Psicológica (2013), 34, 327-342.

Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression

Paula Elosua*[1] and Craig S. Wells[2]

[1] *University of the Basque Country (Spain)*

[2] *University of Massachusetts, Amherst (USA)*

The purpose of the present study was to compare the Type I error rate and power of two model-based procedures, the mean and covariance structure model (MACS) and the item response theory (IRT), and an observed-score based procedure, ordinal logistic regression, for detecting differential item functioning (DIF) in polytomous items. A simulation study was employed in which polytomous data with five ordered categories were generated using Samejima's graded response model under three crossed factors: sample size per group (300-, 500-, and 1,000-examinees), type of DIF (*b*-parameter, *a*-parameter, and *a*- and *b*-parameter DIF), and magnitude of DIF (small and large magnitudes of DIF). The Type I error rate was inflated for IRT based tests and ordinal logistic regression when some of the items contained DIF. For the uniform DIF conditions, MACS and IRT exhibited similar power rates; however, ordinal logistic regression exhibited slightly higher power compared to the other two methods for smaller sample sizes. Lastly, for non-uniform DIF, IRT exhibited much more power compared to MACS and ordinal logistic regression.

Measurement invariance plays a crucial role in interpreting test scores appropriately for individuals from different populations (Raju, Laffitte, & Byrne, 2002). Measurement equivalence or invariance is satisfied across groups (e.g., language, race, cultural, gender, etc.) if persons with the same level of proficiency on the latent variable (measured variable) have the same expected raw score in the observed variable at the item or test score level (Drasgow & Kanfer, 1985). One way of assessing measurement invariance is to examine whether the items are functioning differentially

* This research was supported through grants (PSI2011-30256) by Ministry of Economy and Competitivity and by University of Basque Country (GUI12/32). Correspondence details: Paula Elosua. University of Basque Country. Faculty of Psychology. Tolosa Avenue, 70. 20018 San Sebastián, Spain. E-mail: paula.elosua@ehu.es

(DIF) between the groups of interest. An item exhibits DIF if the probability of answering an item correctly or responding to a particular category differs for individuals from different groups but with the same level of proficiency.

Most of the research to date has examined the detection of DIF for cognitive and psychological assessments and questionnaires that use dichotomous items. However, the use of polytomous items is common in the psychological and educational assessment. Although there are some interesting papers about the detection of DIF in polytomous data (French & Miller, 1996; Kristatkansson, Aylesworth, McDowell, & Zumbo, 2005; Spray & Miller, 1994; Zwick, Donogue & Grima, 1993), the level of development regarding the detection procedures for polytomous items has not been widely studied. The purpose of the present paper was to examine several methods for detecting DIF for polytomous items that are often used on psychological or educational assessments.

DIF Detection Methods

DIF statistics may be classified as either latent model-based or observed-score based. Latent model-based procedures rely on fitting a latent variable model to the data in which the relationship between the item response and underlying latent variable is defined. There are two general types of model-based procedures: linear and non-linear. An example of a linear model-based procedure is the mean and covariance structure model (MACS; Sörbom, 1974). In MACS, a factor linear model is fit to the data for each group in which the relationship between the item response and latent variable is assumed to be linear; measurement invariance is assessed by comparing the factor loadings (i.e., non-uniform DIF) and item thresholds (i.e., uniform DIF) for each item between two or more groups. For non-linear model based procedures, an item response theory (IRT; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1989; Lord, 1980) model is fit to the data for each group in which the relationship between probability of responding to a particular category and the latent variable is described; DIF is assessed by comparing the item parameters (*a*-parameter for non-uniform DIF, and *b*-parameter for uniform DIF) between the two groups.

Observed-score based procedures focus on examining the relationship between item performance and observed scores, often based on total scores. Examples of observed-score based procedures are the Mantel-Haenszel procedure (Mantel, 1963; Potenza & Dorans, 1995), logistic regression (Swaminathan & Rogers, 1990) and discriminant logistic regression (French & Miller, 1996). The advantages of observed-score methods is that they do not rely on a measurement model to fit the data and they do not require parameters within a measurement model to be estimated.

Previous Research

Several studies have compared the MACS approach, which uses a linear factor analytic model, with IRT-based methods that use a nonlinear-model (see Raju, Laffitte, & Byrne, 2002; Meade & Lautenschlager, 2004; Stark, Chernyshenko & Drasgow, 2006). Meade and Lautenschlager (2004) provided a detailed comparison of the likelihood-ratio test within IRT and MACS[1] that included a simulation study that examined the Type I error rate and power of both methods. Polytomous item responses were generated using Samejima's GRM under conditions that varied with respect to type of DIF (uniform and non-uniform DIF) and sample size. The authors found that MACS was unable to detect uniform or non-uniform DIF at the item-level whereas the IRT-based method detected both types of DIF.

Kim and Yoon (2011) compared multiple-group categorical CFA (MG-CFA) to the likelihood-ratio test in IRT in detecting measurement invariance in simulating data. Ordered categorical item responses (dichotomous and polytomous) were generated using the MG-CFA model for various sample sizes and sources of DIF. The MG-CFA was implemented using a backward procedure where the chi-square difference tests were compared to a fully invariant baseline model. The Type I error and power rate were compared across the two models. A Dunn-Bonferroni correction was employed to control the overall Type I error rate. The authors found that the likelihood-ratio test exhibited better Type I error rate control compared to the MG-CFA approach. Both methods exhibited similar power rates. However, because power is conditional on controlling the Type I error rate, it is questionable to interpret the detection of DIF items using the MG-CFA model as power since it exhibited inflated Type I error rates.

Stark, Chernyshenko, and Drasgow (2006) compared MACS and the IRT on detecting DIF for dichotomous and polytomous item responses across various types and amount of DIF, sample sizes, and latent mean differences. Contrary to Meade and Lautenschlager (2004), Stark et al. (2006) found that MACS was able to detect uniform DIF. The differences illustrated in these two studies may be related to the way MACS was implemented. Meade and Lautenschlager (2004) used a constrained-

[1] Meade and Lautenschlager used the terminology confirmatory factor analysis with latent means and intercepts. MACS are included within such models.

baseline model approach whereas Stark et al. (2006) used the free-baseline model approach. Another important difference between Meade and Lautenschlager (2004) and Stark et al. (2006) is that the latter study used a linear common factor model to generate the item responses instead of Samejima's GRM. In this study, we will be using Samejima's GRM to simulate the item responses and the free-baseline approach to testing measurement invariance.

While the model-based approaches such as MACS and IRT are useful for examining DIF and measurement invariance, they are limited by the fact that a measurement model must fit the data and that a sufficient sample size is required to estimate the model parameters. Therefore, observed-score based methods such as logistic regression and Mantel-Haenszel are attractive alternatives because they do not implement a measurement model to detect DIF. Furthermore, many of the observed-score based methods provide an effect size (e.g., pseudo-R^2 for logistic regression) that may be used to judge the magnitude of DIF.

The purpose of the present study was to compare the Type I error rate and power of three approaches for the DIF detection using data generated under Samejima's GRM: IRT, MACS, and ordinal logistic regression. We chose the model-based approaches in order to delve deeper into the similarities and differences between them when the data are polytomous. Ordinal logistic regression was selected because it is flexible and also the most general form of the contingency tables and generalized linear modeling approaches to DIF detection (Zumbo & Hubbley, 2003). A secondary purpose of the study was to provide practitioners guidance about which method is most effective and useful for detecting DIF.

METHOD

Study Design

A Monte Carlo simulation study was conducted to examine the Type I error and power rates of the MACS, likelihood-ratio test, and ordinal logistic regression in detecting DIF for polytomous data. Polytmous item response data were generated using Samejima's GRM to represent a test with 15 items, each comprised of five categories (e.g., Likert-type items). The test length of 15 items was selected according to the number of items that are often observed on personality inventories (e.g., 16PF, Cattell's Personality Inventory). The generating item parameter values, reported in Table 1, were obtained from real data from an actual personality inventory. The data were generated for two groups, a reference and focal group. The generating item parameter values for items 1 to 3 shown in Table 1 were altered to generate DIF for the focal group.

Table 1. Generating Item Parameter Values.

Item	α	β_1	β_2	β_3	β_4
1	1.35	-2.59	-1.18	0.18	2.01
2	1.10	-2.29	-0.46	1.04	3.14
3	0.37	-2.74	0.44	3.71	8.36
4	0.61	-2.80	-1.13	0.28	4.52
5	2.32	-2.12	-0.73	0.40	1.81
6	1.86	-1.48	-0.61	0.38	1.91
7	1.72	-1.68	-0.57	0.71	2.25
8	1.54	-1.62	-0.93	0.08	1.41
9	1.74	-2.08	-0.66	0.41	1.87
10	1.47	-2.50	-0.77	0.48	2.37
11	0.91	-3.85	-1.85	-0.30	2.28
12	1.26	-1.65	-0.07	1.31	3.17
13	2.33	-1.82	-0.73	0.15	1.40
14	2.10	-1.84	-0.61	0.60	2.07
15	1.50	-1.60	-0.91	0.04	1.40

The data were generated under three crossed factors: sample size per group (300-, 500-, and 1,000-examinees), type of DIF (*b*-parameter, *a*-parameter, and *a*- and *b*-parameter DIF), and magnitude of DIF (small and large magnitudes of DIF). The three sample sizes were chosen to represent a small, medium and large sample size used in empirical research in psychology. The three types of DIF were selected to represent uniform and non-uniform DIF. Small and large magnitudes of DIF were selected to examine the power rates for each DIF detection method across realistic magnitudes of DIF. For the small magnitudes of DIF, the *a*- and *b*-parameter values for the focal group were altered by 0.25. For the large magnitudes of DIF, the *a*- and *b*-parameter values were altered by 0.40 and 0.50, respectively. In addition, a non-DIF condition was examined in which the generating item parameter values for the focal and reference group were identical. Therefore, data were generated for 21 conditions: 3 sample sizes

X 3 types of DIF X 2 magnitudes of DIF plus 3 sample sizes for the non-DIF condition. For each condition, 100 replications were performed. The proficiency parameters were sampled from the standard normal distribution for the reference and focal groups, $\theta \sim N(0,1)$.

DIF Detection

MACS. To evaluate measurement invariance, multiple group confirmatory factor analysis (MG-CFA) is typically performed, as well as the computation of the chi-square difference test for nested models. The first step in the detection of DIF is to define the baseline model which is referred to as the free model. To test the invariance for item *i*, the factor loading associated with item *i* is constrained to be equal between the groups. The fit of this new model is then compared to the fit of the free baseline model by taking the difference in chi-square fit statistics. If the model with the additional constraint fits significantly worse than the free baseline model, then item *i* is considered to function differentially between the groups. The degrees of freedom equal the difference in degrees of freedom between the two models which equals the number of parameters being compared. For each item and for each sample replication, one chi-square difference test was computed between both the baseline model and the item-constrained model. We made one comparison for every item for each data replication, excluding the reference item (for identification purposes, the 15th item was defined as the reference item; Lubke & Muthén, 2004). The MACS procedure was implemented using Mplus 5.0 (Muthén & Muthén, 2005) via ML estimation.

Ordinal logistic regression. The ordinal logistic regression was implemented using proportional odds ordinal logistic regression via maximum likelihood estimation. The analyses were performed using the *lrm* function within the software package R. The matching variable was defined as the total score obtained by summing all the individual items. Each item was evaluated for DIF by comparing the fit of two models. The first model was the baseline model, which only included one predictor, the total score. The second model included as predictors the total score plus two additional parameters, the grouping parameter and the interaction between group and observed score. After estimating both models the difference between the -2Log Likelihood from the model with more parameters and the model with less parameters was used to evaluate if the fit was better when information on the group was included in the model. This difference is a likelihood ratio test G_L^2 which is distributed as a χ^2 with degrees of freedom equal to the difference in the number of parameters estimated in the compact model and augmented model (i.e., *df* = 2).

IRT. The likelihood-ratio (LR) test is one of the more popular IRT procedures for detecting DIF due to its control of Type I error rate and acceptable power rates (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinber, & Wainer, 1988). In addition, the LR test can be used to detect uniform and non-uniform DIF independently. The LR test essentially compares the fit of a compact and augmented model to test for DIF between a reference and focal group. The compact model constrains the parameter values to be equal between the reference and focal groups (i.e., assumes no DIF is present). The augmented model allows the parameter values for one item (or a set of items) to be freely estimated in each group, constraining the remaining items to be equal between groups. DIF is assessed by comparing the overall fit of both models. If the item being tested contains DIF (i.e., the parameter values are not equal between the groups), then the overall fit for the augmented model will be much better than the overall fit for the compact model. The overall fit of the respective model is provided by -2 times log likelihood (-2 Log *L*). The IRT likelihood ratio test was implemented using the computer program IRTLRDIF v. 2.0 (Thissen, 2001).

Data Analysis

An item was flagged for DIF if the difference of the chi-square values between two models was significant using an alpha level of 0.05. The empirical Type I error rate was defined by the proportion of non-DIF items that were flagged across replications. The empirical power rate was defined by the number of times that the manipulated (i.e., DIF) item was flagged across replications.

RESULTS

Non-DIF Conditions

Table 2 reports the Type I error rate for the non-DIF condition (i.e., no items were generated to exhibit DIF). The three statistics did not exhibit inflated Type I error rates across the three sample sizes.

Table 2. Type I Error Rate for Non-DIF Conditions.

Sample Size	MACS	IRT	Logistic Regression
N=300	0.02	0.06	0.05
N=500	0.03	0.05	0.05
N=1,000	0.04	0.05	0.06

Uniform DIF (*b*-parameter)

Table 3 reports the Type I error rate when uniform DIF was present for items 1 to 3. The Type I error rate was based on the detection rate for items 4 to 15 (i.e., those not simulated to exhibit DIF). MACS appeared to exhibit controlled Type I error rates whereas logistic regression exhibited noticeably inflated Type I error rates for the large magnitude of DIF conditions with *N*=500 and *N*=1,000. The Type I error rate for the likelihood-ratio test also appeared to be slightly influenced by the magnitude of DIF. It is likely that the presence of DIF contaminated the covariate (i.e., total score) in logistic regression that was used to match the reference and focal groups. For the likelihood-ratio test, the presence of DIF negatively influenced the concurrent calibration. The reason MACS did not suffer an inflated Type I error rate was likely because the reference item was not simulated to exhibit DIF.

Table 4 reports the empirical power rates for detecting uniform DIF. The cells corresponding to conditions in which the Type I error rate was noticeably influenced are X-ed out since power is conditional on controlling the Type I error rate. For conditions in which the Type I error rate was controlled, logistic regression exhibited slightly more power in detecting uniform DIF compared to MACS and the likelihood-ratio test. MACS and the likelihood-ratio test were comparable across the conditions.

Non-uniform DIF (*a*-parameter)

Table 5 reports the Type I error rate when non-uniform DIF was present for items 1 to 3. The Type I error rate was based on the detection rate for items 4 to 15 (i.e., those not simulated to exhibit DIF). Each statistic procedure exhibited controlled Type I error rates when some of the items exhibited non-uniform DIF. It is interesting that although the presence of uniform DIF had a negative influence on the Type I error rate, non-uniform did not appear to negatively influence the covariate in logistic regression or linking procedure (i.e., concurrent calibration) used in the likelihood-ratio test.

Table 3. Type I Error Rate in the Presence of Uniform (*b*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
Sample Size	N=300	Small	0.05	0.06	0.03
		Large	0.04	0.06	0.08
	N=500	Small	0.04	0.05	0.07
		Large	0.04	0.08	0.12
	N=1,000	Small	0.04	0.06	0.09
		Large	0.03	0.09	0.21

Table 4. Power Rates for Detecting Uniform (*b*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
Sample Size	N=300	Small	0.20	0.22	0.24
		Large	0.59	0.62	0.64
	N=500	Small	0.28	0.31	0.37
		Large	0.72	0.72	~~0.75~~
	N=1,000	Small	0.51	0.52	0.57
		Large	0.80	0.74	~~0.76~~

Cells where the Type I error rate was inflated have been X-ed out.

Table 6 reports the power rates for detecting non-uniform DIF (items 1 to 3). It was apparent that the likelihood-ratio test exhibited far more power in detecting non-uniform DIF compared to MACS and the logistic regression procedure. Even in the small magnitude DIF conditions, the likelihood-ratio test exhibited considerable power. MACS and logistic regression exhibited comparable power across the conditions.

Table 5. Type I Error Rate in the Presence of Non-uniform (*a*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
	N=300	Small	0.04	0.05	0.04
		Large	0.03	0.05	0.05
Sample Size	N=500	Small	0.03	0.06	0.04
		Large	0.03	0.05	0.04
	N=1,000	Small	0.04	0.05	0.05
		Large	0.04	0.05	0.04

Table 6. Power Rates for Detecting Non-uniform (*a*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
	N=300	Small	0.09	0.66	0.08
		Large	0.08	0.98	0.10
Sample Size	N=500	Small	0.08	0.80	0.08
		Large	0.13	1.00	0.11
	N=1,000	Small	0.15	0.97	0.11
		Large	0.21	1.00	0.16

Non-uniform DIF (*a*- and *b*-parameter)

Table 7 reports the Type I error rate when the *a*- and *b*-parameter values differed between the reference and focal groups for items 1 to 3. The Type I error rate was based on the detection rate for items 4 to 15 (i.e., those not simulated to exhibit DIF). MACS appeared to exhibit controlled Type I error rates. Logistic regression exhibited noticeably inflated Type I error rates for the large magnitude of DIF conditions with *N*=500 and *N*=1,000 as well as the small magnitude of DIF conditions with *N*=1,000. The likelihood-ratio test exhibited inflated Type I error rates for the large magnitude of DIF condition with *N*=1,000. The observed inflated Type I error rates are not surprising considering the uniform DIF results where the Type I error rates were also inflated.

Table 8 reports the empirical power rates for detecting non-uniform DIF in which the *a*- and *b*-parameter both exhibited DIF. The cells corresponding to conditions in which the Type I error rate was noticeably influenced are X-ed out since power is conditional on controlling the Type I error rate. For conditions in which the Type I error rate was controlled, the likelihood-ratio test exhibited considerably more power than MACS and logistic regression; MACS and logistic regression exhibited comparable power for the same conditions. However, MACS exhibited controlled Type I error rate and reasonable power rates as the sample size increased (however, this was most likely because the reference item was not simulated to exhibit DIF).

Table 7. Type I Error Rate in the Presence of Non-uniform (*a*- and *b*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
	N=300	Small	0.04	0.06	0.06
		Large	0.04	0.08	0.04
Sample Size	N=500	Small	0.03	0.06	0.08
		Large	0.03	0.09	0.16
	N=1,000	Small	0.05	0.07	0.12
		Large	0.03	0.13	0.31

Table 8. Power Rates for Detecting Non-uniform (*a*- and *b*-parameter) DIF.

		Magnitude of DIF	MACS	IRT	Logistic Regression
	N=300	Small	0.28	0.69	0.32
		Large	0.86	1.00	0.78
Sample Size	N=500	Small	0.42	0.84	0.45
		Large	0.97	1.00	~~0.89~~
	N=1,000	Small	0.77	1.00	~~0.72~~
		Large	1.00	~~1.00~~	~~0.98~~

Cells where the Type I error rate was inflated have been X-ed out.

DISCUSSION

The purpose of the present study was to compare two model-based procedures, MACS and IRT, and an observed-score based procedure, ordinal logistic regression, for detecting DIF in polytomous items. Although each method exhibited controlled Type I error rates when none of the items exhibited DIF, the IRT based likelihood-ratio test and ordinal logistic regression exhibited inflated Type I error rates when DIF was present in some of the items; the Type I error rate for MACS did not appear to be influenced by the presence of DIF. This appears to be related to how group differences are equated before testing for DIF. The IRT based likelihood-ratio test uses concurrent calibration to place the item parameter estimates for the two groups onto the same scale so that they may be compared. If some of the items used in the concurrent calibration function differentially, the linking of the two scales may be corrupted making it difficult to interpret differences in item parameter estimates. Logistic regression uses an estimate of proficiency such as the total score to control for differences between the groups. When the estimate of proficiency is influenced by DIF, group comparisons may be invalid. MACS, on the other hand, equates the groups by fixing one of the factor loadings to 1.0 in both groups (i.e., reference item); therefore, the scale of the latent variable for both groups is determined by the same variable, controlling for measurement error. If the reference item functions differentially between the groups, the remaining comparisons will likely be invalid. In this study, the reference item was free of DIF.

Given the importance of selecting appropriate anchor items or a reference item, it is worthwhile examining a selection method of anchor items when testing for DIF. For example, Woods (2009) developed and investigated an empirical procedure for selecting appropriate anchor items for tests of DIF using the likelihood-ratio test. Essentially, the method is a rank-based strategy in which the items with the smallest DIF statistics are selected to represent the anchor. Once those items are selected, they are used to link the scales for the two groups that are then used to test for DIF. Woods found that the rank-based method improved DIF detection. Although the method was developed using the likelihood-ratio test, it could easily be extended to other methods such as MACS or logistic regression.

For uniform DIF, MACS and IRT exhibited similar power rates. This is particularly interesting since the IRT procedure would be expected to have an advantage since it implemented the same model that was used to generate the data (i.e., Samejima's GRM). These results are consistent with Stark et al. (2006) but contradict Meade and Lautenschlager (2004). In this study, we used the free-baseline model, which was also used in Stark et al. (2006). This is evidence supporting the use of the free-baseline approach when assessing measurement invariance. Furthermore, logistic regression exhibited slightly more power than MACS and IRT for the smaller sample size conditions. This may be due to the fact that the model-based procedures require larger sample sizes in order to obtain accurate model parameter estimates that are used in testing DIF. This problem may be exacerbated in real data where the models will not fit the data perfectly.

For non-uniform DIF (*a*-parameter only), only IRT exhibited reasonable power rates. This is consistent with Meade and Lautenschlager (2004) who found that MACS was unable to detect *a*-parameter DIF. Logistic regression also exhibited very low power rates for detecting non-uniform DIF. However, when the item was simulated to exhibit DIF in the *a*- and *b*-parameter, MACS and logistic regression exhibited reasonable power rates, most likely because they were detecting DIF in the *b*-parameter.

For future research, it would be interesting to compare the measurement model approaches (e.g., MACS and the likelihood-ratio) to other observed-score procedures such as the Mantel-Haenszel procedure. For example, Fidalgo and Madeira (2008) and Fidalgo and Scalon (2010) illustrated a unified framework for testing DIF in multiple groups for both dichotomous and polytomous items using the generalized Mantel-Haenszel method. A primary advantage of the Mantel-Haenszel procedures is that they provide an effect size that summarizes the magnitude of the DIF. The disadvantage of the Mantel-Haenszel procedures is that they cannot detect non-uniform DIF. It would be interesting to compare these procedures to model-based procedures under conditions of model fit and model misfit.

RESUMEN

Detección de DIF en ítems politómicos por medio de MACS, TRI y Regresión Logística Ordinal. El objetivo de este trabajo fue comparar el error Tipo I y la potencia de tres métodos de detección de funcionamiento diferencial del ítem en respuestas politómicas. Se compararon dos procedimientos basados en los modelos de estructuras de medias y covarianzas (MACS) y la teoría de respuesta al ítem (IRT) con un tercer procedimiento de puntuación observada, la regresión logística ordinal. Se utilizó simulación Montencarlo para generar datos según el modelo de respuesta graduada de Samejima. Se manipularon tres factores: tamaño de la muestra por grupo (300-, 500-, y 1,000- sujetos), tipo de DIF (*b*-parámetro, *a*-parámetro y *a*- y b parámetros), y magnitud de DIF (pequeño y grande). El error tipo I en presencia de DIF fue mayor que el esperado para la TRI y la regresión logística ordinal. Para la condición de DIF uniforme, MACS y TRI mostraron potencias similares, sin embargo, la regresión logística ordinal mostró una potencia algo superior al resto para tamaños de muestra pequeños. En las condiciones de DIF no uniforme, la potencia de la TRI fue mayor que MACS y la regresión logística ordinal.

REFERENCES

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogoneous populations. *Journal of Applied Psychology, 70,* 662-680

Embretson, S. E., and. Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fidalgo, A. M. & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 58*(6), 940-958.

Fidalgo, A. M. & Scalon, J. D. (2010). Using generalized Mantel-Haenszel statistics to assess DIF among multiple groups. *Journal of Psychoeducational Assessment, 28*(1), 60-69.

French, A. W., and Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

Hambleton, R. K., Swaminathan, H. & Rogers,H.J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logistic Regression* (2nd edn). Wiley.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

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*(6), 935-953.

Kim, E. S. & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*, 212-228.

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

Lubke, G. H., and Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534.

Mantel N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of American Statistic Association, 58*, 690-700.

Meade, A. W., and Lautenschlager, G. J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177-185.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 521-543.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus users's guide*. Los Angeles: Muthén and Muthén.

Potenza, M. T., & N. J. Dorans (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement 19*(1), 23-37.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517-529.

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

Sörbom, D. (1974) A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Spray, J. A., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items*. (American College Testing Research Report Series 94-1) Iowa City, IA: American College Testing Program.

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.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (http://www.unc.edu/~dthissen/dl.html).

Thissen, D., Steinberg, L, & Gerard, M. (1986). Beyond group mean differences: the concept of item bias.Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., & Wainer, H.. (1988). Use of item respone theory in the study of group differences in trace lines. In H.Wainer and H. I. Braun (Eds). *Test validity*. (pp.147-169). NJ: Lawrence Earlbaum associates.

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): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Otawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D., & Hubley, A.M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment* (pp. 505-509). Thousand Oaks, CA: Sage Press

Zwick, R., Donogue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

(Manuscript received: 1 June 2012; accepted: 6 November 2012)