Psicothema 2010. Vol. 22, nº 4, pp. 1018-1025
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Use of multilevel logistic regression to identify the causes of differential item functioning

Nekane Balluerka, Arantxa Gorostiaga, Juana Gómez-Benito*
and María Dolores Hidalgo**
Universidad del País Vasco, * Universidad de Barcelona and ** Universidad de Murcia

Given that a key function of tests is to serve as evaluation instruments and for decision making in the fields of psychology and education, the possibility that some of their items may show differential behaviour is a major concern for psychometricians. In recent decades, important progress has been made as regards the efficacy of techniques designed to detect this differential item functioning (DIF). However, the findings are scant when it comes to explaining its causes. The present study addresses this problem from the perspective of multilevel analysis. Starting from a case study in the area of transcultural comparisons, multilevel logistic regression is used: 1) to identify the item characteristics associated with the presence of DIF; 2) to estimate the proportion of variation in the DIF coefficients that is explained by these characteristics; and 3) to evaluate alternative explanations of the DIF by comparing the explanatory power or fit of different sequential models. The comparison of these models confirmed one of the two alternatives (familiarity with the stimulus) and rejected the other (the topic area) as being a cause of differential functioning with respect to the compared groups.

Utilización de la regresión logística multinivel para identificar las causas del funcionamiento diferencial de los ítems. Dada la relevancia de los tests como instrumentos de evaluación y de toma de decisiones en los campos de la psicología y de la educación, la posibilidad de que algunos de sus ítems presenten un comportamiento diferencial constituye una preocupación central de los psicómetras. En las últimas décadas se han producido importantes avances con respecto a las técnicas diseñadas para detectar el funcionamiento diferencial de los ítems (DIF). Sin embargo, los hallazgos son escasos en lo que respecta a identificar las causas que lo explican. El presente trabajo aborda este problema desde la perspectiva del análisis multinivel. Partiendo del estudio de un caso del ámbito de las comparaciones transculturales, se utiliza la regresión logística multinivel para: 1) identificar las características de los ítems asociadas a la presencia de DIF; 2) estimar la proporción de la variación en los coeficientes de DIF explicada por tales características; y 3) evaluar explicaciones alternativas para el DIF comparando la capacidad explicativa o el ajuste de diferentes modelos. La comparación entre tales modelos permitió confirmar una de las dos alternativas (la familiaridad con el estímulo) y descartar la otra (el tema de estudio) como causa del funcionamiento diferencial de los ítems en los grupos comparados.

Since, in the USA, the first questionnaires were developed for personnel selection among company employees, university students or soldiers, standardised tests have played a key role as evaluation instruments, especially in the fields of psychology and education. This notable increase in the use of tests for decision-making purposes has meant that the potential differential item functioning (DIF) with respect to variables of no relevance to the construct being tested has become a central concern when assessing the validity of psychometric instruments.

Thus, in recent decades, numerous statistical techniques have been developed to analyse this phenomenon (see reviews in Clauser & Mazor, 1998; Gómez-Benito & Hidalgo, 1997; Hidalgo & Gómez-Benito, 1999, 2010; Millsap & Everson, 1993; Osterling & Everson, 2009; Potenza & Dorans, 1995; Roussos & Stout, 2004; Zumbo, 2007). Most of the proposed methods have focused on detecting DIF in dichotomous items, using either procedures based on item response theory (Lord, 1980; Ordóñez & Romero, 2007; Raju, 1990; Thissen, Steinberg, & Wainer, 1993) or those derived from the analysis of contingency tables and/or regression models (Holland & Thayer, 1988; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). More recently, and due to the development of new evaluation instruments in the educational context (portfolio-based assessment, authentic assessment, etc.) and the need to ensure measurement invariance in psychological assessment and transcultural comparison studies, techniques for detecting DIF have also been adapted to polytomous items (Cohen, Kim, & Baker, 1993; Flowers, Oshima, & Raju, 1999; French & Miller, 1996; Hidalgo & Gómez-Benito, 2000; Kim & Cohen, 1998; Miller & Spray, 1993; Thissen, 2001; Welch & Hoover,

Fecha recepción: 14-8-09 • Fecha aceptación: 17-1-10
Correspondencia: Nekane Balluerka Lasa
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
e-mail: nekane.balluerka@ehu.es

1993; Zumbo, 1999; Zwick, Donoghue, & Grima, 1993; Zwick & Thayer, 1996). This wide range of proposed statistical techniques and of studies designed to test their efficacy (power and Type I error rate) in detecting DIF (e.g. Kristjansson, Aylesworth, McDowell, & Zumbo 2005; Finch, 2005; Kim, Cohen, Alagoz, & Kim, 2007; Su & Wang 2005) constitutes what Zumbo (2007) has called the *second generation of DIF*, characterised by the development of sophisticated statistical models for detecting and classifying DIF; however, these models remain unable to explain its causes. Thus, Kim et al. (2007) point out that neither the models for detecting DIF nor their corresponding measures of effect size provide clues as to why DIF occurs. In general, little progress has been made in this regard (Ferne & Rupp, 2007; Padilla, Pérez, & González, 1998; Zumbo & Gelin, 2005).

One proposed explanation for why DIF occurs is based on the principle of multidimensionality (Ackerman, 1992). From this perspective an item presents DIF because some of its characteristics are not relevant to the trait or latent ability of interest. The research by Stout and co-workers, who proposed the procedures known as SIBTEST (Shealy & Stout, 1993a, 1993b) and POLYSIBTEST (Chang, Mazzeo, & Roussos, 1996) falls within this theoretical framework. More recently, other investigators have applied the structural equation models known as multiple indicators multiple causes (MIMIC) and proposed by Muthén (1989) as a way not only of detecting DIF (Gelin & Zumbo, 2007; Shih & Wang, 2009) but also of explaining its causes (Zumbo & Gelin, 2005). These models are able to investigate whether the characteristics and content of an item exert an influence on the test's behaviour, as well as whether variables from the individual's social and/or psychological context help to explain why an item functions differentially. Multilevel models (Bryk & Raudenbush, 1992; Goldstein, 1987; Snijders & Bosker, 1999) can also be considered from this perspective since they are able to combine the results obtained in a series of logistic regression analyses performed with the items, the aim being to identify consistent sources of DIF and compare them with one another, selecting the model that best explains the observed variation.

The present study focuses specifically on this latter line of research. To this end, a case study is analysed following the approach proposed by Swanson, Clauser, Case, Nungester and Featherman (2002), which formulates a set of two-level models that enable the progressive incorporation of item characteristics so as to explain the variation in item responses that is due to DIF. The level-1 models (subject level) are logistic regression models for the analysis of DIF which are similar to those proposed by Swaminathan and Rogers (1990). In the level-2 models (item level) the regression coefficients from the level-1 models, which include the coefficient that represents each item's DIF, are treated as random variables whose variation could be predicted by certain characteristics of the items. This approach is therefore able to: 1) identify the item characteristics that are associated with the presence of DIF; 2) estimate the proportion of variation in the DIF coefficients that is explained by these characteristics; and 3) evaluate alternative explanations of the DIF by comparing the explanatory power or fit of different models. One of the most important features that distinguishes this approach from traditional procedures for detecting DIF is that it formulates DIF as a random parameter, which in addition to optimising its estimation, enables information to be obtained regarding its causes. As De Boeck (2008) states, although random item parameters are uncommon and their application requires further study, they do make sense theoretically. Furthermore, in accordance with the stance taken in the present study, De Boeck shows that in practice the random item approach is useful for dealing with several issues, one of them being troubleshooting with respect to DIF.

Multilevel logistic regression models for analysing DIF

When the dependent variable is not continuous or does not follow a normal distribution (as in the case of binary variables, proportions, count variables and ordinal variables) the fit of the data from the multilevel approach is tackled by means of an extension of the basic hierarchical linear model: the generalised hierarchical linear model. This is an adaptation of the generalised linear model developed by McCullagh and Nelder (1989) for the analysis of hierarchical data and requires the transformation of the dependent variable and the error distribution.

In the data used here to examine the causes of DIF the dependent variable is binary or dichotomous; it is a variable whose limits are 0 and 1 and which does not follow a normal distribution. In such cases it is necessary to assume that the underlying probability distribution takes the binomial form (or a special case of the binomial distribution, such as that of Bernouilli) and that it has a mean of μ. The estimator of μ is p and is interpreted as the likelihood of a given event occurring. A typical transformation for a binomial model is the *logit* transformation:

$$\log\text{it}(p) = 1\text{n}\left(\frac{p}{1\text{-}p}\right) \tag{1}$$

The *logit* of *p* has no limit and the density of *logit (p)* approaches a normal distribution.

In the context of generalised hierarchical linear models this type of transformation is termed *logit function*. Thus, the *logit* is a linking function that establishes a relationship between the untransformed dependent variable Y_{ij} (in our case, the score obtained by subject *i* on item *j*) and the transformed variable η_{ij}, ensuring that the predictions are located within a given interval of values.

Thus, it is possible to construct a level-1 prediction model in order to associate the transformed predicted value with a set of predictive *Q* variables:

$$\eta_{ij}= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \dots + \beta_{Qj}X_{Qij} \tag{2}$$

Note that there is no term for the level-1 error variance, since in binary variables the variance is completely determined by the mean.

When logistic regression is used in the framework of generalised linear models to analyse the characteristics of items that may generate DIF, the level-1 model (subject level) is given by the following expression:

$$\text{logit}[\text{Prob}(\text{Y}_{\text{ij}}= 1)]= \beta_{0j} + \beta_{1j} * H_i + \beta_{2j} * G_i \tag{3}$$

where Y_{ij} is the score obtained by subject *i* on item *j* (1= correct response, 0= incorrect response); H_i indicates the ability level of subject *i* on the attribute or variable measured by the test; G_i is a dummy variable that indicates whether a person belongs to the group of interest or focal group (G_i= 1), or to a group with which the latter is compared, i.e. the reference group ($G_i = 0$); β_{0j} reflects the *log* of the odds (log-odds) for the item difficulty in the reference group; β_{1j} stands for the item discrimination (in this model the same value has been established in the reference and focal groups) or the ability of the item to discriminate between subjects with high and low scores on the attribute measured by the test; and β_{2j} denotes the deviation in the item difficulty in the focal group with respect to the reference group, in other words, the parameter of uniform DIF.

As pointed out by Swanson et al. (2002), more complex models can be considered by adding an interaction term between ability and group, such that both item discrimination and item difficulty can vary between the focal and reference groups, thus enabling non-uniform DIF to be modelled. Additionally, more than two groups can be compared by using multiple dummy codes to represent the corresponding group membership.

Returning to the case in question, in the level-2 model (item level) the coefficients associated with the intercept and the slope are formulated as random variables whose variation can be predicted by certain item characteristics:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} * I_1 + \gamma_{22} * I_2 + \ldots + \gamma_{2n} * I_n + u_{2j} \end{aligned} \tag{4}$$

where γ_{Q0}-s are the means of the level-1 regression coefficients. Specifically, γ_{00} is the mean of the item difficulty values in the reference group; γ_{10} is the mean of the item discrimination values; and γ_{20} is the mean of the deviation in item difficulty values between the focal and reference groups or the overall DIF parameter for item *j*. The u_{Qj}-s are random variables that represent unexplained variability. Specifically, u_{0j} is the variability shown by items in terms of level of difficulty, u_{1j} is the variability among items as regards the level of discrimination, and u_{2j} denotes the variability among items in the DIF index or the unexplained variation in DIF for item *j* after taking into consideration its characteristics. The variables $I_1, \ldots, I_n$, are dummy or interval variables that reflect item characteristics. γ_{2n} is the last parameter or the coefficient associated with the *n*-th characteristic of the item that predicts the variation in DIF.

Case study

In order to illustrate, by means of a case study, the utility of multilevel logistic regression for analysing the causes of DIF, we generated a matrix of 815 school pupils belonging to two nationalities: Spanish nationals and Moroccan immigrants to Spain. The generated sample size for the reference group (510 Spanish) was different from that of the focal group (305 Moroccan). The data generated comprised the responses given (correct: value of 1; incorrect: value of 0) by these pupils to each of the 36 items on an aptitude test. This test length was selected because most of the scales and questionnaires in psychological assessment include between 20 and 40 items. Item responses were simulated using the two-parameter logistic item response model. The ability of subjects was generated at random, following a normal distribution. Since in applied research the focal group usually underperforms in comparison to the reference group, the ability of this latter group followed a normal standardised distribution (0, 1), while the distribution for the focal group showed values of -1.5 and 1 for the mean and standard deviation, respectively. Another factor manipulated was the proportion of items showing DIF, which was set at 25% (nine items with DIF). The amount of DIF was simulated by varying the difficulty parameter (uniform DIF) by 1.00 and by keeping the discrimination parameters the same for the two groups. The DIF in the nine items favoured the reference group. The difficulty and discrimination parameters for the reference group are shown in Appendix 1.

So as to conduct the analysis using HLM6 software (Raudenbush, Bryk, Cheong, & Congdon, 2004), two data files were generated. In order to generate the level-1 file (subjects level) it was necessary to restructure the variables of the matrix such that the subjects were nested in the items. The resulting matrix comprised 29340 cases (815 subjects × 36 items).

The level-2 file (items level) included one variable referring to the number of the item (with values from 1 to 36) and two variables referring to the item characteristics that could prove useful for predicting DIF. The first of these was a quantitative variable referring to the mean level of linguistic familiarity of the item, evaluated on a 20-point Likert scale (1: No familiarity; 20: Very high familiarity). This variable was centered with respect to the overall mean. The second variable reflected the topic addressed by the item and included three categories: physics (17 items), natural sciences (13 items) and history (6 items). In order to be able to include this categorical variable in the model, two dummy variables were generated, taking the topic area ‘physics’ as the reference category.

As regards the models of analysis we firstly proposed a model of random coefficients such as that shown in equation 3, the aim being to obtain the estimators of the variance of the regression coefficients associated with the intercept, the ability level (total score on the aptitude test) and group membership (nationality of subjects). A series of models of random intercepts and random slopes was subsequently fitted to the data, similar to those shown in equation 4 and which differed from one another in the characteristic (or characteristics) of the item used in each case to predict DIF.

Model of random coefficients

As already pointed out, the estimators of subject ability were transformed into a standard scale and the variable ‘nationality’ was centered over the overall mean. In this way the level-1 intercepts can be interpreted as the *logit* of the likelihood of obtaining a correct response when the subject’s ability level takes the value of 0. Table 1 shows the results obtained when fitting the model of random coefficients (equation 3) to the data. As can be seen in the section corresponding to the fixed effects, the mean intercept for all the items was 0.476, the mean discrimination index of the items was 0.86 and the mean index of DIF associated with nationality was -0.567. Given that Spanish nationals were considered as the reference group this latter value illustrates that, after controlling for the effect of ability, Moroccan pupils obtained worse test results than did their Spanish counterparts.

The section of random effects provides information about the variability among items in the regression coefficients, taking into consideration the error in the estimation of these coefficients. The estimated variance components for the intercept and for the regression coefficient associated with ability level were 1.297 and 0.075, respectively (the square roots of these values reflect the standard deviation of the intercepts and of the ability coefficients).

The estimated variance component for the index of DIF was 0.204 (the square root of this value represents the standard deviation among items for the DIF indices).

Models of random intercepts and random slopes

It can be supposed that one of the possible causes of the DIF associated with nationality is that the Spanish terms used in the item wording have low linguistic familiarity for immigrant subjects (for example, the Moroccan pupils). In order to examine this question we considered the mean level of linguistic familiarity of the item, in the second level, as a predictive variable that could explain the variability in the level-1 coefficients associated with DIF. The results obtained are shown in Table 2. The regression coefficient corresponding to the mean level of linguistic familiarity of the item was 0.489, indicating that the Moroccan pupils performed better on those items that were more familiar to them, in other words, their performance worsened as the level of linguistic familiarity of the terms included in the item decreased. Specifically, an increase of one unit on the scale referring to the linguistic familiarity of the item produces an increment of 0.489 in the log-odds of a correct response being given by Moroccans. The change produced in the variance component associated with nationality when incorporating into the model as a predictive variable the mean level of linguistic familiarity of the item provides an index of effect size that is similar to R^2. In Table 1 this variance component

Table 1
Results obtained when fitting a model of random coefficients to the data

Fixed effects	Regression coefficient	Standard error	t	df	p	Interpretation
Intercept	0.476	0.188	2.535	35	0.016	Mean of the intercepts (mean of the log-odds of a correct response for subjects with an ability level of 0)
Ability	0.860	0.048	17.912	35	0.001	Mean increment in the log-odds of a correct response associated with an increase of one standard deviation in ability level
Nationality (dummy)	-0.567	0.182	-3.115	35	0.004	Mean increment in the log-odds of a correct response for Moroccan subjects (overall DIF parameter)
Random effects	**Standard deviation**	**Variance component**	**df**	**Chi squared**	**p**	**Interpretation**
Intercept	1.140	1.297	35	4381.576	0.001	Variability among items in the intercepts
Ability	0.274	0.075	35	329.638	0.001	Variability among items in the coefficients associated with ability
Nationality (dummy)	0.451	0.204	35	257.216	0.001	Variability among items in the coefficients associated with DIF

Note: The ability estimators were transformed such that their mean and SD take the values 0 and 1, respectively. The dummy variable 'nationality' was centered over the total mean. In this way the intercepts can be interpreted as the log-odds of a correct response when the ability level of the subjects is equal to 0

Table 2
Results obtained when fitting a model of random intercepts and random slopes to the data using the mean level of linguistic familiarity of the items, the aim being to predict the DIF coefficients

Fixed effects	Regression coefficient	Standard error	t	df	p	Interpretation
Intercept	0.476	0.188	2.535	35	0.016	Mean of the intercepts (mean of the log-odds of a correct response for subjects with an ability level of 0)
Ability	0.860	0.048	17.913	35	0.001	Mean increment in the log-odds of a correct response associated with an increase of one standard deviation in ability level
Nationality (dummy)	-0.567	0.182	-3.115	34	0.004	Mean increment in the log-odds of a correct response for Moroccan subjects (overall DIF parameter)
Linguistic familiarity	0.489	0.175	2.794	34	0.008	Change in the log-odds of a correct response in Moroccan subjects for each increment of one unit in the mean level of the linguistic familiarity of the item
Random effects	**Standard deviation**	**Variance component**	**df**	**Chi squared**	**p**	**Interpretation**
Intercept	1.140	1.3	35	4381.58	0.001	Variability among items in the intercepts
Ability	0.274	0.075	35	329.64	0.001	Variability among items in the coefficients associated with ability
Nationality (dummy)	0.397	0.158	34	327.762	0.001	Variability among items in the coefficients associated with DIF

has a value of 0.204, whereas in Table 2 its value is 0.158. If we divide the change produced in the variance component between the two models (0.204-0.158= 0.046) by the initial value (0.204) it can be concluded that the level of linguistic familiarity of the item explains 22.55% of the variance in the DIF coefficients.

A second potential cause of the DIF associated with nationality could be that the performance of Spanish and Moroccan pupils varies according to the topic area referred to by the items. As pointed out earlier, this possibility was examined by using items that referred to three different topic areas (physics, natural sciences

Table 3
Results obtained when fitting a model of random intercepts and random slopes to the data using the topic area to which they refer, the aim being to predict the DIF coefficients

Fixed effects	Regression coefficient	Standard error	t	df	p	Interpretation
Intercept	0.476	0.188	2.536	35	0.016	Mean of the intercepts (mean of the log-odds of a correct response for subjects with an ability level of 0)
Ability	0.860	0.048	17.904	35	0.001	Mean increment in the log-odds of a correct response associated with an increase of one standard deviation in ability level
Nationality (dummy)	-0.611	0.181	-3.371	33	0.002	Mean increment in the log-odds of a correct response for Moroccan subjects (overall DIF parameter)
Natural sciences (dummy)	0.203	0.208	0.974	33	0.338	Change in the log-odds of a correct response in Moroccan subjects when the items refer to natural sciences
History (dummy)	0.208	0.215	0.969	33	0.340	Change in the log-odds of a correct response in Moroccan subjects when the items refer to history
Random effects	**Standard deviation**	**Variance component**	**df**	**Chi squared**	**p**	**Interpretation**
Intercept	1.139	1.298	35	4379.977	0.001	Variability among items in the intercepts
Ability	0.274	0.075	35	329.682	0.001	Variability among items in the coefficients associated with ability
Nationality (dummy)	0.396	0.157	33	327.751	0.001	Variability among items in the coefficients associated with DIF

Table 4
Results obtained when fitting a model of random intercepts and random slopes to the data using the mean level of linguistic familiarity of the items and the topic area to which they refer, the aim being to predict the DIF coefficients

Fixed effects	Regression coefficient	Standard error	t	df	p	Interpretation
Intercept	0.476	0.188	2.535	35	0.016	Mean of the intercepts (mean of the log-odds of a correct response for subjects with an ability level of 0)
Ability	0.860	0.048	17.902	35	0.001	Mean increment in the log-odds of a correct response associated with an increase of one standard deviation in ability level
Nationality (dummy)	-0.599	0.184	-3.255	32	0.003	Mean increment in the log-odds of a correct response for Moroccan subjects (overall DIF parameter)
Linguistic familiarity	0.462	0.173	2.670	32	0.011	Change in the log-odds of a correct response in Moroccan subjects for each increment of one unit in the mean level of the linguistic familiarity of the item
Natural sciences (dummy)	0.187	0.172	1.088	32	0.285	Change in the log-odds of a correct response in Moroccan subjects when the items refer to natural sciences
History (dummy)	0.168	0.195	0.867	32	0.392	Change in the log-odds of a correct response in Moroccan subjects when the items refer to history
Random effects	**Standard deviation**	**Variance component**	**df**	**Chi squared**	**p**	**Interpretation**
Intercept	1.140	1.30	35	4381.58	0.001	Variability among items in the intercepts
Ability	0.274	0.075	35	329.748	0.001	Variability among items in the coefficients associated with ability
Nationality (dummy)	0.407	0.166	32	329.541	0.001	Variability among items in the coefficients associated with DIF

and history), creating two dummy variables and taking 'physics' as the reference category. These variables were used as level-2 predictive variables in order to explain the variability in the level-1 coefficients associated with DIF. The results obtained (see Table 3) show that, after controlling for ability level, the log-odds of a correct response being given by Moroccan subjects increased slightly when the items referred to natural sciences or history. However, none of the coefficients was statistically significant. According to these data there are barely any differences in the variance components associated with nationality in Tables 2 and 3, which demonstrates that the topic area to which the items refer does not explain the variance among the DIF coefficients.

The third model of random intercepts and random slopes, which was developed in order to explain the DIF associated with nationality as a function of certain characteristics of the items, included the two variables considered as predictors in the previous models. The results obtained when fitting this model to the data are shown in Table 4. As in the previous two models (Tables 2 and 3) the regression coefficient associated with the level of linguistic familiarity of the item was statistically significant, whereas those linked to the dummy variables referring to the topic area of the items were of very small magnitude. Likewise, there were barely any differences between the variance components associated with nationality of Tables 3 and 4, thus indicating that the additional predictors did not increase the percentage of variance explained by the model. Therefore, the model shown in Table 2 should be selected as it is the most parsimonious.

Discussion

The aim of this paper was to offer a detailed analysis of the causes of DIF by means of a methodological approach that has proven to be highly effective and versatile in the detection of DIF: logistic regression. To this end we followed the approach proposed by Swanson et al., (2002), in which multilevel logistic regression is used to examine the item characteristics that could explain DIF. The analysis conducted here followed a confirmatory strategy in which two possible causes of DIF were tested by means of four sequential models, into which these causes were progressively introduced as variables that explained the variability in the data. The comparison of these models confirmed one of the two alternatives considered (familiarity with the stimulus) and rejected the other (the topic area) as being a cause of differential functioning with respect to the compared groups. This finding means that great care should be taken to ensure that the groups to whom the instrument is applied show an equivalent degree of familiarity with the stimuli. Obviously, any investigation into the causes of an instrument's DIF should be broad enough to ensure that no potentially relevant explanations go unanalysed. If, in addition, the extent to which the findings can be generalised is also tested, then the results will lead not only to the optimisation of the instrument in question but also to improvements in the guidelines for developing and/or adapting tests with less DIF.

In recent years, and in the context of adapting tests for different languages and cultures, some progress has been made in identifying the sources of DIF and, specifically, the item characteristics that might produce it. Thus, Allalouf, Hambleton and Sireci (1999), examining the possible causes of DIF between the Hebrew and Russian versions of the Israeli Psychometric Entrance Test, found that items involving analogies and sentence completions were more problematic, and they pointed out four possible sources of DIF: changes in the difficulty of the item wording, differences in its cultural relevance, changes in the item format, and changes in the item content. Gierl and Khaliq (2001), comparing the English and French versions of an achievement test in Canadian samples, detected almost exactly the same sources of DIF. More recently, Zumbo and Gelin (2005) recommended that, in addition to item format and content, contextual variables such as school setting, parental style and socio-economic level should be taken into consideration.

However, Ferne and Rupp (2007), in a review of 27 studies carried out between 1990 and 2005 with the aim of examining the state of research on DIF, acknowledge the progress made as regards methods for detecting DIF but highlight the scant results obtained in terms of explaining its causes. The same conclusion was reached by Padilla, Pérez and González (1998) in a review of research that sought to identify the causes of DIF in aptitude and performance tests. These reviews illustrate the limited efficacy of the attempts made so far to explain the reasons for DIF, and underline the need for more detailed analysis of this area from different angles.

The «traditional» paradigm for explaining the causes of DIF follows an inductive-exploratory process. The first step is to identify the items that function differentially for one or more of the observed grouping variables, using adequate statistical techniques, and then, having identified the items that show DIF, to take a substantive approach and make use of the advice provided by experts on the construct that is measured by the test under study. This approach is not immune from problems, mainly in terms of establishing the connection between the substantive explanation and the statistical results. Its efficacy has remained limited and has not improved since the report by Camilli and Shepard (1994), which stated that in studies applied to DIF, it was not possible to offer an interpretation of the cause of DIF for half the items identified as showing high differential functioning.

The proposal of Roussos and Stout (1996), based on the multidimensional-experimental paradigm, starts from a deductive model in which, firstly, the theoretical framework is established and a hypothesis is formulated regarding the presence of DIF, which is then tested empirically. The main problem with this approach has to do with establishing what these authors call the primary and secondary dimensions in the test, as this implies having a substantive framework that enables the structural dimension of the data to be perfectly described a priori, so as to guide the subsequent data analysis. This is not always possible in practice, either because there is an underdeveloped body of theory in certain applied fields, or because, in general, prior knowledge about the possible sources of DIF is limited or no information can be gathered in this respect.

Given the above, one option would be to combine both approaches in a kind of spiralling process, such that through several iterations a range of possible causes of DIF could be elucidated confirming or rejecting the alternative explanations. Multilevel analysis may play an important role in this regard, as it considers the nested nature of data and is better than other techniques when it comes to representing the complexity of psychological or social phenomena. The present study has sought to illustrate this contribution by showing applied researchers how the multilevel approach and, specifically, logistic regression can be used to detect the item characteristics that are potentially responsible for DIF. The model also enables greater complexity by incorporating a third level on which, for example, the contextual variables referred to by Zumbo and Gelin (2005) can be modelled.

Furthermore, the items can be nested in subjects, and this means that researchers can focus specifically on the causes of DIF due to subject characteristics (Cheong & Raudenbush, 2000) or optimise the matching criteria on the basis of these characteristics (Clauser, Nungester, & Swaminathan, 1996). The approach is therefore a promising one. Indeed, the comprehensive evaluation of DIF requires such a multilevel perspective, and taking into account the impact of nested variables will lead to greater accuracy of estimations and a better interpretation of the possible causes of DIF. Finally, it should be noted that working to develop this approach is no trivial matter, since identifying the sources of DIF could be of enormous value in the future and enable researchers in a given field to minimise the number of items that function differentially by predicting and avoiding them.

Acknowledgements

The study was partly funded by a grant from Science and Innovation Ministry of the Spanish Government (PSI2009-07280), by a grant of the Gipuzkoan Network of Science, Technology and Innovation Program (OF94/2008) from the Gipuzkoa Provincial Council and European Union (FEDER), by two grants from the Research Bureau of the University of the Basque Country (General Funding for Research Groups, GIU08/09; Research Projects, EHU08/24), by an award to a consolidated research group of the Government of Catalonia (2009SGR822) and by a grant awarded under the Programme for Generating Scientific Knowledge of the Seneca Foundation, part of the Science and Technology Agency of the Region of Murcia (05725/PHCS/07).

Appendix 1
Difficulty (b) and Discrimination (a) parameters for the reference group

Item	b	a	Item	b	a	Item	b	a
1	0.70	0.56	13	0.10	1.05	25	1.05	0.70
2	0.00	0.90	14	-0.09	0.51	26	0.64	1.02
3	1.00	0.90	15	0.61	0.73	27	2.12	0.48
4	0.10	0.90	16	0.95	0.88	28	0.91	1.01
5	0.40	1.05	17	-0.35	1.11	29	0.87	0.53
6	1.28	1.02	18	0.57	1.32	30	-1.29	0.59
7	0.61	0.82	19	0.59	1.32	31	-0.57	0.86
8	0.42	0.92	20	1.64	1.40	32	0.40	0.56
9	1.68	0.65	21	0.13	0.92	33	-0.93	0.88
10	-0.39	0.90	22	-1.55	0.64	34	0.62	0.96
11	-1.12	0.35	23	0.81	1.01	35	-1.21	0.96
12	-1.37	0.31	24	0.47	0.81	36	-1.01	0.75

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Allalouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of DIF in Translated Verbal Items. *Journal of Educational Measurement, 36*(3), 185-198.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

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

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

Cheong, Y.F., & Raudenbush, S.W. (2000). Measurement and structural models of children's problem behaviors. *Psychological Methods, 5*, 477-495.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items (ITEMS Module). *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Clauser, B.E., Nungester, R.J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*, 453-464.

Cohen, A.S., Kim, S.H., & Baker, E. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*, 335-350.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

Ferne, T., & Rupp, A.A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges and recommendations. *Language Assessment Quarterly*, *4,* 113-148.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29*, 278-295.

Flowers, C.P., Oshima, T.C., & Raju, N.S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

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*(3), 315-333.

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.

Gierl, M.J., & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164-187.

Goldstein, H.I. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Gómez-Benito, J., & Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: una revisión metodológica. *Anuario de Psicología, 74*, 3-32.

Hidalgo, M.D., & Gómez-Benito, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento, 1*, 39-60.

Hidalgo, M.D., & Gómez-Benito, J. (2000). Comparación de la eficacia de la regresión logística politómica y el análisis discriminante logístico en la detección del DIF no uniforme. *Psicothema, 12*(2), 298-300.

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. Elsevier Science & Technology.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer, & H.I. Braun (Eds.): *Test Validity*. Hillsdale, N.J.: Erlbaum.

Kim, S.-H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345-355.

Kim, S.-H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomous scored items. *Journal of Educational Measurement, 44*, 93-116.

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*, 935-953.

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

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models*. Chapman and Hall: London.

Miller, T., & Spray, J. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*(2), 107-122.

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

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Ordóñez, X.G., & Romero, S.J. (2007). XS-DIF: programa para el análisis del funcionamiento diferencial de los ítems en Excel. *Psicothema, 19*, 171-172.

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

Padilla, J.L., Pérez, C., & González, A. (1998). La explicación del sesgo en los ítems. *Psicothema, 2*, 481-490.

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

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Raudenbush, S., Bryk, A., Cheong, Y.F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Scientific Software International.

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

Roussos, L.A., & Stout, W.F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Roussos, L.A., & Stout, W.F. (2004). Differential item functioning analysis. In D. Kaplan (Ed.): *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage.

Shealy, R.T., & Stout, W.F. (1993a). An item response theory model for test bias and differential test functioning. In Holland, P.W., & Wainer, H. (Eds.): *Differential item functioning* (pp. 197-239). Hillsdale, NJ: LEA.

Shealy, R.T., & Stout, W.F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika 58*, 159-194.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London: Sage Publications.

Su, Y.-H., & Wang, W.-C (2005). Efficiency of the Mantel-Haenszel, generalized Mantel-Haesnzel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*, 313-350.

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

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

Thissen, D. (2001). *IRTLRDIF v.2.0b* [Computer program]. University of North Carolina at Chapel Hall: L.L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.): *Differential item functioning* (pp. 67-113). Hillsdale, NJ: LEA.

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

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*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B.D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5,* 1-23.

Zwick, R., & Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187-201.

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