. Simulated item responses were generated following the 1PL item response model.

Four DIF detection methods were performed in the simulation studies: 1) the proposed lasso LR method using ability estimates in the DIF model; 2) the LR Lasso DIF approach (Magis et al., 2015) using test score in the DIF model; 3) traditional LR; 4) MH test. The power values indicating the proportion of DIF items that are correctly flagged as DIF items, and Type I errors indicating the proportion of non-DIF items that are incorrectly flagged as DIF items were recorded as outcome measures for each approach. The full simulation studies were conducted in R (R Development Core Team, 2013).

4. Results

Figure 1 and Figure 2 demonstrate the power and Type I error rates for all four approaches by group size (500, 1000 or 2000) and DIF magnitude (0.4 or 0.8), depending on test length of 20 items (the first row), 40 items (the second row), or 60 items (the last row), and 10% (the first column) or 20% (the second column) DIF items in the test correspondingly.

According to Figure 1, using IRT ability estimates rather than total test score in the regularized LR model leads to higher power in all situations. MH and LR approaches generally have very similar performance in detecting DIF items correctly. Specifically, when sample size or DIF magnitude is large, the proposed lasso LR in this paper exhibits same or greater power in detecting DIF items compared to LR or MH methods. In other situations, that is, when sample size and DIF magnitude are both small, for example, when group size equals to 500 and 1000, and the DIF magnitude is 0.4, MH and LR methods have a much better performance than the other two lasso approaches.

The results are not surprising since the more conservative strategy—BIC criterion was used to select the penalty parameter. Also, previous research examined that when the DIF magnitude is small than or equal to 0.4, there is a very minimal effect on equating and ability estimation (Wells, Subkoviak, & Serlin, 2002). Therefore, when the DIF magnitude is small, it is more important to prevent over identification of DIF items. If in some situations a higher power is desired regardless of the Type I error, other criteria such as AIC or CVs can be considered when determining the cutoff value of the penalty parameter.

In Figure 2, the proposed lasso method has the lowest Type I error rates in all manipulated conditions, and even the largest Type I error is smaller than 0.05. However, there are many false alarms using MH and LR methods in certain situations.

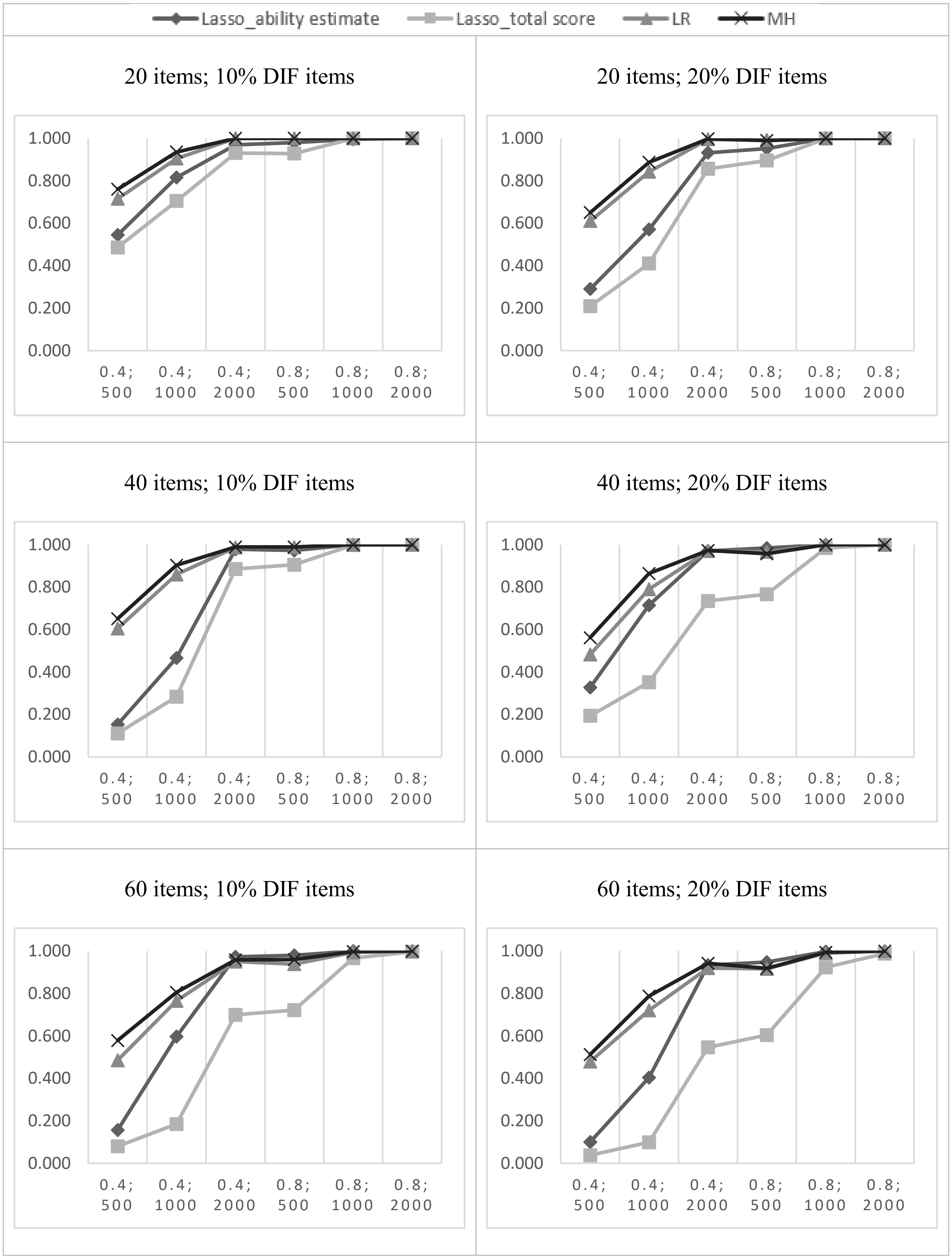

Figure 1. Power

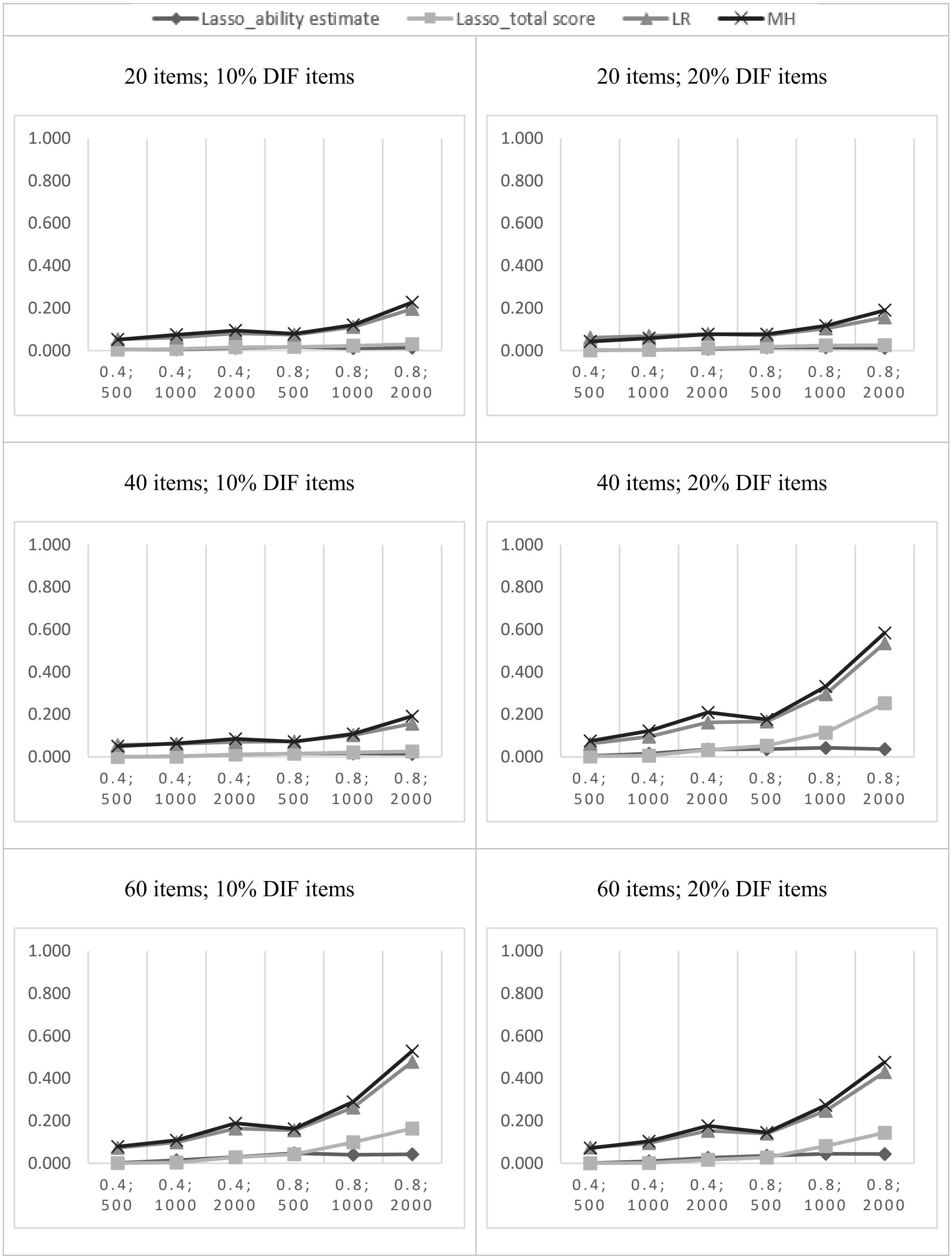

Figure 2. Type I Error Rate

5. Conclusions

The purpose of the study is to present a regularized LR method to DIF which allows simultaneous estimation of all DIF parameters in a single model. The simulation results are encouraging since the proposed method is able to detect DIF items accurately in most situations especially when the DIF magnitude and the sample size are large. At the same time, it has an outstanding performance in controlling for Type I error rates compared to other DIF detection methods which prevent over identification of DIF. Another advantage of the regularized LR method is that it is very flexible since it can be easily generalized to multiple-group comparison with more than two groups or even continuous covariates such as age by modifying the parameter representing group membership in the model. Future research is desired to examine the performance of this method in detecting non-uniform DIF and to investigate its applicability in DIF detection for polytomous items.

References

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

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, NJ: Erlbaum.

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

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111-135.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of educational statistics*, *7*(2), 105-118.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds), *Differential Item Functioning: Theory and Practice* (pp. 255-276). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

R Development Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project.org*.

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

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.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376.

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

Wang, W. C., & 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.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42-57.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.