Anuario de Psicología/The UB Journal of Psychology
2015, vol. 45, nº 1, 55-69

Differential item functioning and cut-off scores: Implications for test score interpretation*

María Dolores Hidalgo[1]
Francisca Galindo-Garre[2]
Juana Gómez-Benito[3]
[1] *Universidad de Murcia*
[2] *University Medical Center, Amsterdam*
[3] *Universitat de Barcelona*

Psychological and educational measurement instruments are used to make decisions that can have an important impact on the person being assessed. It is therefore essential to ensure that tests are free from bias so that the scores they yield provide a fair interpretation. This study aimed to assess the impact that items showing differential functioning may have on test interpretations based on cut-off scores. To this end a simulation study was conducted in which we manipulated the size of the comparison groups (100, 250, 500 and 1000), the magnitude of differential item functioning (DIF) (set at 0.8 for the difference between the difficulty parameters of the two groups) and the degree of test contamination (0%, 10%, 20%, 30% and 40% of items with differential functioning). Overall, the simulation considered 20 conditions, 1000 replications and a 20-item test. Results indicated that the selected cut-off did have an influence, and as the degree of test contamination increased, greater differences between the groups were erroneously detected, both in terms of statistical test significance and effect size obtained. These findings highlight the importance of ensuring that measurement instruments are free from DIF so that the interpretation of scores is both accurate and fair, this being a key aspect of a test's validity.

Keywords: *Differential Item Functioning, DIF, Cut-off Scores, Rasch Model.*

* *Acknowledgements*: This study received funding from the Seneca Foundation, Agency for Science and Technology in the Region of Murcia, Funding Programme for Research Projects in the Humanities and Social Sciences (Project 11917/PHCS/09), and from Spain's Ministry of Science and Innovation (Project PSI2009-07280).

Correspondence: María Dolores Hidalgo, Departamento de Psicología Básica y Metodología, Universidad de Murcia, Campus de Espinardo, Apdo. 4020, 30080-Murcia. Teléfono: +34868883470. E-mail: mdhidalg@um.es

Funcionamiento diferencial del ítem y puntuaciones de corte: implicaciones en la interpretación de las puntuaciones de los tests

Los instrumentos de medida psicológicos y educativos se emplean en la toma de decisiones que afectan de modo relevante a las personas evaluadas. Por ello es clave que se garantice una interpretación equitativa de las puntuaciones obtenidas, mediante la utilización de tests no sesgados. El objetivo del trabajo es valorar el impacto de la presencia de ítems con funcionamiento diferencial en las interpretaciones basadas en puntuaciones de corte. Para ello se diseñó un estudio de simulación en que se manipuló el tamaño muestral de los grupos de comparación (100, 250, 500 y 1000), la magnitud de funcionamiento diferencial del ítem (establecida en 0.8 como diferencias entre los parámetros de dificultad de ambos grupos) y el grado de contaminación del test (0%, 10%, 20%, 30% y 40% de ítems con funcionamiento diferencial). En total se trabajó con 20 condiciones, 1000 réplicas y un test de 20 ítems. Los resultados evidenciaron la influencia del punto de corte seleccionado y mostraron que a mayor grado de contaminación del test se detectan erróneamente mayores diferencias entre los grupos de comparación, tanto a nivel de la prueba de significación como del tamaño del efecto estudiados. Todo ello permite concluir la relevancia de obtener evidencias de ausencia de DIF en los instrumentos de medida para lograr una interpretación precisa y equitativa de sus puntuaciones, en el marco de la validez del test.

Palabras clave: *funcionamiento diferencial del ítem, DIF, puntuaciones de corte, modelo de Rasch.*

Introduction

Tests, questionnaires and surveys are useful tools for measuring variables or attributes in the health and social sciences. These instruments tend to be administered with a specific purpose, and the data derived from them are often used to make decisions that may be of considerable significance for the respondent. Given that tools of this kind are now widely used in areas such as clinical diagnosis, personnel selection, public opinion research, health surveys and the assessment of academic performance, among others, it is of utmost importance to ensure that their application guarantees equal opportunities and the fair treatment of the persons to whom they are administered. In other words, it is essential that the test used does not contain biased items. To this end, it is vital to determine whether these measurement instruments are invariant and, in the event that they are not, what effect the presence of biased items may have on the results obtained. An item that is biased will unfairly favour one group over another, in other words, item response will vary depending on one or more group variables (e.g. ethnicity, gender, socioeconomic status, cultural background or language ability) which in themselves are not relevant to the construct that the test seeks to measure (Angoff, 1993). The technical term for this problem is differential item functioning (DIF). According to Millsap and Meredith (1992), measurement invariance is produced if and only if $P(Y \mid W = w, V = v) = P(Y \mid W = w)$, where P denotes the probability, Y is a random observed variable that is related to or which seeks to measure the random variable W, which is latent, and V is an observable random variable that defines multiple populations of subjects according to their values or categories. In the context of detecting items with DIF, the term *focal group* is used to define the set of individuals, generally a minority, who are the principal target of the study. Conversely, the *reference group*, generally a majority, constitutes the group of subjects with whom the focal group will be compared. For dichotomous items, DIF is said to be present if the probability of a correct response on the item depends on the group to which the subject belongs (focal or reference), despite these groups being matched on the attribute measured by the test (the latent variable or latent trait). The presence of items that behave differentially for one of the groups that are supposedly matched on the attribute measured by the test can have serious repercussions, since the members of these groups will then obtain different scores, leading in turn to erroneous interpretations when comparing the scores obtained (Li & Zumbo, 2009).

Although considerable progress has been made in developing statistical techniques for detecting DIF items and in evaluating which of these methods are the most useful and effective (Hidalgo & Gómez-Benito, 2010; Millsap & Everson, 1993; Osterlind & Everson, 2009; Penfield & Lam, 2000; Sireci & Rios, 2013), far less is known about the consequences that DIF items may have in psychometric test properties and the interpretation of the results derived from it. In research on DIF this question has been examined using different approaches. Meredith (1993), Widaman and Reise (1997) and Wu, Li, and Zumbo (2007) studied the effect of DIF items on an instrument's factor structure, with measurement invariance being evaluated by means of multigroup factor analysis. Other studies have explored the effect of DIF in predictive validity with respect to an external criterion, a phenomenon referred to as *differential prediction* (Drasgow, 1982; Linn, 1984; Roznowski & Reith, 1999; Stark, Chernyshenko, & Drasgow, 2004). In this case, the focus of interest was on the effect of biased items in tests used for personnel selection and promotion purposes, principally in the field of organizational and work psychology, since tests containing such items may not predict performance in the same way across different groups (Dorans, 2004). A further issue in this context is that the hypothesis of differential prediction implies that the regression equations of X (test score) over Y (external criterion, for example, job performance) will differ across the groups considered. On the other hand, Jones and Raju (2000) and Stark et al. (2004) examined the influence of DIF items on interpretations based on cut-off scores, where these are used to make selection decisions with respect to an external criterion. Finally, research has also explored the effect that biased items can have on interpretations derived from total test scores, mainly those involving a comparison of means (Li & Zumbo, 2009), as well as the impact of DIF on latent trait estimation according to item response models (Wells, Subkoviak, & Serlin, 2002).

When working with cut-off scores it is also important to ensure that the assignment or classification of subjects is valid. In this regard, one needs to consider the presence of item invariance due to variables that are not relevant to what the test is seeking to measure, since DIF may have a detrimental effect on the meaning of test scores as well as on the measurement of the latent trait of interest (Roznowski & Reith, 1999). Given that clinical or educational diagnostic decisions will be based on total test scores rather than on individual item scores, it is also important, to evaluate the effect that the presence of DIF items may have on total scores as part of the process of obtaining validity evidence (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 1999). It should be remembered that in the context of criterion-referenced interpretation based on a cut-off score, this cut-off is established on the basis of total test scores, either by means of an empirical methods (e.g. ROC curves) or judgmental methods (Hambleton, 1990).

In sum, the cut-off score is a value on the scale of possible test scores that is used to classify subjects into two categories that reflect different levels of performance in terms of the skills, abilities, traits or domains measured by the test. This classification may be made on the basis of the estimated proportion of items of the evaluated domain which the subject answers correctly, or in relation to a criterion defined as a cut-off on the scale of possible test scores.

The aim of the present study was to determine the effects that DIF items can have when present in tests whose results are interpreted according to cut-off scores. To this end we designed a Monte Carlo simulation study that would enable us to examine the behaviour of the model or statistical test under a range of pre-established and controlled conditions. This simulation study considered two general conditions: 1) a bias- free test, that is, one in which no item manifests DIF, and 2) a test containing biased items, that is, with a variable number of items manifesting DIF. In the first condition we expected that regardless of the group to which the subject belonged (focal or reference, an example being men versus women), the number of subjects above the cut-off would depend on the test score obtained and would not be influenced by group membership; this would mean that the percentage of subjects above a given cut-off is the same in both groups (the selection rate is the same). In the second condition, where the test contains DIF items, we expected that while the number of subjects above the cut-off will still depend on the test score obtained, it will also be influenced by differential item functioning, such that the percentage of subjects above a given cut-off will not be the same in both groups (the selection rate is different). We also hypothesized that the difference in selection rate between the focal and reference groups will increase in line with the degree of test contamination, which in turn will depend on the amount of DIF simulated in the items and on the percentage of DIF items in the test. In the simulation study we also manipulated the sample size of the comparison groups.

From a practical point of view, understanding the effect that DIF items have on selection rates based on a cut-off score is of key interest for applied professionals, since as we have already noted selection decisions are based not on individual items but on total test scores.

Method

Experimental Conditions

The independent variables manipulated were sample size of the reference and focal groups, the amount of DIF and the percentage of DIF items in the test.

Sample size

Four combinations of sample sizes were used for the reference and focal groups: 100/100, 250/250, 500/500 and 1000/1000. These conditions reflect a range of situations that are plausible in practice, from small sample sizes (100/100) that are commonly found in clinical settings to larger samples (1000/1000) that one would expect to find in educational applications. For each of these sample sizes, two standardized normal ability distributions (μ=0, σ=1) were generated in the interval [-3, +3]. The means and standard deviations of the different ability distributions were the same for both the reference group (μ_R, σ_R) and the focal group (μ_F, σ_F) (non-impact condition), such that the difference between means of the two groups (reference and focal) was 0 (μ_d=μ_R-μ_F=0).

Amount of DIF

A single condition was established for the amount of DIF, defined as the difference between the difficulty parameters ($\mathbf{b_j}$) of items in the reference and focal groups. The manipulated difference was 0.8, which indicates that the magnitude of simulated DIF was high for each item (Li & Zumbo, 2009). As Li and Zumbo (2009) point out, items with high DIF would be expected to give rise to greater differences in responses to items between one group and the other, and consequently the combined effect of these items across the test as a whole would produce greater differences in total test scores; these artificial differences would be due to the presence of biased items.

Percentage of DIF items in the test

Another manipulated factor was the percentage of DIF items in the test, with five conditions being established: 0%, 10% (two DIF items in the test) 20% (four DIF items in the test), 30% (six DIF items in the test) and 40% (eight DIF items in the test). The condition of 10% of items manifesting DIF is common in the case of performance and aptitude tests; as Narayanan and Swaminathan (1994) point out, between 10% and 15% of items in these tests may show differential functioning. In adapted tests, however, the percentage of DIF is usually higher, slightly above 20% (Gierl, Gotzmann, & Boughton, 2004). Type I error rates may be affected by the proportion of DIF items in the test, insofar as the greater the number of DIF items in the test that favour one of the groups (e.g. the reference group) the more contaminated will be the total score for the other group (usually the focal group); consequently, there will be differences between the two groups in both the total test scores obtained and the percentage of subjects above the established cut-off. It should also be taken into account that for each test with simulated DIF the level of contamination of the matching variable was different. As Wang and Yeh (2003) point out, although the percentage of DIF items in a test is related to the degree of its contamination, it is the magnitude of test contamination that matters, and this magnitude varies in accordance with both the percentage of DIF items in the test and the amount of DIF in the item. Thus, two tests may have the same proportion of DIF items but different degrees of test contamination, depending on whether the level of DIF detected is high, moderate or low. With this in mind, Wang and Yeh (2003) proposed using the average signed area (ASA) as an index for estimating the amount of DIF in a test. The ASA for the Rasch (1-p) model is defined as:

$$ASA = \sum_{i=1}^{I} (b_{iF} - b_{iR}) / I$$

where I is the number of test items, b_{iF} is the difficulty parameter of item i in the focal group and b_{iR} is the difficulty parameter of the same item in the reference group. With respect to the conditions simulated in this study, the ASA values were as follows: 10% of DIF items, ASA = 0.08; 20% of DIF items, ASA = 0.16; 30% of DIF items, ASA = 0.24; and 40% of DIF items, ASA = 0.32.

The test size in the present study was set at 20 items, given that most scales and questionnaires used in the field of clinical and health assessment contain between 10 and 30 items. Examples of such tests include the NHP38 (Nottingham Health Profile; McDowell & Newell, 1996) and its brief version, the NHP20, which contains 20 dichotomous items (Prieto, Alonso, & Lamarca, 2003); the GDS-15 (Geriatric Depression Scale), comprising 15 dichotomous items (Marc, Rane, & Bruce, 2008); the CAST (Childhood Asperger Syndrome Test), which includes 31 dichotomous items (Scott, Cohen, Bolton, & Brayne, 2002); or scales for assessing depression, such as the 21-item BAS-D (Brief Assessment Scale for Depression; Allen et al., 1994). All of these tests use cut-off scores to diagnose individuals.

A total of 4 (sample sizes) x 5 (percentage of DIF items in the test) conditions were analysed, with 1000 replications being performed for each condition.

Generating the Data Matrices

Item responses were generated using the 1-p model (Rasch, 1980). The difficulty parameters were randomly selected from a normal distribution with mean zero and standard deviation of 1, following Paek (2010). The item parameters used to simulate the responses of the reference and focal groups are shown in table 1.

TABLE 1. *DIFFICULTY PARAMETERS FOR THE REFERENCE AND FOCAL GROUPS FOR EACH OF THE CONDITIONS MANIPULATED IN TERMS OF THE PERCENTAGE OF DIF ITEMS.*

		0% items with DIF	***10% items with DIF***	***20% items with DIF***	***30% items with DIF***	***40% items with DIF***
Item	b_R	b_F	b_F	b_F	b_F	b_F
1	-1.97	-1.97	-1.97	-1.97	-1.97	-1.97
2	-1.19	-1.19	-1.19	-1.19	-1.19	-1.19
3	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97
4	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67
5	-0.49	-0.49	-0.49	-0.49	-0.49	-0.49
6	-0.45	-0.45	-0.45	-0.45	-0.45	**0.35**
7	-0.32	-0.32	-0.32	-0.32	**0.48**	**0.48**
8	-0.03	-0.03	-0.03	**0.77**	**0.77**	**0.77**
9	0.03	0.03	**0.83**	**0.83**	**0.83**	**0.83**
10	0.06	0.06	**0.86**	**0.86**	**0.86**	**0.86**
11	0.31	0.31	0.31	**1.11**	**1.11**	**1.11**
12	0.40	0.40	0.40	0.40	**1.20**	**1.20**
13	0.41	0.41	0.41	0.41	0.41	**1.21**
14	0.56	0.56	0.56	0.56	0.56	0.56
15	0.79	0.79	0.79	0.79	0.79	0.79
16	0.87	0.87	0.87	0.87	0.87	0.87
17	1.06	1.06	1.06	1.06	1.06	1.06
18	1.08	1.08	1.08	1.08	1.08	1.08
19	1.56	1.56	1.56	1.56	1.56	1.56
20	2.06	2.06	2.06	2.06	2.06	2.06

Note: DIF items are shown in bold.

The procedure for data simulation involved two steps: 1) subject ability was randomly generated according to a standardized normal distribution using the *R* program; and 2) item responses were generated independently for each group (focal and reference) using a program written by the authors and which implemented the procedure described by Hambleton and Cook (1983) for dichotomous items, the input being subject parameters and item parameters.

Data Analysis

The effect of the presence of DIF items on the selection rate was analysed by studying the differences in selection rates (according to the cut-off) for subjects in the reference group and subjects in the focal group. This was done by applying the *Z* test for independent samples (reference group vs. focal group) to the difference in the proportions of subjects above the cut-off. We also calculated measures of effect size: ln odds ratio (*ln-OR*) and risk ratio (*RR*). The ln-OR was obtained using the following expression:

$$Ln - OR = Ln\left(\frac{Odds - GR}{Odds - GF}\right)$$

where

$$Odds - GR = \frac{P_R}{1 - P_R}$$

and

$$Odds - GF = \frac{P_F}{1 - P_F}$$

with *P* being the proportion of subjects in a group (focal or reference) that are above the cut-off. According to the criteria of Cohen (1988), an *OR* of 1.44 indicates a small effect size, a value of 2.47 a moderate effect size and a value of 4.25 a strong effect size, which for the *ln-OR* statistic are equivalent to values of 0.3646, 0.9042 and 1.45, respectively.

The RR was obtained using the following expression:

$$RR = \frac{P_R}{P_F}$$

Consequently, a total of five statistics were computed: a) Type I error rate of the *Z* statistic (proportion of times among 1000 replications that the null hypothesis was rejected at the nominal level of .05) in each of the manipulated conditions, b) mean and standard deviation for the values of *ln-OR* across the replications, and c) mean and standard deviation for the values of *RR* across the replications. Two cut-off points were established: 1) domain score corresponding to 50% of the items ($X_c = 10$), and 2) domain score corresponding to 70% of the items ($X_c =$ 14).

We also considered the effect of DIF on total test scores by using the Student's *t* test for independent samples to analyse differences between the mean scores of the two groups. In addition, we calculated the standardized mean in order to assess the effect size of the observed differences. The standardized mean difference was calculated using the following expression:

$$d = \frac{\overline{X}_R - \overline{X}_F}{s}$$

where

$$s = \sqrt{\frac{(n_R - 1)S_R^2 + (n_F - 1)S_F^2}{n_R + n_F - 2}}$$

In this case, three statistics were computed: a) Type I error rate of the Student's *t* test (proportion of times among 1000 replications that the null hypothesis was rejected at the nominal level of .05) in each of the manipulated conditions, b) mean of the *d* index (standardized mean difference) and c) standard deviation of the *d* index.

All data analyses were performed using the *R* program (2012), and all statistical contrasts were one-tailed.

Results

In the no-DIF conditions the Type I error rate was well controlled. All Type I error rates were close to the nominal value of .05, regardless of the sample size and the cut-off point used (see table 2). Similar results were obtained when comparing the means of the two groups using the Student's *t* test.

TABLE 2. *TYPE I ERROR RATES FOR THE Z AND T TESTS AND MEANS AND STANDARD DEVIATIONS FOR THE EFFECT SIZE INDICES IN EACH OF THE MANIPULATED CONDITIONS.*

					50% cut-off, X_c= 10					***70% cut-off, X_c=14***				
			d			*ln-OR*		*RR*			*ln-OR*		*RR*	
Sample size	***Items with DIF***	t	Mean	SD	Z	Mean	SD	Mean	SD	Z	Mean	SD	Mean	SD
100/100	0	0.040	0.000	0.140	0.053	-0.014	0.306	1.013	0.225	0.020	0.027	0.813	1.267	1.223
	2	0.165	0.150	0.141	0.130	0.295	0.316	1.273	0.302	0.027	0.109	0.806	1.417	1.263
	4	0.551	0.310	0.147	0.376	0.585	0.344	1.613	0.441	0.036	0.246	0.801	1.605	1.358
	6	0.872	0.460	0.149	0.638	0.834	0.360	1.994	0.612	0.040	0.439	0.785	1.902	1.471
	8	0.970	0.559	0.146	0.873	1.135	0.384	2.619	0.930	0.047	0.470	0.798	1.967	1.567
250/250	0	0.050	0.000	0.093	0.044	-0.019	0.198	0.996	0.137	0.045	0.027	0.628	1.175	0.862
	2	0.380	0.150	0.087	0.288	0.309	0.202	1.267	0.188	0.044	0.136	0.632	1.400	1.089
	4	0.936	0.320	0.092	0.782	0.604	0.215	1.600	0.267	0.082	0.420	0.694	1.910	1.520
	6	1.000	0.447	0.094	0.964	0.821	0.224	1.918	0.348	0.124	0.618	0.707	2.351	1.947
	8	1.000	0.558	0.089	0.999	1.115	0.235	2.481	0.511	0.151	0.701	0.688	2.513	1.945
500/500	0	0.041	0.000	0.063	0.037	-0.013	0.136	0.995	0.094	0.041	0.001	0.407	1.084	0.487
	2	0.607	0.150	0.062	0.499	0.291	0.138	1.243	0.126	0.044	0.130	0.552	1.235	0.552
	4	0.997	0.310	0.061	0.978	0.590	0.147	1.572	0.177	0.148	0.421	0.480	1.705	1.090
	6	1.000	0.440	0.065	1.000	0.806	0.156	2.877	0.233	0.232	0.626	0.502	2.111	1.384
	8	1.000	0.559	0.067	1.000	1.108	0.168	2.440	0.345	0.285	0.691	0.494	2.233	1.319
1000/1000	0	0.040	0.000	0.043	0.039	0.000	0.095	1.000	0.067	0.035	-0.019	0.273	1.017	0.278
	2	0.918	0.150	0.046	0.841	0.304	0.098	1.251	0.090	0.061	0.1403	0.295	1.195	0.354
	4	1.000	0.314	0.045	1.000	0.590	0.104	1.565	0.125	0.272	0.434	0.328	1.612	0.570
	6	1.000	0.444	0.047	1.000	0.812	0.111	1.881	0.166	0.429	0.590	0.335	1.882	0.652
	8	1.000	0.375	0.705	1.000	1.099	0.118	2.410	0.235	0.513	0.689	0.351	2.090	0.799

Note: t: Student's *t* test; *d*: standardized mean difference; *Z*: *Z* test for difference in proportions; *ln-OR*: ln(odds ratio); *RR*: risk ratio; *SD*: standard deviation; X_c: cut-off score on the test.

In DIF conditions as the degree of test contamination increased (i.e. a greater number of items manifested DIF), greater differences between the reference and focal groups were erroneously detected, both as regards selection rates (percentage of subjects above the cut-off in each group) and mean scores. This effect was more marked when working with the 50% cut-off, as compared with the 70% criterion. It should be noted, however, that the data were simulated following a normal distribution with mean zero and standard deviation 1, and that the percentage of subject above the 70% cut-off was very small in both groups (focal and reference).

As expected, the Type I error rate was higher in the conditions with larger sample sizes. With samples of 2000 subjects (NR = NF = 1000) and only 10% of items with DIF, statistically significant differences were observed between the two groups (Type I error rate above the nominal level) on both the test of differences between means and the test of differences in proportions; in the latter case independently of the cut-off score used.

It is important, however, to interpret the results described above in conjunction with their corresponding effect sizes. In the DIF conditions and with respect to the differences in proportions (selection rate) between groups, a large effect size was not observed in any of the situations considered (cut-off, amount of DIF in the test and sample size). Applying the criteria of Cohen (1988) an effect size was regarded as small when the value of *ln-OR* was 0.3646, moderate when this value was 0.9042 and high when it was 1.45. In the worst case observed the average effect size had only a moderate magnitude, although the effect size did increase (table 2) in line with greater test contamination.

The same trend was observed when examining the effect sizes for differences between means. On average, effects were moderate in situations with a higher percentage of DIF items, and small with smaller degrees of test contamination.

Discussion

The aim of this study was to examine the impact that items showing differential functioning may have on test interpretations based on cut-off scores, in other words, on the selection rates of the groups. As noted by Li and Zumbo (2009), relatively little is known about the impact of DIF on subsequent statistical conclusions when the total test score is used in data analyses. Moreover, even less is known about the effects of using this contaminated score as a variable in the corresponding hypothesis test.

The results of this study indicate that as the degree of test contamination increases, greater differences between the reference and focal groups are erroneously detected, both as regards selection rates (percentage of subjects above the cut-off in each group) and the mean score. In addition, a higher Type I error rate was found in conditions with a larger sample size. Under these sample conditions and with just 10% of items with DIF the Type I error rate was above the nominal level for both the test of differences between means and the test of differences in proportions. These results are in line with those reported by Li and Zumbo (2009) for the effect on the differences between means, and highlight the need to assess DIF in a test prior to conducting any subsequent analysis. Another study by Zumbo (2003), based on multigroup confirmatory analysis, found that the presence of DIF items did not have important consequences at the test level, since both the factor loadings and the error variance were statistically equivalent between the reference and focal groups, even when the percentage of DIF items in the test was high. Nevertheless, Zumbo (2003) points out that the presence of DIF items reduces the validity of total test scores for any application of interest, since it introduces a systematic bias into these scores and limits their usefulness.

In sum, the presence of DIF items may undermine the validity of a test (Li & Zumbo, 2009), and it should therefore be taken into account in tests used for clinical or educational diagnosis, due to the consequences that derive from the use of results of this kind. In educational assessments, for example, cut-offs are used for various purposes including the classification of students and for deciding whether they fulfil the requirements to move to a higher level. In this regard, the use of these scores has repercussions not only for students but also for teachers and the educational institutions in question. Other contexts in which cut-off scores are used include clinical diagnosis, career promotion and the certification of competencies, as well as in research as a way of establishing subgroups. Clearly, then, the use of cut-off scores has an important scope and impact, and evaluating the validity of such scores is therefore a key step in the process of test validation (Davis-Becker & Buckendahl, 2013; Sireci, Hauger, Wells, Shea, & Zenisky, 2009). In this regard, obtaining evidence of an absence of DIF among a test's items is one way of supporting the internal and external validity of cut-off scores.

One limitation of the present study concerns the fact that the simulated tests fitted the Rasch model. Although in this case the observed total score on the test is a sufficient estimator of ability or the latent trait (DeMars, 2008), this does not hold for multiparametric item response models, where the items may vary in their discrimination parameter. Thus, when items fit the 2-p or 3-p model it should not be assumed that subjects ordered according to the observed total score will be ordered according to the expected score on the latent trait (DeMars, 2008). For these models, therefore, it is necessary to consider the cut-off not only with respect to the observed score but also with respect to the level of the latent trait.

A further limitation is that the study design assumes that the null hypothesis of no differences in selection rates between the reference and focal groups is true, the objective being to evaluate the extent to which this null hypothesis is rejected when it really is true, and to observe if the Type I error rate is higher than expected according to the level of significance established in the test used to compare proportions between groups (focal and reference). However, in future studies it would be necessary to consider not only the Type I error rate but also the Type II error rate, and also to evaluate not only potential differences in selection rates between the focal and reference groups but also whether these rates are as expected according to the ability distribution of the groups and the cut-offs used.

Further research is required to extend the scope of the results obtained here. For example, one could examine other cut-offs, the effect of impact between groups, different sample sizes for the reference and focal groups, conditions in which DIF is cancelled out at the test level, and different magnitudes of DIF. It would also be useful to extend the study to other item response models, covering both dichotomous and polytomous items, so as to determine the corresponding effect when working in these conditions.

REFERENCES

Allen, N., Ames, D., Ashby, D., Bennetts, K., Tuckwell, V., & West, C. (1994). A brief sensitive screening instrument for depression in late life. *Age and Ageing,* 23: 218.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W.H. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: LEA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Davis-Becker, S.L., & Buckendahl, C.W. (2013). Identifying and evaluating external validity evidence for passing scores. *International Journal of Testing, 13*, 50-64.

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.

Dorans, N.J. (2004) Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41,* 43-68.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92,* 526-531.

Gierl, M.J., Gotzmann, A., & Boughton, K.A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*, 241-264.

Hambleton, R. K. (1990). Criterion-referenced testing methods and practices. In T. Gutkin & C. Reynolds (Eds.), *Handbook of school psychology* (2nd ed.; pp. 388-414). New York: Wiley.

Hambleton, R.K., & Cook, L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Hidalgo, M.D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd edition), vol. 4, pp. 36-44. USA: Elsevier - Science & Technology.

Jones, J.A., & Raju, N.S. (2000). *Differential item and test functioning and cutoff scores in personnel decision making*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30*, 343-370.

Linn, R.L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33-47.

Marc, L.G., Rane, P.J., & Bruce, M.L. (2008). Screening performance of the Geriatric Depression Scale (GDS-15) in a diverse elderly home care population. *American Journal of Geriatric Psychiatry, 16,* 914-921.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires (2nd edition)*. New York: Oxford University Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

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

Millsap, R.E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias*. Applied Psychological Measurement, 16*, 389-402.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd edition). Thousand Oaks, California: Sage Publications, Inc.

Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH Chi-square test in DIF applications. *Applied Psychological Measurement, 34,* 539-548.

Penfield, R.D., & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5-15.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes, 1:27.* http://www.hqlo.com/content/1/1/27.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment test.* Chicago: University Chicago Press.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248-269.

Scott, F.J., Cohen, S.B., Boston, P., & Brayne, C. (2002). The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school age children. *Autism, 6*, 9-31.

Sireci, S., Hauger, J., Wells, C., Shea, C., & Zenisky, A. (2009). Evaluation of the standard setting on the 2005 grade 12 national assessment of educational progress mathematics test. *Applied Measurement in Education, 22*, 339-358.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*, 170-187.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89, 497*-508.

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, 479-499.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wu, A.D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research & Evaluation, 12*(3). Retrieved March 5, 2013 from: http://pareonline.net/getvn.asp?v=12&n=3

Zumbo, B.D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20,* 136-147.