Psicothema

Psicothema 2016, Vol. 28, No. 1, 83-88
doi: 10.7334/psicothema2015.142

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2016 Psicothema
www.psicothema.com

A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests

María Dolores Hidalgo[1], María Dolores López-Martínez[1], Juana Gómez-Benito[2] and Georgina Guilera[2]
[1] University of Murcia and [2] University of Barcelona

Abstract

Background: Short scales are typically used in the social, behavioural and health sciences. This is relevant since test length can influence whether items showing DIF are correctly flagged. This paper compares the relative effectiveness of discriminant logistic regression (DLR) and IRTLRDIF for detecting DIF in polytomous short tests. **Method:** A simulation study was designed. Test length, sample size, DIF amount and item response categories number were manipulated. Type I error and power were evaluated. **Results:** IRTLRDIF and DLR yielded Type I error rates close to nominal level in no-DIF conditions. Under DIF conditions, Type I error rates were affected by test length DIF amount, degree of test contamination, sample size and number of item response categories. DLR showed a higher Type I error rate than did IRTLRDIF. Power rates were affected by DIF amount and sample size, but not by test length. DLR achieved higher power rates than did IRTLRDIF in very short tests, although the high Type I error rate involved means that this result cannot be taken into account. **Conclusions:** Test length had an important impact on the Type I error rate. IRTLRDIF and DLR showed a low power rate in short tests and with small sample sizes.

Keywords: Differential item functioning, polytomous items, short tests, discriminant logistic regression, IRTLRDIF.

Resumen

Funcionamiento diferencial del ítem en tests breves: comparación entre regresión logística discriminante e IRTLRDIF. **Antecedentes:** en ciencias sociales, del comportamiento y de salud es habitual usar tests breves. El tamaño del test puede afectar a la correcta identificación de ítems con DIF. Este trabajo compara la eficacia relativa de la Regresión Logística Discriminante (RLD) e IRTLRDIF en la detección del DIF en tests cortos politómicos. **Método:** se diseñó un estudio de simulación. Se manipuló tamaño del test, tamaño de la muestra, cantidad DIF y número de categorías de respuesta al ítem. Se evaluó el Error Tipo I y la potencia. **Resultados:** en las condiciones de no-DIF IRTLRDIF y RLD mostraron tasas de Error Tipo I cercanas al nivel nominal. En tests con DIF las tasas de Error Tipo I dependieron del tamaño del test, de la muestra, cantidad de DIF, contaminación del test y número de categorías del ítem. RLD presentó mayor tasa de Error Tipo I que IRTLRDIF. La potencia estuvo afectada por la cantidad de DIF y tamaño de la muestra. En tests muy cortos RLD mostró mayor potencia que IRTLRDIF. **Conclusiones:** en tests cortos y con DIF las tasas de Error Tipo I fueron elevadas. La potencia de IRTLRDIF y RLD fue relativamente baja en tests cortos y tamaños muestrales pequeños.

Palabras clave: funcionamiento diferencial del ítem, ítems politómicos, tests breves, regresión logística discriminante, IRTLRDIF.

Test score interpretation may be invalidated by the presence of differential item functioning (DIF) among different groups of respondents based on characteristics such as gender, country or the language version of the test used. DIF is present if groups have different probabilities of success on an item after being matched on the attribute measured by the test.

DIF detection for polytomously scored items has attracted much attention in recent decades, with a wide variety of statistical techniques being proposed for the identification of DIF. Logistic regression (LR) for polytomous items (French & Miller, 1996) is a popular non-parametric procedure, although item response theory (IRT) procedures such as likelihood ratio methods (Thissen, Steinberg, & Wainer, 1988) have also been used. The efficiency and effectiveness of these procedures in detecting DIF have been studied under different simulated conditions in the educational assessment context, where tests generally have a large number of items (Gelin & Zumbo, 2007). In the health sciences, however, scales tend to have only a small number (3-6) of items (Scott et al., 2010; Teresi, 2006). Short scales and brief test versions of this kind are increasingly popular, as they are quick to apply and easy to score, thus making them well suited for use in screening processes, clinical assessment, survey research, and other assessment contexts.

Test length can influence whether DIF items are correctly flagged, the main problem being the psychometric quality of the matching score. Another factor that may produce misleading results regarding DIF detection is matching criterion contamination.

Received: May 31, 2015 • Accepted: November 6, 2015
Corresponding author: María Dolores Hidalgo
Faculty of Psychology
University of Murcia
30080 Murcia (Spain)
e-mail: mdhidalg@um.es

When the matching variable used to detect DIF is the total test score or an estimate of test ability, and one or more items are biased, this can lead to an inaccurate ability estimates and, therefore, false DIF identification. Thus, under this condition, if the test is also short, this only exacerbates the problem because it increases the risk of detecting items with pseudo-DIF (Scott et al., 2009), that is, items flagged as showing DIF due to the high degree of test contamination. This adds to the difficulty of interpreting DIF and of explaining its causes.

However, the effectiveness of DIF detection methods in relation to test length has not been sufficiently explored. Scott et al. (2009), using ordinal logistic regression, simulated scale lengths of 2, 3, 4, 5, 10 and 20 items, and sample sizes of 500 or smaller, and found that the impact of the number of scale items was relatively small, and that DIF was successfully detected. Donoghue, Holland, and Thayer (1993), using the Mantel-Haenszel (MH) procedure and the standardization statistic, simulated tests with 4, 9, 19 and 39 items and concluded that DIF analyses of tests with 4 and 9 items was not recommended because the results were too dependent on confounding factors. Paek and Wilson (2011), comparing MH and IRT-DIF methods under the Rasch model, used the same test lengths as Donoghue et al. (1993) and small sample sizes (100/100, 200/200 and 300/300). They found that the IRT-DIF methods performed well and achieved higher statistical power than did the MH procedure. Although LR has been shown to be more flexible and efficient than other procedures for detecting DIF, as well as being easy to implement, the use of an observed test score as the matching variable may not be adequate, particularly for short scales, and as a number of studies suggest, latent trait estimation using IRT may be more appropriate (Bolt, 2002; Wang & Yeh, 2003).

Despite the fact that the use of DIF analyses with short scales can be problematic, numerous applied studies have sought to analyse DIF in such scales (Scott et al., 2010). Consequently, it is important to determine the relative effectiveness of parametric and non-parametric methods for detecting DIF. To this end the present paper compares the effectiveness of IRTLRDIF (a technique based on item response theory) and discriminant logistic regression (an observed matching criteria method) for detecting DIF in short polytomous tests.

In general, there are two forms of DIF, uniform and non-uniform. Although both DIF types are important (Sireci & Ríos, 2013), non-uniform occurs substantially less frequently than does uniform DIF (Camilli & Shepard, 1994). Thus, the present study focuses in polytomous uniform constant DIF pattern.

IRTLR Test (Item Response Theory Likelihood Ratio Test)

Thissen, Steinberg, and Wainer (1988) propose a likelihood ratio test for detecting DIF using IRT. In the IRTLR test, the null hypothesis of no differences in item parameters between groups is tested using a model comparison strategy. In the first IRT model (compact model), item parameters are constrained to be equal in the two groups. This model is fitted by constraining the studied item and an anchor item (or set of anchor items) that is DIF-free to have the same parameters in both groups. In the second IRT model (augmented model), the same anchor item (or set of items) is again constrained to be equal in both groups, whereas no between-group equality constraints are applied to the item under study. Given that compact model (C) is nested within the augmented model (A), a G^2 goodness-of-fit statistic is calculated for each model. The significance test of the null hypothesis is obtained by comparing C and A model as follows:

$$G^2 = -2LL_c - (-2LL_A)$$

which follows a central chi-square distribution with l degrees of freedom (df), where l is the difference between the number of item parameters estimated in model C and model A, respectively. If the value obtained is greater than the theoretical value of the chi-square distribution with l df, then we reject the null hypothesis and, by implication, model C, concluding that the specified item or items show DIF.

Discriminant Logistic Regression (DLR)

Miller and Spray (1993) proposed DLR as a method for evaluating DIF in polytomous items. DLR is basically a LR analysis for dichotomous variables, where the dependent variable is group membership, with two levels (focal and reference), and the predictor variables are item response (Y) (polytomous), observed test score (X) and the interaction XY. The discriminant function for non-uniform DIF is formulated as follows:

$$P(G \mid X,Y) = \frac{\exp(G-1)(\beta_0 + \beta_1 X + B_2 Y + \beta_3 XY)}{1+\exp(\beta_0 + \beta_1 X + \beta_2 Y + \beta_3 XY)}$$

This procedure is able to evaluate the existence of both uniform and non-uniform DIF, which can be modelled in the same equation, and it is possible to separately test the coefficients for each. These hypotheses are normally tested using a conditional likelihood ratio test (Hidalgo, Gómez-Benito, & Padilla, 2005). Thus, the likelihood ratio statistic estimated in the absence of DIF (model 1 that included the X variable) is compared with that obtained when the model is adjusted for the presence of DIF (i.e. the full model with X, Y and XY). If the difference between the two statistics is significant, the item is considered to show DIF. If the effect of the item response variable (Y) is significant but that of the interaction is not (model 2), the item shows uniform DIF.

Method

Data generation

Item responses were generated using the graded response model (Samejima, 1969). The boundary category characteristic curves (BCCCs) were defined to represent the cumulative probability ($P^*_{jk}(\theta)$) of a response above category k. The BCCC is given as:

$$P^*_{jk}(\theta) = \frac{\exp\left(a_j\left(\theta - b_{jk}\right)\right)}{1+\exp\left(a_j\left(\theta - b_{jk}\right)\right)}$$

where a_j is the discrimination parameter for item j, b_{jk} is the location parameter for item j in the boundary category k, and θ is the ability parameter.

Experimental conditions

Four variables were manipulated:

1) Sample size for the reference and focal groups (250/250, 500/500 and 1000/1000) reflected situations that are more likely in practice and involved small and large sample sizes for comparison groups. However, it is also frequent to find unbalanced sample sizes across groups, with smaller sample sizes for the focus groups. In these cases, equal sample sizes can be reached by extracting a random subsample from the majority group.
2) Amount of DIF: Two levels of differences in location parameters were manipulated (0.4 and 0.8), indicating that the magnitudes of DIF were simulated to be moderate and large for each item. IRT parameters for the simulated DIF effect are shown in Table 1.
3) Test length: Simulated tests comprised 4, 5, 8 or 10 items. A scale length of 10 items is representative of numerous short scales typically found in behavioural sciences. Shorter test lengths were used in order to consider other scales frequently used in screening studies. The percentage of DIF items in a test (0%, 10%, 12.5%, 20% and 25%) was also considered. The condition of 10% reflects a general situation in test use, although in test adaptations, the percentage of DIF items is frequently higher than 20%. This condition was manipulated in combination with the test length, such that simulated tests had only one item with DIF.
4) Number of response categories per item: As found in behavioural and health sciences, items were simulated to represent ordered categorical data with four or three categories.

A total of 48 plus 36 conditions were manipulated. Under each condition 100 replications were made.

Data analysis

DIF detection by DLR was estimated using a software program created by two of the authors (M.D.H. and J.G.B.). Total score was used as the matching criterion. Items were flagged as having significant DIF when the G^2 comparison statistic of Model 2 with respect to Model 1 was significant at $p \leq .05$.

DIF detection by IRTLR was estimated using the IRTLRDIF software (Thissen, 2001). A one-stage strategy was applied and all items except the item under study were used as anchors. Items were flagged as having significant DIF when the G^2 comparison statistic was significant at $p \leq .05$.

Finally, the empirical Type I error rate for each DIF-free item was assessed by the proportion of false positives. In conjunction, the empirical statistical power rate of each DIF item was assessed by the proportion of true positives. According to Bradley's (1978) liberal criterion of robustness, a test can be considered robust if its empirical rate of Type I error, $\dot{\alpha}$, is within the interval $0.5\ \alpha \leq \dot{\alpha} \leq 1.5\ \alpha$. Thus, for the nominal level $\alpha = 0.05$, the empirical Type I error rate should be within the interval $0.025 \leq \dot{\alpha} \leq 0.075$. If the empirical average Type I error was located beyond this liberal interval, it was declared to be inappropriate; thus, the corresponding power rate of DIF detection was meaningless.

Table 1
Item parameters for the reference group

	k = 4				*k* = 3		
Item	a_R	b_{1R}	b_{2R}	b_{3R}	a_R	b_{1R}	b_{2R}
1	0.99	-1.95	-0.19	2.57	0.99	-1.95	-0.19
2	0.37	-0.64	0.77	1.66	0.37	-0.64	0.77
3	0.90	-0.91	0.21	0.98	0.90	-0.91	0.21
4	0.88	-2.25	-1.80	1.66	0.88	-2.25	-1.80
5	0.63	-2.11	-0.54	0.74	0.63	-2.11	-0.54
6	0.99	-1.95	-0.19	2.57	0.99	-1.95	-0.19
7	0.37	-0.64	0.77	1.66	0.37	-0.64	0.77
8	0.90	-0.91	0.21	0.98	0.90	-0.91	0.21
9	0.88	-2.25	-1.80	1.66	0.88	-2.25	-1.80
10	0.63	-2.11	-0.54	0.74	0.63	-2.11	-0.54

Note: Items 1-4 for test length = 4; Items 1-5 for test length = 5; Items 1-8 for test length = 8; Items 1-10 for test length = 10. In all cases the item manipulated with DIF was item 1; *a*: discrimination parameter, *b*: threshold parameter; R: reference group; *k*: number of item response categories

Results

Type I Error Rate

Table 2 shows the Type I error rates of both procedures for detecting DIF in each condition. Regarding the condition of no DIF items, Type I error rates were, as expected, close to the nominal level across all sample sizes, test lengths and both DIF detection techniques. Specifically, Type I error rates ranged from .023 to .068 and were lower than .075 and higher than .025 in all conditions, except when IRTLRDIF was used with small sample sizes and short test lengths. In general, the number of item response categories had no effect: when k= 4, Type I error rates for DLR ranged from .035 to .068, whereas when k=3, they ranged from .034 to .068. A similar pattern was observed when IRTLRDIF was used: when k= 4, Type I error rates ranged from .023 to .049, whereas when k= 3, they ranged from .025 to .065.

Under DIF conditions, Type I error rates were affected not only by test length but also by amount of DIF, sample size and number of item response categories. Type I error rates differed according to the DIF detection technique used. When DLR was used to detect DIF, Type I error rates were above .075 in all conditions when DIF was manipulated in very short tests. Type I error rates ranged from .19 to .56 when DIF amount was 0.8, and from .08 to .31 when it was 0.4. Regarding the number of categories per item, Type I error rate ranged from .12 to .56 for k= 4, and from .08 to .43 for k= 3. In general, Type I error rates were higher when sample size was larger.

When IRTLRDIF was used to detect DIF in very short tests, Type I error rates were higher than the liberal criterion when DIF amount was 0.8, NR= NF= 1000, and k= 4, and also when DIF amount was 0.8 and the total sample size was 1000. Type I error rates were higher when sample size was larger.

Type I error rates decreased as test length increased. This effect was more notable for DLR than for IRTLRDIF. When DIF was manipulated in tests with n= 8, Type I error rates ranged from .07 to .17 when DIF amount was 0.8, and from .05 to .08 when it was 0.4. In tests with n= 8 and k= 4, Type I error rates ranged from .05 to .17, whereas in tests with k= 3, they ranged from .05 to .11. As in very short tests, Type I error rates were higher when

sample size was larger. When IRTLRDIF was used in tests with eight items, Type I error rates were higher than the liberal criterion in the condition of a larger amount of DIF (0.8), N= 2000 and k= 4. More specifically, Type I error rates ranged from .05 to .09 when the amount of DIF was 0.8, and from .04 to .07 when it was 0.4. These results were similar independently of the number of categories per item, as in both conditions, the average Type I error rate was lower than the liberal criterion. In general, Type I error rate was below .07 when the total sample size was 500 or 1000, and it ranged from .04 to .09 when the sample size was 2000. This pattern was similar when a longer test was considered. In very short tests, Type I error rates were lower for IRTLRDIF than for DLR, whereas the two procedures performed similarly with longer tests.

Table 2
Type I error rates at 5% for all conditions

Sample Size Reference/Focal	n	Amount of DIF	k = 4		k = 3	
			DLR	IRTLRDIF	DLR	IRTLRDIF
250/250	4	0.0	.065	.023	.068	.040
		0.4	**.120**	.033	**.307**	.060
		0.8	**.190**	.057	**.430**	.067
	5	0.0	.056	.046	.046	.040
		0.4	**.078**	.030	.053	.038
		0.8	**.103**	.058	**.083**	.048
	8	0.0	.061	.054	.056	.059
		0.4	.050	.053	.056	.046
		0.8	.070	.069	.069	.064
	10	0.0	.047	.045	.049	.056
		0.4	.059	.063	.056	.050
		0.8	.066	.060	.049	.062
500/500	4	0.0	.068	.033	.058	.030
		0.4	**.120**	.037	**.077**	.030
		0.8	**.327**	**.110**	**.197**	.053
	5	0.0	.046	.030	.034	.030
		0.4	**.113**	.060	.070	.033
		0.8	**.240**	**.078**	**.135**	.055
	8	0.0	.054	.058	.041	.035
		0.4	.066	.069	.053	.039
		0.8	**.097**	.063	**.080**	.059
	10	0.0	.050	.042	.048	.056
		0.4	.041	.047	.059	.050
		0.8	.069	.054	.058	.041
1000/1000	4	0.0	.035	.025	.045	.025
		0.4	**.220**	.070	**.150**	.063
		0.8	**.557**	**.210**	**.380**	**.077**
	5	0.0	.060	.040	.040	.040
		0.4	**.135**	.045	**.085**	.025
		0.8	**.340**	**.155**	**.205**	.053
	8	0.0	.045	.045	.050	.065
		0.4	**.076**	.047	.060	.044
		0.8	**.166**	**.086**	**.110**	.051
	10	0.0	.059	.048	.047	.054
		0.4	**.077**	.052	.063	.047
		0.8	**.118**	**.104**	**.076**	.039

Note: In bold, Type I error rate that exceeded the liberal criterion; *n*: test length; *k*: number of item response categories; DLR: discriminant logistic regression

Power rate

Table 3 shows that power rates were affected by DIF amount, sample size, and DIF technique used, but not by test length. When very short tests were considered (n= 4) and DLR was used, power rates were above .80 when DIF amount was 0.8, regardless of sample size and the number of item response categories. However, when DIF amount was 0.4, power rates ranged from .59 to 1, and they were below .80 when total sample sizes were 500 or 1000, irrespective of the number of response categories. When IRTLRDIF was used and DIF amount was 0.8, power rates were below .80 for very short tests and small sample sizes. When DIF amount was 0.4, power rates were only above .80 with larger sample sizes.

When longer tests were considered, both DLR and IRTLRDIF achieved power rates below .80 when total sample size was lower

Table 3
Power rate at 5% for all conditions manipulated

Sample Size Reference/Focal	n	Amount of DIF	k= 4		k= 3	
			DLR	IRTLRDIF	DLR	IRTLRDIF
250/250	4	0.4	.59*	.23	.60*	.03
		0.8	.99*	.64	1.00*	.00
	5	0.4	.63*	.38	.40	.34
		0.8	1.00*	.79	.89*	.77
	8	0.4	.65	.44	.39	.20
		0.8	.99	.87	.97	.74
	10	0.4	.55	.32	.36	.32
		0.8	1.00	.86	.96	.90
500/500	4	0.4	.76*	.60	.60*	.36
		0.8	1.00*	.83*	1.00*	.83
	5	0.4	.84*	.50	.64	.42
		0.8	1.00*	.87*	1.00*	.95
	8	0.4	.87	.62	.68	.54
		0.8	1.00*	.97	1.00*	.97
	10	0.4	.87	.64	.75	.61
		0.8	1.00	.98	1.00	.99
1000/1000	4	0.4	1.00*	.93	.93*	.80
		0.8	1.00*	.97*	1.00*	.93*
	5	0.4	.99*	.93	.94*	.86
		0.8	1.00*	.93*	1.00*	.99
	8	0.4	.98*	.96	.96	.90
		0.8	1.00*	1.00*	1.00*	.99
	10	0.4	.99*	.96	.93	.87
		0.8	1.00*	1.00*	1.00*	1.00

Note: An asterisk indicates that the power was meaningless because its corresponding average Type I error was inflated; *n*: test length; *k*: number of item response categories; DLR: discriminant logistic regression

than 2000, irrespective of the amount of DIF and the number of item response categories. However, under the above conditions, the power rate for both procedures was higher when the number of item response categories was k= 4 rather than k=3. Thus, when DIF was manipulated in tests with eight items and the technique used was DLR, power rates ranged from .36 to .65 when DIF amount was 0.4, and from .68 to .87 when it 0.8. In tests with eight items and with four categories per item, power rates were .65 (sample size 250/250) and .87 (sample size 500/500), whereas in tests with three categories per item, they were .39 (sample size 250/250) and .75 (sample size 500/500). When IRTLRDIF was used, power rates ranged from .20 to .44 when DIF amount was 0.4, and from .54 to .64 when it was 0.8. In tests with eight items and four categories per item, power rates were .44 (sample size 250/250) and .62 (sample size 500/500), whereas in tests with three categories per item, they were .20 (sample size 250/250) and .54 (sample size 500/500).

DLR achieved higher power rates than did IRTLRDIF in short tests, but this result cannot be taken into account due to the high Type I error rate involved. As expected, in tests with 8 or 10 items, DLR showed a higher power rate than did IRTLRDIF with smaller sample sizes, irrespective of the magnitude of DIF manipulated.

Discussion

Short scales are commonly used in the behavioural and health sciences, and this paper compared the effectiveness of DLR and IRTLRDIF for detecting DIF in short tests.

The main results concerning Type I error rates are consistent with previous research. First, both DLR and IRTLRDIF yielded Type I error rates close to the nominal level under no-DIF conditions. Conversely, in DIF conditions, Type I error rates were affected not only by the test length but also by the amount of DIF, the magnitude of matching criterion contamination, the sample size, and the number of item response categories. For both procedures, Type I error rates were higher than the nominal level in short tests with four item response categories and when the sample size and the amount of DIF were larger. It should be noted that in this simulation study, only one item with DIF was simulated in each test, and when test sizes were short (4 or 5 items), the degree of matching criterion contamination was consequently high (25% or 20% of DIF items in the test). However, Type I error rates were lower when IRTLRDIF was used and, in general, IRTLRDIF showed better control of Type I error rate than did DLR.

IRTLRDIF was, however, shown not to be adequate if the anchor test included a large amount of DIF. This is consistent with the findings of Finch and French (2008), González-Betanzos and Abad (2012), and Wang and Yeh (2003), all of whom found that Type I error rate is greatly influenced by the level of contamination in the anchor items. The results obtained in the present study were consistent with previous research when the number of item response categories was 4. However, with k= 3, Type I error rates were lower than the nominal alpha level. Type I error rate was particularly inflated when the amount of DIF was large.

When DLR was used to detect DIF, Type I error rates were affected by sample size, test length and the amount of DIF. In addition, there was an interaction between test length and the amount of DIF (higher Type I error rate with a greater amount of DIF in the test and shorter tests) and between sample size and the amount of DIF (higher Type I error rate with larger sample sizes and a greater amount of DIF).

The main findings regarding power were that rates were affected by the amount of DIF and sample size, but not by test length. In general, DLR achieved higher power rates than did IRTLRDIF in short tests, although this result cannot be taken into account because of the high Type I error rate involved. In tests with 8 or 10 items and small sample sizes, DLR showed a higher power rate than did IRTLRDIF. As in other studies, a larger proportion of DIF items, that is, a high level of contamination in the anchor items, was associated with a decrease in power for IRTLRDIF (Wang & Yeh, 2003).

Finally, as theory states, both the reliability of the matching variable and the variability in total test scores will be lower with short tests, such that Type I error rate may increase. However, when the MH procedure is used, test length has a minimal effect on the error rate and power when detecting DIF in tests between of 20 and 40 items, with results being worse with fewer than 20 items (Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca, 2013). Scott et al. (2009), using ordinal logistic regression, also found that test length was not relevant for tests of between 5 and 20 items. By contrast, the magnitude of matching criterion contamination does have an important effect on DIF detection (Guilera et al., 2013).

The disagreement found between the two detection methods was expected. The matching variable used by DLR is raw score, which are sufficient statistics of — and monotonically related to — IRT scales under the Rasch model or the partial credit model, but not under the 3-p model or the graded response model (DeMars, 2008). As for parametric approaches to DIF detection such as IRTLRDIF, the problem here, as Bolt (2002) points out, concerns model misspecification and the need for larger samples sizes in order to avoid inflated Type I error rates due to model misfit. In IRT-DIF methods, the power to detect DIF increases with increases in sample size, whereas methods such as LR (non-parametric approach) are powerful enough with relatively small sample sizes (Scott et al., 2009). In general, the election between IRTLRDIF and DLR should be guided not only by the sample size but also by the availability of software and statistical expertise. It is important to note that the DLR is available in most statistical software packages and requires relatively small sample sizes, but IRT-based likelihood ratio methods require relatively larger sample sizes and more restrictive model assumptions, they are likely to be best when sample sizes for all groups are large enough for stable parameter calibration item (Sireci & Ríos, 2013).

At all events, and regardless of the DIF detection method used, the main problem with short scales is the difficulty of identifying which item is causing DIF, as one item with DIF can contaminate the other items through their contribution to the matching criteria (Scott et al., 2009). Although purification procedures can be applied, they may be less suitable for scales with only a small number of items, as removing items can affect the precision of the matching variable (Scott et al., 2010). For scales of this kind, we would recommend the use of mixed methods (Benítez, Padilla, Hidalgo, & Sireci, in press) and an effect size statistic in order to make decisions about eliminating/changing DIF items in a test (Gómez-Benito, Hidalgo, & Zumbo, 2013). Multilevel LR may also be considered (Balluerka, Gorostiaga, Gómez-Benito, & Hidalgo, 2010).

Although the findings of the present study provide some practical advices regarding DIF detection in short tests and scales,

it has several limitations mainly related to the simulated conditions considered. Thus, further research is needed to determine the effect of unbalanced sample sizes for reference and focal groups, the detection of non-uniform DIF or the extension of results to other patterns of DIF. In the latter case, it is important to note that DIF in polytomous items are much more complex than in dichotomous items, and several DIF patterns depending on the response model can be found (Penfield, 2007; Penfield, Alvarez & Lee, 2009), so it would be interesting to extend this study manipulating DIF patterns (i.e., balanced, shift-low or shift-high DIF patterns) and using Differential Step Functioning framework (Penfield, 2007).

References

Balluerka, N., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema, 22,* 1018-1025.

Benítez, I., Padilla, J. L., Hidalgo, M. D., & Sireci, S. G. (in press). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education*.

Bolt, D. M. (2002). A Monte-Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Psychological Measurement, 15,* 113-141.

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

DeMars, C. E. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the Raw Score. *Educational and Psychological Measurement, 68*, 379-386.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Finch, W. H., & French, B. F. (2008). Anomalous Type I Error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68,* 742-759.

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. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods, 6,* 573-588.

Gómez-Benito, J., Hidalgo, M. D., & Zumbo, B. (2013). Efficiency of combination of R-square and Odds-ratio using discriminant logistic function for detecting DIF in polytomous items. *Educational and Psychological Measurement, 73*, 875-897.

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of Differential Item Functioning with the likelihood ratio test. *Methodology, 8,* 134-145.

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I Error and statistical power of Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18,* 553-571.

Hidalgo, M. D., Gómez-Benito, J., & Padilla, J. L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem [Logistic regression: Analytic strategies in differential item functioning detection]. *Psicothema, 17*, 509-515.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Paek, I., & Wilson, M. (2011). Formulating the Rasch Differential Item functioning model under the marginal maximum Likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71,* 1023-1046.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20,* 335-355.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of Differential Step Functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*, 61-78.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 17.*

Scott, N. W., Fayer, P. M., Aaronson, N. K., Bottomley, A., Graeff, A., Groenvold, M., …, Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) using short scales. *Journal of Clinical Epidemiology, 62,* 288-295.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A., Groenvold, M., …, Sprangers, M. A. G. (2010). Differential Item Functioning (DIF) analyses of health related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes, 8*, 81.

Sireci, S. G., & Ríos, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal of Theory and Practice, 19*, 170-187.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44,* 39-49.

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* [Computer software and manual]. University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In Wainer, H. & Braun, H. I. (Eds.), *Test Validity*, pp 147-169. Hillsdale, N.J.: Erlbaum.

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,* 479-498.