Vol. 2 (1), pp. 22-39, Jan, 2015
http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

A Comparison of Logistic Regression Models for Dif Detection in Polytomous Items: The Effect of Small Sample Sizes and Non-Normality of Ability Distributions

Yasemin KAYA[1], **Walter L. LEITE**, **M. David MILLER**

Research and Evaluation Methodology Program, School of Human Development and Organizational Studies in Education, University of Florida

Abstract

This study investigated the effectiveness of logistic regression models to detect uniform and non-uniform DIF in polytomous items across small sample sizes and non-normality of ability distributions. A simulation study was used to compare three logistic regression models, which were the cumulative logits model, the continuation ratio model, and the adjacent categories model. The results revealed that logistic regression was a powerful method to detect DIF in polytomous items, but not useful to distinguish the type of DIF. Continuation ratio model worked best to detect uniform DIF, but the cumulative logits model gave more acceptable type I error results. As sample size increased, type I errors increased at cumulative logits model results. Skewness of ability distributions reduced power of logistic regression to detect non-uniform DIF. Small sample sizes reduced power of logistic regression.

Article Info

Received
06 January 2015

Accepted:
19 January 2015

Keywords: DIF, logistic regression, polytomous items, non-normality, uniform, non-uniform

1. Introduction

The logistic regression method (Swaminathan & Rogers,1990) has become a widely used method to detect differential item functioning (DIF) during the last two decades (e.g., Scott, Fayers, Aaronson, Bottomley, DeGraeff, Groenvold, Koller, Peterson & Sprangers, 2007 ; Hauger & Sireci, 2008 ; Crane, Hart, Gibbons & Cook, 2006 ; Gelin & Zumbo,2003). Its capability to detect both uniform and non-uniform DIF in either dichotomous or polytomous items (Zumbo, 1999) makes logistic regression a powerful method for investigating DIF. Unlike the Mantel-Haenszel (MH) method, logistic regression allows the ability variable to take continuous values, and to interact with the group variable (Swaminathan & Rogers, 1990). Furthermore, because logistic regression is a model based approach to study DIF, it is possible to include additional covariates in the model to help explain the mechanism behind DIF (Swaminathan & Rogers, 1990).

DIF detection methods with logistic regression for polytomous items are extensions of the logistic regression model for dichotomous items. Different extensions have been proposed by several researchers (Miller & Spray, 1993; Welch & Hoover, 1993; French &Miller, 1996; Zumbo, 1999), which are all based on three logistic models for polytomous items described by Agresti (2002): the cumulative logits model, the adjacent categories model, and the continuation ratio model. Each model uses a different coding procedure to compare pairs of

[1] Corresponding Author Email: yaseminkaya@ufl.edu

response categories. The cumulative logits model reformulates the response categories into two categories from 1 to j and from j+1 to J, where J is the total number of categories. In the adjacent categories model, the probability of each response category is compared with the probability of the nearby response category. With the continuation ratio model, the probability of each response category is compared with the probability of all higher response categories.

French and Miller (1996) were the first to use the cumulative logits, adjacent categories, and continuation ratio logistic regression models as DIF detection methods for polytomous items and to compare these models with a simulation study. They concluded that cumulative logits and continuation ratio models provided higher power to detect DIF than the adjacent categories model. The primary purpose of our study is to expand on French and Miller's research by investigating the feasibility of using the three logistic regression models for DIF detection in polytomous items in the case of small sample sizes and non-normal ability distributions.

Unlike IRT-based DIF detection methods (Swaminathan & Gifford, 1985), logistic regression models do not assume normality of group ability distributions. Therefore, researchers have not paid much attention to whether non-normality of ability distributions has any effect on the power of logistic regression to detect uniform or non-uniform DIF. Skewness of group ability distributions in DIF detection is an important factor that needs to be considered, because in educational and psychological research it is possible to meet with non-normal distribution cases. Knowledge about how skewness of ability distributions affects power may assist applied researchers in selecting the most appropriate DIF detection method given their available sample size. Furthermore, skewness of ability distributions may have an effect on the Type I error rates of logistic regression.

DIF detection in polytomous items with small samples is another condition requiring further study. Previous research on the effect of small sample sizes for detecting DIF in polytomous items showed different results for different DIF detection methods. Bolt (2002) compared parametric (i.e., Graded Response Model- Likelihood Ratio test, and Graded Response Model- Differential Functioning of Items and Tests) and nonparametric (i.e., Poly-SIBTEST) DIF detection methods in polytomous items. Her study demonstrated that when sample size is small (i.e., 300 per group), the parametric methods were more powerful to detect DIF than poly-SIBTEST. Also, type I error rate increased in small sample sizes for Poly-SIBTEST. The study concluded that the parametric methods might be more desirable than non parametric methods for small sample sizes. Furthermore, Chang and his colleagues (1996) found that changes in sample sizes had a positive impact on the Type I error rates of MH and Standardized mean difference (SMD) methods, but they found no effect on the Type I error rate of Poly-SIBTEST. Spray and Miller (1994) indicated that Logistic Discriminant Function Analysis procedure was affected by small sample size in their study. In order to obtain an adequate level of power, the minimum sample size recommend in DIF studies with either dichotomous or polytomous is approximately 250 per group (Swaminathan & Rogers, 1990; Scott, Fayers, Aaronson, Bottomley, Graeff, Groenvold, Gundy, Koller, Petersen & Sprangers, 2009; Zumbo, 1999). None of the previous studies addressed the effect of small sample sizes on DIF detection in polytomous items with logistic regression models. The present study investigated the feasibility of logistic regression for polytomous items for small samples by simulating sample sizes of 100 and 250 per group.

The research questions addressed in this study were: Do the cumulative logits, continuation ratio and adjacent categories models differ with respect to power to detect DIF in polytomous items? Do these models differ with respect to Type I error when used to detect DIF in polytomous items? Is the performance of logistic regression models for polytomous items affected by non-normality of the ability distribution? How are logistic regression models for polytomous items affected by small sample sizes? These questions were addressed through a Monte Carlo simulation study.

The logistic regression method of DIF detection

Using logistic regression for DIF detection in dichotomous items was proposed by Swaminathan and Rogers (1990). Based on the standard logistic regression model (Agresti, 2002), the authors used as the following model to detect DIF:

$$P(u=1|\theta,g)=\frac{e^{z}}{1+e^{z}} \tag{1}$$

where

$$z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g) \tag{2}$$

In this model, θ is the ability of an individual from a particular group, *g* is the group membership defined as 0 for the reference group and 1 for the focal group, and the p θg is the product of the two variables. The respondent's total scores on the scale is commonly used to represent the ability level θ. The regression coefficient τ_2 is the group difference in performance on the item, and τ_3 is the interaction term between group membership and ability. Swaminathan and Rogers (1990) indicated that if $\tau_2 \neq 0$ and $\tau_3 = 0$, the items shows uniform DIF. Furthermore, if $\tau_3 \neq 0$, the item shows non-uniform DIF, no matter whether $\tau_2 = 0$ or not.

Swaminathan and Rogers (1990) compared the logistic regression procedure with the MH procedure in terms of power by manipulating sample size, test length, and type of DIF. The results indicated that logistic regression was as powerful as the MH in detecting uniform DIF. Furthermore, logistic regression is able to detect non-uniform DIF, but the MH method is not.

The logistic regression DIF detection method for dichotomous items has been extended to polytomous items by several researchers (Miller & Spray, 1993; Welch & Hoover, 1993; French &Miller, 1996; Zumbo, 1999). French and Miller (1996) conducted a simulation study to examine the usefulness of different logistic regression models extended to polytomous data. Their study compared three extensions of the logistic modeling to polytomous data, which are the cumulative logits, adjacent categories, and continuation ratio logit models. The cumulative logits model is (Agresti,2002):

$$\text{logit}[P(Y\leq j|x)]=\log\frac{P(Y\leq j|x)}{1-P(Y\leq j|x)}=\log\frac{\pi_1(x)+......+\pi_j(x)}{\pi_{j+1}(x)+......+\pi_J(x)}, \tag{3}$$

with *j*=1,….., *J*-1, where *J* is the number of the response categories in an item. This model reformulates the multiple categories to two categories (i.e., 1 to *j*; *j*+1 to *J*), and compares the probability of a 1 to j response (i.e., the numerator) with a *j*+1 to *J* response (i.e., the denominator). This is a model that does not lose data during the dichotomization of the score categories.

The adjacent categories model is defined as (Agresti, 2002):

$$\text{logit}[P(Y=j|x)]=\log\frac{\pi_j}{\pi_{j+1}}, \tag{4}$$

where j=1,……, J-1. In the adjacent categories model, the probability of every category is compared with the probability of the adjacent category (French and Miller, 1996).

The continuation ratio model is described as (Agresti, 2002):

$$\log \frac{\pi_j}{\pi_{j+1} + \ldots\ldots + \pi_J}, \quad j = 1, \ldots\ldots, J\text{-}1. \tag{5}$$

or as

$$\log \frac{\pi_{j+1}}{\pi_1 + \ldots\ldots + \pi_j}, \quad j = 1, \ldots\ldots. J\text{-}1. \tag{6}$$

In this model, the probability of a response category is compared with the probability of the categories either above or below it.

In their study, French and Miller (1996) evaluated the power of logistic regression DIF detection method in polytomously scored items across the three logistic regression models described above. A 25-item test was simulated with a single item including DIF; and every item had four potential score categories ranging from 0 to 3. The authors created four conditions with respect to location of DIF: In three conditions, they varied discrimination parameters across groups to create non-uniform DIF, and in one condition they varied the difficulty parameters to create uniform DIF. They also simulated two levels of sample size, 500 and 2,000. Their results can be summarized into three major points: Firstly, as expected, sample size had a substantial effect on power, which was higher with the largest sample size (i.e., 2,000). Second, the magnitude of the difference in item parameters between two groups had an impact on the likelihood of detection of uniform and non-uniform DIF. Finally, the cumulative logits and continuation ratio models had high power to detect uniform DIF and non-uniform DIF. The adjacent categories model had the lowest power, due to loss of data when comparing each pair of categories separately from the other categories. On the other hand, the study stated that the adjacent categories model might be useful to determine the location of DIF.

Zumbo (1999) proposed a DIF detection method that combines the Ordinal Logistic regression (OLR) method, which works with cumulative logits model, and an R^2 measure of effect size in order to detect DIF and determine its magnitude in polytomous items. Zumbo (1999) proposed a DIF effect size measure for ordinal items based on the R^2. In three steps, he showed the application of the method on two items from a simulated 20-item ordinal test data for gender-based DIF. In the first step of modeling, only the conditioning variable (i.e. the total score) was entered in the model. In the second step, the group variable (i.e., gender) was added to the model. In the third and the last model, the interaction term between gender and total score was added to the model. DIF tests were performed by using the chi-square values obtained from each step of the analysis: The difference between the chi-square statistic from step 1 and the chi-square statistic from step 3 gave the simultaneous statistical test of uniform and non-uniform DIF. The same procedure was repeated between step 1 and step 2 to identify of uniform DIF, and between step 2 and step 3 to identify non-uniform DIF. The estimation of the effect size of DIF was obtained by taking the differences in R^2 between step 3 and step 1, step 2 and step 1, and step 3 and step 2. R^2values at or above .002 were considered to show meaningful levels of DIF.

The effect of skewed ability distributions on DIF detection methods

Kristjansson, Aylesworth, McDowell, and Zumbo (2005) evaluated the effect of skewness on the performance of four DIF detection methods for polytomous items, which are the MH, generalized MH, logistic discriminant function analysis, and unconstrained cumulative logits ordinal logistic regression. These authors simulated conditions where the skewness levels where either -.75 (moderate negative) for both groups and zero. Their study's results indicated that the skewness level simulated did not have a notable effect on the efficiency of the four methods examined.

Welkenhuysen-Gybels (2004) examined the performance of different DIF detection methods for dichotomously scored items by varying group ability distribution. To test the robustness of the logistic regression method, the author simulated three different ability distribution conditions: (1) a normal distribution for both groups, (2) a normal distribution for the reference group and a positively skewed distribution for focal group (with a beta distribution of 1.5, 5), and (3) a normal distribution for the reference group and a negatively skewed distribution for the focal group (with a beta distribution of 5, 1.5). In the case of uniform DIF, the logistic regression method results indicated that both the false positive rate and the false negative rate for the normal / positively skewed ability distribution condition had a higher value than for the normal / normal ability condition. On the other hand, the false negative rate for the normal / negatively skewed ability distribution condition was lower than for the normal / normal ability distribution condition. In the case of nonuniform DIF, a skewed distribution always produced higher false positive rate than the normal distribution. However, the false negative rate of any skewed distribution condition was not significantly different than the normal distribution condition.

The majority of simulation studies about logistic regression have been done by simulating a normal ability distribution. As we mentioned above, there are only a few studies for all DIF detection methods with skewed ability level conditions. However, real samples are not always normally distributed, as shown in some application studies from medical and educational fields (e.g., Wang & Lane, 1996). Therefore, there have not been enough simulation studies examining the robustness of DIF detection methods to skewness of ability distributions. In particular, there have been no studies about the effect of skewed ability distribution on the performance of logistic regression models for DIF detection in polytomously scored.

The effect of small sample sizes

Many studies have been published about how effective DIF methods work with different sample sizes. In particular, studies about logistic regression for DIF detection in dichotomous items have addressed sample sizes as small as 250. More specifically, Swaminathan and Rogers (1990), comparing MH and Logistic Regression techniques for dichotomously scored items, found that for samples of 250, the logistic regression method resulted in 75% correct detection of uniform DIF and 50 % correct detection of nonuniform DIF. With 500 respondents per group, the logistic regression procedure resulted in 100 % accurate uniform DIF detection and 75% accurate nonuniform DIF detection. Rogers and Swaminathan (1993) also found a strong effect of sample size on the DIF detection rate of logistic regression. Their results showed a 19% increase on the DIF detection rate of logistic regression as the sample size increased from 250 to 500.

A Monte Carlo simulation study performed by Herrera and Gomez (2008) detected the effect of the different sample sizes of reference and focal groups on the power and type I error of logistic regression. Their study manipulated 12 conditions with two different sample sizes for the reference group (i.e., 500, 1,500) and six different ratios of sample sizes between the focal group and the reference group (i.e., 1/5, 1/4, 1/3, 2/5, 1/2, 1/1). Error mean squares were used as the accuracy index. Surprisingly, the results of their study indicated that the highest type I error rate of the logistic regression procedure resulted from the condition with a sample size of 1,500 (i.e., the largest sample size) with equal group sizes.

Welkenhuysen-Gybels (2004) examined the effect of sample size on logistic regression for detection of uniform and non-uniform DIF. They examined conditions with large and equal sample size for the focal and reference groups (i.e., 1,000), as well as a smaller sample size for the focal group (i.e., 300) than the reference group (i.e., 1,000). In the case of uniform DIF, the results for logistic regression indicated that when the sample size decreased for the focal group, the false positive rate decreased but the false negative rate increased. For nonuniform DIF, the logistic regression model produced an increase on the false negative rate as the sample size decreased.

Scott et al. (2009) studied the effect of sample size on DIF detection in polytomous items with ordinal logistic regression. They found that for a power level of 80% or higher, ordinal logistic regression requires at least 200 observations per group. For a p-value of .001 instead of .05, these authors suggested a minimum sample size of 500per group. Based on the findings in the literature, Zumbo (1999) suggested at least 200 observations per group to achieve adequate levels of power for DIF detection with ordinal logistic regression.

2. Method

The logistic regression procedure was used in this study to conduct differential item functioning (DIF) analyses in polytomously scored items. A Monte Carlo simulation study was conducted to compare the performances of three different logistic regression models for DIF detection in polytomous items: the continuation ratio logits model, the cumulative logits model, and the adjacent categories model. The performances of the three ordinal logistic regression models for DIF detection were compared in terms of the power and type I error rates. Our study’s design is a partial replication and extension of the study by French and Miller (1996).

We simulated scores for a test with 25 items of four score categories under the graded response model of Samejima (1969; 1996), which is an appropriate model for items with ordered polytomous responses. Similar to French and Miller's (1996) study, only the 25th item in each test condition had uniform or non-uniform DIF and was used to investigate the power of the logistic regression model for DIF detection. Therefore, items 1 to 24 did not contain DIF, and were used to estimate Type I error rates.

Factors Manipulated

We simulated DIF in the 25th item according to four conditions reflecting different combinations of type of DIF (i.e., uniform or non-uniform) and magnitude of DIF (i.e., the differences between item parameters). The first three conditions contained nonuniform DIF of increasing magnitude, and the last condition contained uniform DIF. These conditions will be detailed in the data generation section.

In all conditions, the sample sizes for the reference group and the focal group were equal. Three levels of sample size were investigated. The sample size of 500 for each group was used to represent a medium sample size. Small sample sizes were chosen based on the findings in the literature, which recommends that 200 per group is the smallest sample size for the logistic regression procedure (Zumbo, 1999; Scott et.al. 2009). For the two conditions with small sample size, we defined the number of examinees in each group as 100 and 250.

The sample size of 250 was a common sample size in simulation studies investigating DIF with small samples. Furthermore, we investigated the feasibility of the sample size of 100.

The effect of skewness of ability distributions has not been thoroughly studied in previous research. To examine the effect of skewness on power to detect DIF in polytomous items, we simulated three levels of skewness of ability distributions: normal, moderate negative skewness, and high negative skewness. Normally-distributed ability values with a mean of zero and standard deviation of one were simulated by using the *rnorm* function from the R statistical software (R development core team, 2010). Fleishman's (1978) power transformation method was used for simulating skewed distributions. The levels of skewness and kurtosis were simulated by using the coefficient values on Fleishman's Power Method Weights table. Moderate negative skew on ability distribution was studied by Kristjansson et al. (2005) at the level of -.75 and it showed a slight effect on the performance of DIF detection methods. In the present study, we set the skewness levels of ability distributions as -.75 for moderate skewness, and -1.75 for high skewness. The kurtosis level of the ability distributions was fixed at 3.75.

Data Generation

The item parameters used by French and Miller (1996) to generate the data were also used in this study. In all conditions, the item parameters remained the same for the first 24 items, which are the items with no DIF. For the first 24 items, five different values of item discrimination parameter (*a*) were simulated. The values were .50 for questions 1 to 5; .75 for questions 6 to 10; 1.00 for questions 11 to 15; 1.25 for questions 16 to 20; and 1.50 for questions 21 to 24. Since the value of the first score category was 0, French and Miller (1996) set the threshold parameters to 0 for the first score category (b_1). Because in a four-category item only three thresholds are needed, we did not use b1 parameters in the simulation. The parameters b2, b3 and b4 were different for each item, and they were in increasing order within every item.

Four different DIF conditions were simulated for the 25th item. From condition 1 to condition 3, the difference between the item discrimination parameters of the focal and the reference groups were increased by .5 in each condition, but item threshold parameters were held constant at in order to create non-uniform DIF. In condition 4, item discrimination parameters of the focal and the reference groups were held constant at the same level, which was 1.0, but the difference between the item threshold parameters were increased by 1.0 for b_2 and b_3 in order to create uniform DIF. b_4 remained constant at 2.0.

The simulation design consisted of four DIF conditions (i.e., three uniform and one non-uniform DIF), three sample size conditions (i.e.,100, 250, 500), and three level of skewness (i.e., normal, moderately skewed and highly skewed), and three logistic regression models (i.e., cumulative logits, continuation ratio, adjacent categories). For each condition, 1,000 datasets were generated. Data was simulated and analyzed with the R 2.10.1 statistical software (R Development Core Team, 2010).

Data Analysis

We used the VGAM package for categorical data analysis (Yee, 2010) of the R statistical software to fit the logistic regression models. We evaluated DIF with each ordinal logistic regression model using the three steps recommended by Zumbo (1999) mentioned earlier. These steps require fitting each logistic regression model three times to each dataset: a first model with only the total score as a predictor of item response; a second model with total score and group variables as predictors; a third model with total score, group variable, and the interaction term as predictors. DIF was evaluated by testing the differences in deviance chi square values between these models given their difference in degrees of freedom. Three different deviance chi-square tests were performed. To test the existence of any kind of DIF, we subtracted the deviance chi square value of third model from the first model. To test uniform DIF, the deviance value of the second model was subtracted from the first model. And finally, to test the non-uniform DIF, the deviance chi-square value of the third model was subtracted from the second model.

The VGAM package for categorical data analysis (Yee, 2010) of the R statistical software is able to fit the logistic regression models for ordinal responses with or without the parallelism (or proportionality) assumption. The parallelism assumption states that logistic regression models for different categories have equal slopes for a given predictor (Armstrong & Sloan, 1989; Cole & Ananth, 2001). We compared the model fit between the ordinal logistic regression models with or without the parallelism assumption. We found that the parallelism assumption only held with the model where the single predictor was the total score. Therefore, we decided to fit the ordinal logistic regression models in the study without making the parallelism assumption.

After the data generation and analysis were completed, the p-values of each logistic regression models were collected. Power was estimated by calculating the proportion of iterations DIF was correctly detected in the 25th item. The Type I error was calculated by computing the proportion of iterations DIF was falsely detected in items 1 to 24. Similarly to French and Miller's (1996) study, the alpha level of .05 was divided by the number of items,25, to control the family-wise Type I error rate, resulting in an alpha level of .002 per test. The results of each model were compared across all the conditions to determine the best performing model.

3. Results

The results of the study are composed of the power values for the 25th item and Type I error values for 24 non-DIF items in each condition. It was expected that the power levels would increase from condition one through condition three because increasing differences between item discrimination parameters would make the shape of item characteristic curves (ICCs) more distinguishable (Embretson & Reise, 2000). It was also expected that increases in sample size would result in higher power for DIF detection. Finally, skewness of ability distributions was expected to reduce the power for detecting DIF. We present the results of this study in Tables 1 to 6. Overall, the results show that most of the expectations were met.

Table 1 contains power values of the three ordinal logistic regression methods to flag the presence of any kind of DIF in item 25 across all the levels of sample sizes, ability distributions, and magnitude of DIF. In condition 1, which had the smallest difference between item parameters, none of the models with sample size of 100 provided sufficient power to detect DIF due to a 0.5 difference between discriminations. However, sufficient power values were provided in the groups with sample size of 250 and high power values with sample size of 500. We found that three models slightly differed in their performance to detect any kind of DIF in condition 1. With the smallest level of DIF, the cumulative logits model provided slightly lower power than other two models with sample size of 100 in condition 1 and 2. Models did not differ in conditions 3 and 4. In condition 2, power values were sufficient for sample size of 100 and perfect power values were observed for sample sizes of 250 and 500. Conditions 3 and 4 provided almost perfect power results for all models with all sample sizes. The results gathered from conditions 2 to 4 showed that a small sample size of 100 worked well in the case of conditions with large difference between the item parameters of the focal and the reference groups. The continuation ratio and adjacent categories models performed similarly in all the conditions. Power values did not differ substantially between normal and skewed distributions.

Table 1. Power of ordinal logistic regression models to detect any kind of DIF in item 25

Ability Distribution	Model	Sample size	Condition 1	Condition 2	Condition 3	Condition 4
Normal Distribution	Cumulative logits	100	0.059	0.798	0.978	0.852
		250	0.789	0.999	0.993	0.988
		500	0.996	1	1	0.996
	Continuation ratio	100	0.205	0.873	0.997	0.936
		250	0.812	1	1	1
		500	0.997	1	1	1
	Adjacent categories	100	0.205	0.869	0.995	0.939
		250	0.795	1	1	1
		500	0.997	1	1	1
Moderately Skewed	Cumulative logits	100	0.144	0.789	0.985	0.622
		250	0.768	1	0.999	0.641
		500	0.996	0.999	0.997	0.51
	Continuation ratio	100	0.203	0.876	0.997	0.961
		250	0.801	1	1	1
		500	0.996	1	1	1
	Adjacent categories	100	0.209	0.883	0.995	0.96
		250	0.814	1	1	1
		500	0.996	1	1	1
Highly Skewed	Cumulative logits	100	0.158	0.84	0.992	0.93
		250	0.795	1	0.999	0.996
		500	1	1	0.991	0.999
	Continuation ratio	100	0.229	0.897	0.999	0.962
		250	0.815	1	1	1
		500	1	1	1	1
	Adjacent categories	100	0.227	0.902	0.999	0.962
		250	0.815	1	1	1
		500	1	1	1	1

For conditions 1 to 3, Table 2 illustrates power values of three logistic regression methods to flag the presence of non-uniform DIF in item 25 across all the levels of sample size, ability distribution, and magnitude of DIF. Because in condition 4 the 25th item had uniform DIF, the values in the condition 4 column of table 2 are type I error rates for detecting non-uniform DIF when uniform DIF exists. Results in Table 2 showed that power rates increased from condition 1 through condition 3, as expected. We found that none of the logistic regression models had sufficient power to detect non-uniform DIF with a sample size of 100 or 250 in condition 1. Only the continuation ratio model provided acceptable power values with a sample size of 500 for groups with normal and moderately skewed ability distributions in condition 1. For groups with normal and moderately skewed ability distributions, sample sizes of 250 and 500 provided sufficient power for all the models in condition 2. For groups with highly skewed distribution and 500 sample size, only continuation ratio model had sufficient power in condition 1. In condition 3, for groups with normal and moderately skewed ability distributions, all the models worked well to detect non-uniform DIF in the sample sizes of 250 and 500. The continuation ratio model provided sufficient power in sample size of 100. For groups with highly skewed distribution, the cumulative logits and adjacent categories models had sufficient power values in sample size of 500. However, the continuation ratio model provided sufficient power values in sample sizes of 250 and 500. None of the models worked well in sample size of 100 for groups with highly skewed ability distribution in condition 3. Skewness of ability distribution effected power of the models for non-uniform DIF detection to some extent. As skewness increased, the power values decreased gradually from normal to highly skewed ability distribution. In general, power levels for DIF detection for groups with highly skewed ability distributions were not at an adequate level except for condition 3 with a sample size of 500. The ordinal logistic regression models performed differently than what was expected based on French and Miller's (1996) results. In most of the cases, the continuation ratio model performed best in non-uniform DIF detection. Moreover, in some condition settings, the continuation ratio model was the only model that had the sufficient power to detect non-uniform DIF. On the other hand, the cumulative logits and the adjacent categories models performed similarly. In condition 4, we could not find any difference among the power values.

Table 2. Power of ordinal logistic regression models to detect non-uniform DIF in item 25

Ability Distribution	Model	Sample size	Condition 1	Condition 2	Condition 3	Condition 4
Normal Distribution	Cumulative logits	100	0.017	0.152	0.491	0.002
		250	0.088	0.659	0.964	0
		500	0.339	0.976	1	0.005
	Continuation ratio	100	0.059	0.441	0.865	0.005
		250	0.254	0.961	1	0.007
		500	0.694	1	1	0.03
	Adjacent categories	100	0.031	0.203	0.564	0.002
		250	0.112	0.742	1	0.003
		500	0.384	0.987	1	0.014
Moderately Skewed	Cumulative logits	100	0.034	0.189	0.524	0.128
		250	0.077	0.611	0.968	0.218
		500	0.18	0.95	0.997	0.338
	Continuation ratio	100	0.035	0.272	0.657	0.007
		250	0.18	0.864	0.997	0.012
		500	0.502	1	1	0.016
	Adjacent categories	100	0.023	0.148	0.431	0.003
		250	0.096	0.633	0.97	0.01
		500	0.267	0.977	1	0.006
Highly Skewed	Cumulative logits	100	0.006	0.029	0.159	0.006
		250	0.016	0.08	0.435	0.001
		500	0.028	0.318	0.823	0.002
	Continuation ratio	100	0.017	0.085	0.292	0.003
		250	0.046	0.39	0.861	0.006
		500	0.134	0.857	0.998	0.004
	Adjacent categories	100	0.007	0.035	0.12	0.003
		250	0.017	0.126	0.491	0.003
		500	0.042	0.461	0.929	0.001

Table 3 illustrates percentages of flagging uniform DIF in item 25 for three logistic regression methods across all the levels of sample sizes and ability distributions. Item parameters were chosen to create non-uniform DIF in conditions 1 to 3, and uniform DIF in condition 4. Thus, the values in the condition 4 column of the table are power levels, but the values in the columns for conditions 1, 2, and 3 are Type I error rates for detecting uniform DIF when non-uniform DIF exists. Although conditions 1 to 3 were simulated to present non-uniform DIF, the results in Table 3 indicate that all of the logistic regression models in conditions 2 and 3 detected uniform DIF in item 25 in most of the iterations. This extreme inflation of type I error rates leads to the conclusion that the ordinal logistic regression models are of limited use in distinguishing between uniform and non-uniform DIF. Only condition 4 was designed to create a uniform DIF item. Study results in Table 3 show that all three logistic regression models were powerful to detect uniform DIF in item 25 for all three sample size with normal or highly skewed distributions. However, we detected an unexpected behavior of the cumulative logistic model with the moderately skewed distribution. In condition 4, the power values of cumulative logits model for moderately skewed groups were lower than with normal or highly skewed distributions. These results were accompanied by a lower inflation of Type I error rates for the cumulative logits model in conditions 3.

Table 3. Power of ordinal logistic regression models to detect uniform DIF in item 25

Ability Distribution	Model	Sample Size	Condition 1	Condition 2	Condition 3	Condition 4
Normal Distribution	Cumulative logits	100	0.158	0.72	0.92	0.958
		250	0.748	1	0.997	1
		500	0.997	1	1	1
	Continuation ratio	100	0.146	0.646	0.88	0.972
		250	0.642	0.992	0.999	1
		500	0.974	1	1	1
	Adjacent categories	100	0.205	0.797	0.96	0.976
		250	0.76	0.999	1	1
		500	0.995	1	1	1
Moderately Skewed	Cumulative logits	100	0.172	0.661	0.758	0.768
		250	0.767	0.902	0.653	0.73
		500	0.993	0.905	0.463	0.58
	Continuation ratio	100	0.168	0.753	0.957	0.986
		250	0.709	1	1	1
		500	0.988	1	1	1
	Adjacent categories	100	0.217	0.842	0.98	0.986
		250	0.796	1	1	1
		500	0.994	1	1	1
Highly Skewed	Cumulative logits	100	0.226	0.87	0.936	0.986
		250	0.878	0.996	0.937	1
		500	0.999	0.999	0.946	1
	Continuation ratio	100	0.259	0.889	0.996	0.99
		250	0.83	1	1	1
		500	0.996	1	1	1
	Adjacent categories	100	0.286	0.925	0.998	0.991
		250	0.874	1	1	1
		500	0.998	1	1	1

Tables 4 to 6 contain Type I error rates for the 24 DIF-free items. In these tables, the Type I error rate is considered adequate if it is under 0.002. Table 4 illustrates Type I error results for the test of any kind of DIF detection, no matter which type of DIF exists, with all the logistic regression models, sample sizes, and ability distribution levels. Table 5 presents Type I error rates for the test of uniform DIF detection with all the conditions. Table 6 shows Type I error rates for the test of non-uniform DIF detection in all conditions. At the alpha level of 0.002, all three tables showed common features: Results indicated that only the cumulative logits model had Type I error rates at acceptable levels. However, it was observed that as sample size increased, the type I error rates of the cumulative logits model increased as well. Type I error rates with the continuation ratio and the adjacent categories models were mostly higher than the alpha level. No clear effect of skewness or sample size appeared on the Type I error rates with the continuation ratio and adjacent categories models.

Table 4. Type I error rates of ordinal logistic regression models for the test for any kind of DIF in items 1 to 24

Ability Distribution	Model	Sample size	Condition 1	Condition 2	Condition 3	Condition 4
Normal	Cumulative logits	100	0.0005	0.0009	0.0002	0.0004
		250	0.0009	0.0015	0.0012	0.0014
		500	0.0017	0.0021	0.0019	0.0022
	Continuation ratio	100	0.0034	0.0032	0.0026	0.0030
		250	0.0020	0.0026	0.0025	0.0021
		500	0.0024	0.0023	0.0021	0.0026
	Adjacent categories	100	0.0032	0.0032	0.0029	0.0031
		250	0.0017	0.0024	0.0025	0.0020
		500	0.0023	0.0023	0.0020	0.0024
Moderately Skewed	Cumulative logits	100	0.0009	0.0007	0	0.0003
		250	0.0009	0.0011	0.0010	0.0007
		500	0.0015	0.0010	0.0010	0.0008
	Continuation ratio	100	0.0028	0.0034	0.0028	0.0030
		250	0.0024	0.0022	0.0020	0.0021
		500	0.0021	0.002	0.0020	0.0015
	Adjacent categories	100	0.0030	0.0036	0.0030	0.0030
		250	0.0024	0.0024	0.0020	0.0023
		500	0.0025	0.0020	0.0019	0.0017
Highly Skewed	Cumulative logits	100	0.0003	0.0095	0.0009	0.0011
		250	0.0014	0.0015	0.0012	0.0010
		500	0.0016	0.0017	0.0014	0.0013
	Continuation ratio	100	0.0027	0.0030	0.0023	0.0028
		250	0.0029	0.0022	0.0021	0.0019
		500	0.0017	0.0019	0.0020	0.0021
	Adjacent categories	100	0.0027	0.0028	0.0026	0.0031
		250	0.0027	0.0023	0.0019	0.0019
		500	0.0018	0.0019	0.0019	0.0020

Table 5. Type I error rates for the tests of uniform DIF detection for items 1 to 24

Ability Distribution	Model	Sample size	Condition 1	Condition 2	Condition 3	Condition4
Normal Distribution	Cumulative logits	100	0.0011	0.0010	0.0013	0.0010
		250	0.0017	0.0013	0.0017	0.0020
		500	0.0018	0.0025	0.0023	0.0016
	Continuation ratio	100	0.0026	0.0021	0.0025	0.0026
		250	0.0019	0.0018	0.0020	0.0022
		500	0.0020	0.0025	0.0020	0.0018
	Adjacent categories	100	0.0025	0.0023	0.0029	0.0024
		250	0.0020	0.0020	0.0019	0.0021
		500	0.0022	0.0027	0.0021	0.0015
Moderately Skewed	Cumulative logits	100	0.0026	0.0009	0.0013	0.0010
		250	0.0010	0.0016	0.0018	0.0017
		500	0.0014	0.0020	0.0019	0.0013
	Continuation ratio	100	0.0026	0.0028	0.0020	0.0022
		250	0.0017	0.0020	0.0018	0.0022
		500	0.0019	0.0020	0.0021	0.0015
	Adjacent categories	100	0.0027	0.0030	0.0027	0.0024
		250	0.0017	0.0019	0.0021	0.0020
		500	0.0017	0.0020	0.0023	0.0018
Highly Skewed	Cumulative logits	100	0.0015	0.0013	0.0014	0.0017
		250	0.0018	0.0026	0.0014	0.0015
		500	0.0019	0.0019	0.0014	0.0017
	Continuation ratio	100	0.0028	0.0027	0.0026	0.0025
		250	0.0025	0.0029	0.0020	0.0020
		500	0.0018	0.0019	0.0080	0.0020
	Adjacent categories	100	0.0028	0.0027	0.0025	0.0028
		250	0.0023	0.0028	0.0020	0.0022
		500	0.0019	0.0020	0.0013	0.0019

Table 6. Type I error rates for the tests of non-uniform DIF detection for items 1 to 24

Ability Distribution	Model	Sample size	Condition 1	Condition 2	Condition 3	Condition 4
Normal Distribution	Cumulative logits	100	0.0005	0.0007	0.0008	0.0002
		250	0.0010	0.0010	0.0010	0.0010
		500	0.0019	0.0017	0.0015	0.0019
	Continuation ratio	100	0.0036	0.0035	0.0025	0.0023
		250	0.0023	0.0027	0.0029	0.0020
		500	0.0028	0.0025	0.0019	0.0030
	Adjacent categories	100	0.0035	0.0028	0.0023	0.0024
		250	0.0018	0.0022	0.0024	0.0018
		500	0.0023	0.0023	0.0019	0.0024
Moderately Skewed	Cumulative logits	100	0.0009	0.0007	0.0001	0.0002
		250	0.0009	0.0007	0.0004	0.0004
		500	0.0015	0.0006	0.0010	0.0010
	Continuation ratio	100	0.0031	0.0034	0.0027	0.0035
		250	0.0018	0.0028	0.0020	0.0020
		500	0.0020	0.0020	0.0023	0.0024
	Adjacent categories	100	0.0035	0.0036	0.0030	0.0033
		250	0.0020	0.0026	0.0019	0.0020
		500	0.0022	0.0020	0.0020	0.0024
Highly Skewed	Cumulative logits	100	0.0005	0.0009	0.0013	0.0007
		250	0.0012	0.0012	0.0016	0.0010
		500	0.0017	0.0018	0.0017	0.0014
	Continuation ratio	100	0.0028	0.0036	0.0033	0.0033
		250	0.0022	0.0023	0.0025	0.0022
		500	0.0019	0.0021	0.0023	0.0021
	Adjacent categories	100	0.0026	0.0031	0.0025	0.0029
		250	0.0022	0.0017	0.0023	0.0020
		500	0.0021	0.0020	0.0020	0.0018

4. Discussion

In summary, we detected that the cumulative logits model had the lowest power among the models in the test of any kind of DIF. Since we obtained almost perfect power when the magnitude of DIF increased, we could not detect any difference for the test of any kind of DIF, no matter which type of DIF exists. The three logistic regression models differed on their performances when the test for non-uniform DIF was applied as well. For a non-uniform DIF test, the continuation ratio model produced the most power to detect DIF, while the cumulative logits and the adjacent categories models did not differ. These results disagree with French and Miller's (1996) findings, in which cumulative logits and continuation ratio performed similarly, and the adjacent categories model had the lowest power compared to the other models.

As the magnitude of DIF increased, the power for detecting non-uniform DIF increased as well. In other words, the increase in the difference between item discrimination parameters made it easier to detect DIF in an item. This was an expected result, and confirmed French and Miller’s (1996) findings.

DIF detection with small sample sizes was less powerful than with moderate sample sizes which confirmed the previous findings (French &Miller, 1996). Non-uniform DIF detection in groups of 250 and 500 was powerful enough with logistic regression when the difference between item discrimination parameters was 1.00 or above. Furthermore, in testing the presence of any kind of DIF, logistic regression models were powerful enough for the groups with sample size of 100 when the difference between the item discrimination parameters for the focal and the reference groups was 1.00, or larger. For the test of non-uniform DIF, logistic regression did not have sufficient power with a smallest sample size of 100 even if the magnitude of DIF was large. For the case of uniform DIF, logistic regression models were powerful with small sample sizes (i.e., 100 and 250). Thus, as opposed to previous studies (Zumbo, 1999; Scott et.al. 2009), this study's results show that logistic regression has sufficient power to detect uniform DIF in datasets with a sample size of 100 per group. However, this result is tempered by the finding of extremely large Type I error rates for the test of uniform DIF when only non-uniform DIF exists. Therefore, we concluded that logistic regression models were not able to distinguish non-uniform DIF from uniform DIF in polytomous items, when the test of uniform DIF was applied to an item with non-uniform DIF. This inference confirms French and Miller's (1996) findings.

Logistic regression models differed on their Type I error rates. The cumulative logits model was the only model whose type I errors were all within the acceptable level. The Type I error rates with all models were found to increase as sample size increases, which disagrees with Herrera and Gomez's (2008) findings. The continuation ratio and the adjacent categories models mostly had high type I error rates given the alpha level of .002. Finally, skewness of ability distribution has an effect on the power of detecting non-uniform DIF with logistic regression, when the skewness level is as high as -1.75.

Based on the overall results, it can be concluded that the logistic regression method is a powerful method to detect DIF in polytomous items, but not useful to distinguish the type of DIF. Sample size is a factor that affects the power of the logistic regression to detect DIF. However, if the difference between item discrimination parameters is equal to or larger than 1.0, logistic regression provides sufficient power to detect DIF with small samples, such as 100 per group.

We found that the continuation ratio model is the most powerful logistic regression model to detect non-uniform DIF in polytomous items. The cumulative logits model has the lowest power among the models in the test of any kind of DIF. On the other hand, even though the cumulative logits model gives the lowest power among the models, it has the lowest type I error rates. The type I error rate of the cumulative logits model increases as the sample size increases, but it remains at acceptable levels.

5. Conclusion

French and Miller (1996) indicated that running separate regressions for each model was time consuming in logistic regression. However, recent improvements in statistical programs allow us to run all the separate regressions at the same time. Therefore, this is no longer a disadvantage for logistic regression in polytomous items. Previous research indicated that likelihood-ratio DIF detection test for polytomous items was not powerful for small sample sizes as small as 500 per group (Ankenmann, Witt & Dunbar, 1999). However, we found that logistic regression is powerful with sample size of 250, and even with a sample size of 100 in the case of large differences between item discrimination parameters. Nevertheless, the IRT-

LR allows direct omnibus tests of DIF hypotheses for all the item parameters, which is not possible with logistic regression method.

Generalized MH and logistic regression are both similarly powerful to detect uniform DIF in polytomous items (Kristjansson et. al, 2005). Logistic regression's capability to detect both uniform and non-uniform DIF makes it advantageous over MH, because the MH method is not able to detect non-uniform DIF. On the other hand, logistic regression with polytomous items is not able to distinguish nonuniform DIF from uniform DIF. Another disadvantage of logistic regression is that skewness of ability distributions can reduce the power of logistic regression for non-uniform DIF detection, but it does not affect the power of uniform DIF and any kind of DIF detection tests. Finally, small sample sizes reduce the power of most DIF methods, but logistic regression can attain sufficient power with small sample sizes if the difference between item parameters is large, even with a sample size as small as 100.

The necessity of dichotomization of polytomous response categories in order to compare the probabilities to answer a question as correct for different groups in logistic regression method causes the loss of some amount of data, which makes logistic regression less advantageous. Nevertheless, as French and Miller (1996) point out, the separate comparisons of score categories in the adjacent categories model helps us to identify the location of DIF in polytomous items, which is a unique feature of logistic regression in polytomous items.

The continuation ratio model is the most powerful logistic regression model to detect non-uniform DIF in polytomous items. However, high rates of type I error occur in all the test results with the continuation ratio model. The cumulative logits model is the only model that provides acceptable type I error rates in every condition. Hence, in non-uniform DIF detection, the continuation ratio model can be used due to its high power to detect non-uniform DIF. On the other hand, since the power of the models do not differ in the test of uniform DIF detection, the cumulative logits model is a more appropriate model for uniform DIF detection tests due to its low type I error rate. On the other hand, since the cumulative logits model performs worse than other two models in non-uniform DIF detection, using the continuation ratio or the adjacent categories models can be preferred to detect non-uniform DIF.

As with any Monte Carlo simulation study, this study had limitations due to our particular choice of conditions. One limitation of this study is that only a single item was simulated including DIF within a test of 25 items. With multiple items showing DIF, the total score may be contaminated and a purification process may be necessary. Another limitation is that we simulated only one level of uniform DIF, which meant that the change of power levels with the increase of uniform DIF magnitude was not examined. Moreover, DIF conditions were generated by changing only one of the item parameters, and fixing the other ones to a certain value. There was no condition simulated in which both *a* and *b* parameters were changing. Another limitation of this study is that we controlled the family-wise Type I error rate using a Bonferroni correction of the alpha level, which tends to result in conservative tests (Kim, 2010). An improvement in the method to control for the family-wise Type I error rate would be to use the Benjamini and Hochberg False Discovery Rate method (Benjamini & Hochberg, 1995).

We chose to focus on negatively skewed distributions, so the effect of positive skewness of ability distributions on DIF detection was not examined in this study. We also set the kurtosis to a fixed value. The combination of different kurtosis levels with different skewness levels might show an effect on power. Thus, future research should examine the effect of positively skewed ability distributions and changing kurtosis values as well. Finally, the focal and the reference groups were simulated with equal sample sizes, but unequal sample sizes for groups could produce differences between the three ordinal logistic regression models examined.

6. References

Agresti A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley.

Ankenmann R.D., Witt E.A. & Dunbar S.B., (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*, 277–300.

Armstrong B.G. & Sloan M., (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, *129*, 191–204.

Bock R.D., & Aitkin M., (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*, 443–459.

Bolt D.M., (2002). A monte carlo comparison of parametric and nonparametric polytomous DIF detection dethods. *Applied Measurement in Education*, *15*, 113–141.

Chang H.H., Mazzeo J. & Roussos L., (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333–353.

Cole S.R. & Ananth C.V., (2001). Regression models for unconstrained, partially or fully constrained continuation odds ratio. *International Journal of Epidemiology*, *30*, 1379–1382.

Crane P.K., Hart D.L., Gibbons L.E. & Cook K.F., (2006). A 37-item shoulder functional status item pool had negligible differential item functioning. *Journal of Clinical Epidemiology*, *59*, 478–484.

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.

Embretson S.E. & Reise S.P., (2000). *Psychometric methods: Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fleishman A.I., (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532.

French A.W., & Miller, T. R., (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315–332.

Gelin M.N. & Zumbo B.D., (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement*, *63*, 65–74.

Herrera A.N. & Gomez J., (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity*, *42*, 739–755

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

Kim J., (2010). Controlling Type 1 Error Rate in Evaluating Differential Item Functioning for Four DIF Methods: Use of Three Procedures for Adjustment of Multiple Item Testing. *Educational Policy Studies Dissertations*, 67.

Kristjansson E., Aylesworth R., McDowell I. & Zumbo B.D., (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*, 933–953.

Monaco M.K., (1997). A Monte Carlo assessment of skewed theta distributions on differential item functioning indices. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *58*(5-B), 2746.

R development Core Team, (2010).*R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rogers H. and Swaminathan H., (1993). A comparison of logistic regression and Mantel–Haenszel procedures for detecting differential item functioning, *Applied Psychological Measurement 17*, 105–116.

Roussos L.A. & Stout W.F., (1996). Simulation studies of the effects of small sample and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Samejima F., (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*.

Samejima F., (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*, 17–35.

Scott N.W., Fayers P.M., Aaronson N.K., Bottomley A., DeGraeff A., Groenvold M., Gundy C., Koller M., Petersen M.A. & Sprangers M.A.G., (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62,* 288–295.

Scott N.W., Fayers P.M., Aaronson N.K., Bottomley A., DeGraeff A., Groenvold M., Gundy C., Koller M., Petersen M.A. & Sprangers M.A.G., (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research, 16,* 115–129.

Spray J. & Miller T., (1994). *Identifying nonuniform DIF polytomously scored test items (94-1)*. ACS Research Report Series.

Swaminathan H. & Gifford J.A., (1985). Bayesian estimation in the two parameter logistic model. *Psychometrika, 50*, 349–364.

Swaminathan H. and Rogers H., (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27*, 361–370.

Vaughn B.K., (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach* (Unpublished Doctoral dissertation). Florida State University, Tallahassee, FL.

Wang N. & Lane S., (1996). Detection of Gender-Related Differential Item Functioning in a Mathematics Performance Assessment. Applied Measurement in Education, 9, 175–199.

Welch C.J. & Hoover H.D., (1993). Procedures for extending item bias techniques to polytomously scored items. *Applied Measurement in Education*, *6*, 1–19.

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality & Quantity, 38,* 681–702.

Yee T.W., (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software, 32*.

Zumbo B.D., (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.