Behavior Research Methods
2010, 42 (3), 847-862
doi:10.3758/BRM.42.3.847

A general framework and an R package for the detection of dichotomous differential item functioning

David Magis
Katholieke Universiteit Leuven, Leuven, Belgium
and University of Liège, Liège, Belgium

Sébastien Béland
University of Quebec, Montreal, Quebec, Canada

Francis Tuerlinckx
Katholieke Universiteit Leuven, Leuven, Belgium

and

Paul De Boeck
Katholieke Universiteit Leuven, Leuven, Belgium
and University of Amsterdam, Amsterdam, The Netherlands

Differential item functioning (DIF) is an important issue of interest in psychometrics and educational measurement. Several methods have been proposed in recent decades for identifying items that function differently between two or more groups of examinees. Starting from a framework for classifying DIF detection methods and from a comparative overview of the most traditional methods, an R package for nine methods, called *difR*, is presented. The commands and options are briefly described, and the package is illustrated through the analysis of a data set on verbal aggression.

The present article addresses the psychometric issue of differential item functioning (DIF). An item is said to function differently (i.e., to be a *DIF item*) when subjects from different groups but with the same ability level have, nevertheless, different probabilities of answering the item correctly. DIF items can lead to biased measurement of ability because the measurement is affected by so-called *nuisance factors* (Ackerman, 1992). The presence of DIF jeopardizes the ideal of a correct measurement procedure.

Detection methods have been developed to identify DIF items, so that these items can be removed from a test. Early on, detection methods were suggested by Angoff and Ford (1973), Cardall and Coffman (1964), Cleary and Hilton (1968), Lord (1976), and Scheuneman (1979). Although of historical interest, these methods are not much used anymore. In this article, we focus on the methods that have gained considerable interest in recent decades. They are referred to here as *the traditional methods*.

Early reviews of DIF detection methods were published by Berk (1982), Ironson and Subkoviak (1979), Rudner, Getson, and Knight (1980), and Shepard, Camilli, and Averill (1981). More recent overviews have been proposed by Camilli and Shepard (1994), Clauser and Mazor (1998), Millsap and Everson (1993), Osterlind and Everson (2009), and Penfield and Camilli (2007). The distinction between methods based on item response theory (IRT) and those not based on IRT plays a major role in their classification. A second important distinction is that between uniform and nonuniform DIF. In our overview, single versus multiple focal groups will also play a role, as well as whether or not a purification procedure is followed. The framework is described in "A General Framework for DIF Analysis," below, and the methods are explained in "Detection Methods," below. Our overview and the associated R package will be restricted to methods for dichotomous items.

Commonly, each method comes with its own software tool, and there is no common software available that can be used for several methods or for a comparison of detection results. Some examples are the DICHODIF software (Rogers, Swaminathan, & Hambleton, 1993), which focuses on the Mantel–Haenszel (MH; Mantel & Haenszel, 1959) method; the IRTDIF program (Kim & Cohen, 1992); the IRTLRDIF (Thissen, 2001) and DFITPU (Raju, 1995) programs, which calculate DIF statistics based on

D. Magis, david.magis@ulg.ac.be

IRT models; and the simultaneous test bias (SIBTEST) program (Li & Stout, 1994) for the method of the same name. An exception is the DIFAS program (Penfield, 2005), which compares the methods of Mantel–Haenszel and Breslow–Day, as well as some methods for polytomous items. In this article, we present a new package for the software R (R Development Core Team, 2008), called *difR* (version 2.2), which can perform several traditional DIF detection procedures for dichotomous items. The commands of the package have a structure similar to those for all DIF detection methods, and the user can choose between several IRT-based or non-IRT-based methods. Some specific tuning parameters related to specific methods are also available. The basic working of the package is described in "An R Package for DIF," below, and is illustrated in "Example," below, by detecting DIF items in a data set of verbal aggression information.

A General Framework for DIF Analysis

The framework for describing the DIF detection methods to select from consists of four dimensions: the number of focal groups, the so-called *methodological approach* (IRT-based or non-IRT-based), the type of the DIF effect (uniform or nonuniform), and whether or not item purification is used.

Number of focal groups. The usual setting consists of comparing the responses of a reference group with those of a focal group. It can happen in practice that more than one focal group is considered. This occurs, for instance, when the performance of students from several types of schools is to be compared with that of students from a reference type of school. In other cases, none of the groups is a reference group, but one is still interested in a comparison. The common approach is to perform pairwise comparisons between each focal group and the reference group—or between all groups, if there is not a reference group. However, multiple testing has several disadvantages. First, it requires controlling for significance level, by means of a Bonferroni correction (Miller, 1981), for instance. Second, the power to detect DIF items is usually lower than the power of a single test comparing all groups simultaneously (see, e.g., Penfield, 2001). A few methods, such as the generalized MH approach and the generalized Lord's test, have been specifically developed to deal with multiple groups. They are extensions of the usual approaches for one focal group to the case of more than one focal group. Very recently, Bayesian statistical approaches were developed by Soares, Gonçalves, and Gamerman (2009), which are promising but are not discussed here, because they are based on newly formulated IRT models.

Methodological approach. There are two methodological approaches for the DIF detection methods: those relying on an IRT model, and those not relying on IRT. For the former, the estimation of an IRT model is required, and a statistical testing procedure is followed, based on the asymptotic properties of statistics derived from the estimation results. For the latter, the detection of DIF items is usually based on statistical methods for categorical data, with the total test score as a matching criterion. We refer to these classes of methods as *IRT methods* and *non-IRT methods*, respectively. Some authors use the terms *parametric* and *nonparametric* instead.

For dichotomously scored items, the usual IRT models are the logistic models with one, two, or three parameters. We further denote them by 1PL, 2PL, and 3PL models, respectively. The 3PL model can be written as

$$\Pr\left(Y_{ij}=1 \,\middle|\, \theta_i, a_j, b_j, c_j\right) = c_j + \left(1-c_j\right)\frac{\exp\left[a_j\left(\theta_i - b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i - b_j\right)\right]}, \quad (1)$$

where Y_{ij} is the binary response of subject i to item j; θ_i is the ability of subject i; and a_j, b_j, and c_j are, respectively, the discrimination, difficulty, and pseudoguessing parameters of item j. The 2PL model can be obtained from Equation 1 by fixing c_j to 0; the 1PL model comes from additionally fixing a_j to 1.

Type of DIF effect. The next concept to be introduced is concerned with the type of DIF effect. By *DIF effect*, one usually means the difference (between subjects from different groups but with the same ability level) in the probabilities of answering the tested item correctly, once these probabilities are transformed by using the model link function. If this difference between the transformed probabilities is independent of the common ability value, then the DIF effect is said to be *uniform*. On the other hand, if the difference in success probabilities (or their link function transform) is not constant across the ability levels but depends on it, then one refers to a *nonuniform* or *crossing* DIF effect. In the IRT approach, the choice of a particular model influences the type of DIF effect that is assumed (Hanson, 1998). Consider, for instance, the 1PL model obtained from Equation 1 by fixing a_j to 1 and c_j to 0. The link function of this model being the logistic (or logit) transformation, the logit of probability (Equation 1) is given by

$$\text{logit}\,\Pr(Y_{ijg}=1\,|\,\theta_i, b_{jg}) = \theta_i - b_{jg}, \quad (2)$$

where subscript g refers to the group membership, with $g=R$ for the reference group and $g=F$ for the focal group. Thus, for 2 subjects i and i^* from two different groups but having the same ability level (i.e., $\theta_i = \theta_{i^*} = \theta$), the difference in logits of probabilities is equal to $b_{jR} - b_{jF}$ and does not depend on the ability level. Therefore, the 1PL model can be used to detect uniform DIF. Also the 2PL and 3PL models can be used for that purpose, and they are also appropriate models for the detection of nonuniform DIF because they contain discrimination parameters (2PL and 3PL) and pseudoguessing parameters (3PL). *Crossing DIF* refers to the crossing of 2PL or 3PL item characteristic curves of the same item in focal and reference groups (see, e.g., Narayanan & Swaminathan, 1996). *Nonuniform* is more general and is not linked to an IRT approach.

Item purification. An important practical issue when investigating DIF is that the presence of one or several DIF items may influence the results of tests for DIF in other items. Thus, some items that are not functioning dif-

ferently can wrongly be identified as DIF items, which indicates an unwanted increase of the Type I error of the method. This is especially the case if some DIF items are included in the set of a priori non-DIF items. Such a priori non-DIF items are usually called *anchor* or *DIF-free* items. For non-IRT methods, this implies that the total test scores, which are used as proxies for ability levels, are influenced by the inclusion of DIF items. For IRT methods, the DIF items have an unwanted effect on the scaling of the item parameters used to obtain a metric (see "IRT Methods," below).

To overcome this potential confounding problem, several authors (Candell & Drasgow, 1988; Clauser, Mazor, & Hambleton, 1993; Fidalgo, Mellenbergh, & Muñiz, 2000; Holland & Thayer, 1988; Lautenschlager & Park, 1988; Wang & Su, 2004; Wang & Yeh, 2003) have suggested an iterative elimination of the DIF items, which is now commonly called *item purification*. Its principle can be sketched by using the following stepwise process.

1. Test all items one by one, assuming they are not DIF items.
2. Define a set of DIF items on the basis of the results of Step 1.
3. If the set of DIF items is empty after the first iteration, or if this set is identical to the one obtained in the previous iteration, then go to Step 6. Otherwise, go to Step 4.
4. Test all items one by one, omitting the items from the set obtained in Step 2, except when the DIF item in question is being tested.
5. Define a set of DIF items on the basis of the results of Step 4 and go to Step 3.
6. Stop.

To execute Step 4 for IRT-based methods, DIF items are discarded during the rescaling of the item parameters to a common metric. For non-IRT-based methods, the DIF items are discarded from the calculation of the total test scores and related DIF measures. Note that there is no guarantee that the iterative process will end with two successive identical sets of items, which is the stopping rule of the algorithm. To overcome this drawback, one usually sets a maximal number of iterations, and the process is stopped when this number is reached.

The alternative for item purification is a procedure that stops at Step 2 in the purification process. It is a one-step or simultaneous procedure (the detection is simultaneous for all items), and it has, therefore, the drawback that the assumption of no DIF for the other items may distort the result, but it always ends in a nonambiguous result.

Detection Methods

Table 1 lists the traditional methods, according to the number of groups, the methodological approach, and the type of DIF. Each of these methods can be used with or without purification. A general presentation of these methods follows, and their names, as displayed in Table 1, are given in italics.

Non-IRT methods for uniform DIF. Most traditional methods belong to the class of non-IRT methods and are designed to detect uniform DIF. The MH, standardization, and SIBTEST procedures are based on statistics for contingency tables. Logistic regression can be seen as a bridging method between IRT and non-IRT methods, as noticed by Camilli and Shepard (1994).

The MH method (Mantel & Haenszel, 1959) is very popular in the DIF framework (Holland & Thayer, 1988). It aims at testing whether there is an association between group membership and item response, conditionally upon the total test score (or sum score). More precisely, let J be the number of items of the test. Let T_j be the number of examinees (from both groups) with sum score j (where j is

Table 1
Traditional Methods for Detecting Differential Item Functioning (DIF)

		Number of Groups	
Framework	DIF Effect	2	>2
Non-IRT	Uniform	Mantel–Haenszel*	Pairwise comparisons
		Standardization*	Generalized Mantel–Haenszel*
		SIBTEST	
		Logistic regression*	
Non-IRT	Nonuniform	Logistic regression*	Pairwise comparisons
		Breslow–Day*	
		NU.MH	
		NU.SIBTEST	
IRT	Uniform	LRT*	Pairwise comparisons
		Lord*	Generalized Lord*
		Raju*	
IRT	Nonuniform	LRT*	Pairwise comparisons
		Lord*	Generalized Lord*
		Raju*	

Note—NU.MH, modified Mantel–Haenszel for nonuniform DIF; NU.SIBTEST, modified SIBTEST for nonuniform DIF; LRT, likelihood ratio test. *Currently implemented in difR package (Version 2.2).

taken between zero and J). Then, for any tested item, the T_j examinees are cross-classified into a 2×2 contingency table with group membership and type of response (correct or incorrect) as entries. Let A_j, B_j, C_j, and D_j be the four cell counts of this table, in which A_j and B_j refer to the numbers of correct and incorrect responses, respectively, to the tested item in the reference group. The quantities C_j and D_j refer to the corresponding numbers of correct and incorrect responses, respectively, in the focal group. Let n_{Rj} and n_{Fj} be the number of responses among examinees in the reference group and the focal group, respectively, with sum score j (so $n_{Rj} = A_j + B_j$, and $n_{Fj} = C_j + D_j$), and define m_{1j} and m_{0j} as the number of correct and incorrect responses, respectively, among examinees with sum score j (so $m_{1j} = A_j + C_j$, and $m_{0j} = B_j + D_j$). With this notation, the MH statistic can be written as

$$\mathrm{MH} = \frac{\left(\left| \sum_j A_j - \sum_j \mathrm{E}(A_j) \right| - 0.5 \right)^2}{\sum_j \mathrm{Var}(A_j)}, \tag{3}$$

where the sums over index j are restricted to sum scores that are actually observed in the data set, and where $\mathrm{E}(A_j)$ and $\mathrm{Var}(A_j)$ are given by

$$\mathrm{E}(A_j) = \frac{n_{Rj}\, m_{1j}}{T_j}$$

and

$$\mathrm{Var}(A_j) = \frac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^2 (T_j - 1)}. \tag{4}$$

Under the null hypothesis of no conditional association between item response and group membership, which corresponds to the hypothesis of no DIF, the MH statistic follows asymptotically a chi-square distribution with one degree of freedom. An item is therefore classified as DIF if the MH statistic value is larger than a critical value based on the asymptotic null distribution, which is the chi-square distribution. The correction -0.5 in Equation 3 is a continuity correction factor to improve the approximation of the chi-square distribution, which is especially needed for small frequencies.

An alternative statistic associated with the same method, which can also be used as a basis for an effect-size measure, is the common odds ratio across all j values, α_{MH} (Mantel & Haenszel, 1959), given by

$$\alpha_{\mathrm{MH}} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}. \tag{5}$$

The logarithm of this estimate, $\lambda_{\mathrm{MH}} = \log(\alpha_{\mathrm{MH}})$, is asymptotically normally distributed (see, e.g., Agresti, 1990). Values around zero indicate that the item is non-DIF. Several forms for the variance of λ_{MH} were proposed (Breslow & Liang, 1982; Hauck, 1979; Philips & Holland, 1987; Robins, Breslow, & Greenland, 1986). According to Penfield and Camilli (2007), the most commonly used variance is Philips and Holland's proposal. The log odds ratio λ_{MH} is commonly used for the DIF effect size of the item. More precisely, Holland and Thayer (1985) proposed computing $\Delta_{\mathrm{MH}} = -2.35\lambda_{\mathrm{MH}}$ and classifying the effect size as negligible if $|\Delta_{\mathrm{MH}}| \leq 1$, moderate if $1 < |\Delta_{\mathrm{MH}}| \leq 1.5$, and large if $|\Delta_{\mathrm{MH}}| > 1.5$. This is often referred to as the *ETS Delta scale* (Holland & Thayer, 1988).

A second method is *standardization* (Dorans & Kulick, 1986), which relies on an approach similar to the MH method. In the standardization method, the proportions of a correct response in each group and for each value of the total test score are compared. The standardized p difference (ST-p-DIF) is the resulting test statistic, and it can be seen as a weighted average of the differences of success rates (at each level of the test score) between focal and reference groups. Using the previous notations, the ST-p-DIF statistic takes the following form:

$$\text{ST-}p\text{-DIF} = \frac{\sum_j \omega_j \left(P_{Fj} - P_{Rj} \right)}{\sum_j \omega_j}, \tag{6}$$

where $P_{Fj} = C_j / n_{Fj}$ and $P_{Rj} = C_j / n_{Rj}$ are the proportions of successes among the focal group and the reference group, respectively, and ω_j is a weighting system. Usually ω_j is chosen as the proportion of subjects from the focal group with a total test score j, but several alternatives exist (Dorans & Kulick, 1986). The ST-p-DIF statistic can take values from -1 to $+1$. Values close to zero indicate that the item does not function differently.

Although a formula for the standard deviation of the ST-p-DIF statistic has been proposed (Dorans & Holland, 1993), the null hypothesis distribution has not yet been derived. The usual classification rule consists, therefore, in fixing a threshold *thr*, such that the item is classified as DIF if the ST-p-DIF statistic is larger than *thr*. Common choices for *thr* are .05 or .10. In addition, Dorans, Schmitt, and Bleistein (1992) proposed the absolute value of the ST-p-DIF statistic as a basis to interpret the size of DIF: negligible DIF if $|\text{ST-}p\text{-DIF}| \leq .05$, moderate DIF if $.05 < |\text{ST-}p\text{-DIF}| \leq .10$, and large DIF if $|\text{ST-}p\text{-DIF}| > .10$. Because the contingency table structure is similar to that for the MH method, it is not surprising that Dorans (1989) has shown some important similarities between the two methods. Finally, note that Dorans and Holland also proposed an adapted formulation of the standardization test for the case of multiple-choice items and a correction for guessing.

The SIBTEST method can be seen as a generalization of the standardization technique (Shealy & Stout, 1993). The corresponding SIBTEST statistic has several structural advantages with respect to the ST-p-DIF. Among others, it can test for DIF of a set of items, rather than testing each item separately, and a statistic with an asymptotic standard normal distribution is available to test the null hypothesis of no DIF.

To explain the SIBTEST, let us start from the assumption that the reference group and the focal group have

equal average ability levels. The SIBTEST statistic takes the following form:

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}\left(\hat{\beta}_U\right)}, \tag{7}$$

where $\hat{\beta}_U$ is given by

$$\hat{\beta}_U = \sum_j \mathrm{F}_j \left(\bar{Y}_{\mathrm{R}j} - \bar{Y}_{\mathrm{F}j}\right) \tag{8}$$

where F_j is the proportion of subjects from the focal group with total test score j, and $\bar{Y}_{\mathrm{R}j}$ and $\bar{Y}_{\mathrm{F}j}$ are the average scores of the subjects with total score j, from the reference and the focal group, respectively, on the set of tested items. To see the similarity with the standardization test, note that the numerator of this statistic is the same as for the latter, except for the fact that, now, an item set is considered. The term $\hat{\sigma}(\hat{\beta}_U)$ is the estimated standard error of $\hat{\beta}_U$, and its formula can be found in Shealy and Stout (1993, p. 169, Equation 19). Under the null hypothesis—that is, that the set of tested items does not function differently—the statistic B follows an asymptotic standard normal distribution.

Recall, however, that Equation 7 holds only when the two groups of examinees have the same average ability level. In practice, this is an unrealistic assumption. Therefore, Shealy and Stout (1993) suggested a regression-based correction for the average ability difference. This correction mainly consists of obtaining regression-based estimates $\bar{Y}^*_{\mathrm{R}j}$ and $\bar{Y}^*_{\mathrm{F}j}$, F^*_j, and $\hat{\sigma}(\hat{\beta}_U)^*$. These corrected values are plugged into Equations 7 and 8 instead of the corresponding uncorrected quantities. For further details, see Shealy and Stout.

In addition to its use for significance testing, the $\hat{\beta}_U$ statistic gives an indication of the DIF effect size. Roussos and Stout (1996) developed the following classification, which is derived from the ETS Delta scale for the MH procedure: negligible DIF if $|\hat{\beta}_U| \leq .059$, moderate DIF if $.059 \leq |\hat{\beta}_U| \leq .088$, and large DIF if $|\hat{\beta}_U| > .088$.

Finally, following the *logistic regression* approach (Swaminathan & Rogers, 1990), a logistic model is fitted for the probability of answering the tested item correctly, based on the total test score, group membership, and the interaction between these two. A uniform DIF effect can be detected by testing the main effect of group, and a nonuniform DIF effect can be detected by testing the interaction. Formally, the full logistic regression model has the following form:

$$\text{logit}\,(\pi_i) = \beta_0 + \beta_1\,\mathrm{S}_i + \beta_2\,\mathrm{G}_i + \beta_3\,(\mathrm{SG})_i, \tag{9}$$

where π_i is the probability of person i endorsing the item, S_i is the total test score, G_i is the group membership (focal or reference), and $(\mathrm{SG})_i$ is the interaction of S_i and G_i. Model parameters $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ are estimated and tested through the usual statistical test procedures (e.g., Wald test, likelihood ratio test, etc.). The null hypothesis of no DIF is rejected on the basis of β_3 for nonuniform DIF and on the basis of β_2 for uniform DIF. Zumbo and Thomas (1997) proposed ΔR^2 as an effect-size measure, defined as the difference between Nagelkerke's R^2 coefficients (Nagelkerke, 1991) of the two nested logistic models. For instance, the full model, with parameters $\{\beta_0, \beta_1, \beta_2, \beta_3\}$, and the reduced model, with parameters $\{\beta_0, \beta_1\}$, are to be compared when uniform and nonuniform DIF are considered simultaneously. Zumbo and Thomas proposed the following interpretation: negligible DIF if $\Delta R^2 \leq .13$, moderate DIF if $.13 < \Delta R^2 \leq .26$, and large DIF if $\Delta R^2 > .26$. Jodoin and Gierl (2001) have proposed a less conservative scale with cutoff scores of .035 and .07, instead of .13 and .26, respectively.

For multiple groups, any of the aforementioned methods (MH, standardization, SIBTEST, logistic regression) can be used for pairwise comparisons between each focal group and the reference group, or just between all groups ("Pairwise comparisons" in Table 1). Among the non-IRT methods, the MH method has been generalized to a simultaneous test for multiple groups (Penfield, 2001; Somes, 1986), indicated as the "generalized Mantel–Haenszel" method in Table 1, as suggested by Penfield (2001). The logistic regression method can also be generalized using multiple group indicators in the regression equation. This has been suggested by Millsap and Everson (1993), but it has not yet been included in a published empirical study of DIF.

Non-IRT methods for nonuniform DIF. As explained above, the logistic regression approach can also be used as a method for detecting a nonuniform DIF, but it is not the only approach. Several alternatives exist. The Breslow–Day (BD) test (Breslow & Day, 1980) determines whether the association between item response and group membership is homogeneous across the range of total test scores. If it is not, then a nonuniform DIF is present (Penfield, 2003). With the same notations as for the MH method, the BD statistic can be written as

$$\mathrm{BD} = \sum_j \frac{\left[A_j - \mathrm{E}\left(A_j\right)\right]^2}{\mathrm{Var}\left(A_j\right)}. \tag{10}$$

In Equation 10, the expected value of A_j is the positive root of a quadratic equation and equals the positive value among the two following roots:

$$\mathrm{E}\left(A_j\right) = \frac{\hat{\alpha}\left(n_{\mathrm{R}j} + m_{1j}\right) + \left(n_{\mathrm{F}j} - m_{1j}\right) \pm \sqrt{\rho}}{2\left(\hat{\alpha} - 1\right)}, \tag{11}$$

where $\hat{\alpha}$ is an estimate of the common odds ratio—for instance, as given by Equation 5—and

$$\rho = \left[\hat{\alpha}\left(n_{\mathrm{R}j} + m_{1j}\right) + \left(n_{\mathrm{F}j} - m_{1j}\right)\right]^2 - 4\hat{\alpha}\left(\hat{\alpha} - 1\right) n_{\mathrm{R}j} m_{1j}. \tag{12}$$

The variance of A_j is given by

$$\mathrm{Var}\left(A_j\right) = \left(\frac{1}{\mathrm{E}\left(A_j\right)} + \frac{1}{n_{\mathrm{R}j} - \mathrm{E}\left(A_j\right)} + \frac{1}{m_{1j} - \mathrm{E}\left(A_j\right)} + \frac{1}{n_{\mathrm{F}j} - m_{1j} + \mathrm{E}\left(A_j\right)}\right)^{-1} \tag{13}$$

(for further details, see Aguerri, Galibert, Attorresi, & Marañón, 2009). The BD statistic has an asymptotic chi-square distribution with as many degrees of freedom as the number of total test scores that are taken into account in the sum in Equation 10.

Second, several authors have proposed adapting a method for detecting uniform DIF for the case of nonuniform DIF. Modified versions of MH (Mazor, Clauser, & Hambleton, 1994) and SIBTEST (Li & Stout, 1996) are available (see also Finch & French, 2007; Narayanan & Swaminathan, 1996). They are indicated in Table 1 as NU.MH and NU.SIBTEST, respectively, with NU referring to nonuniform DIF.

For multiple groups and nonuniform DIF, and apart from the recent Bayesian approaches mentioned earlier, there seem to be no methods described in the literature. One possible approach is to extend the generalized MH method to the context of a nonuniform DIF, similar to the way Mazor et al. (1994) did for the MH technique for uniform DIF. Alternatively, the logistic regression method can be used for more than one focal group, as is mentioned in "Non-IRT methods for uniform DIF," above.

IRT methods. IRT methods can be used to detect both uniform DIF and nonuniform DIF effects. The 1PL can be used only to detect a uniform DIF, and the 2PL and 3PL are suitable for the identification of uniform and nonuniform DIF. There are three main types of IRT methods.

The first is the LRT (likelihood ratio test) method (Thissen, Steinberg, & Wainer, 1988). It consists of fitting two IRT models: a compact model with item parameters being identical for both groups of subjects and an augmented model with item parameters that are allowed to vary between the groups of examinees. The significance of these additional parameters is tested by means of the usual likelihood ratio test. Although conceptually close to the logistic regression method, this LRT technique is built upon the fitting of an item response model. According to the selected IRT model, only the item difficulties (1PL model), or also discriminations (2PL model), and pseudoguessing parameters (3PL model) can vary between the groups.

The second approach is called *Lord's chi-square test* (Lord, 1980) and is based upon the null hypothesis of equal item parameters in both groups of subjects and a statistic with a chi-square distribution under the null hypothesis. Any type of item response model can be fitted, but the item parameters must be scaled with a common metric prior to statistical testing. This issue is discussed by Candell and Drasgow (1988) and Lautenschlager and Park (1988), among others. The Q_j statistic used for this method has the following form:

$$Q_j = (v_{j\mathrm{R}} - v_{j\mathrm{F}})'(\Sigma_{j\mathrm{R}} - \Sigma_{j\mathrm{F}})^{-1}(v_{j\mathrm{R}} - v_{j\mathrm{F}}), \qquad (14)$$

where $v_{j\mathrm{R}} = (a_{j\mathrm{R}}, b_{j\mathrm{R}}, c_{j\mathrm{R}})$ and $v_{j\mathrm{F}} = (a_{j\mathrm{F}}, b_{j\mathrm{F}}, c_{j\mathrm{F}})$ are the vectors of item discrimination, difficulty, and pseudoguessing estimates of item j in the reference group and focal group, respectively, and $\Sigma_{j\mathrm{R}}$ and $\Sigma_{j\mathrm{F}}$ are the corresponding variance–covariance matrices. The Q_j statistic has an asymptotic chi-square distribution and relies on the asymptotic normality of the maximum likelihood estimates of the item parameters. The degrees of freedom correspond to the number of estimated parameters in the model. Note that, under the 1PL model, the statistic in Equation 14 has the simple form

$$Q_j = \frac{\left(b_{j\mathrm{R}} - b_{j\mathrm{F}}\right)^2}{\hat{\sigma}^2_{j\mathrm{R}} + \hat{\sigma}^2_{j\mathrm{F}}}, \qquad (15)$$

where $\hat{\sigma}_{j\mathrm{R}}$ and $\hat{\sigma}_{j\mathrm{F}}$ are the estimated standard errors of item difficulty in the reference group and focal group, respectively.

Kim, Cohen, and Park (1995) extended Lord's test to more than one focal group in a procedure called the *generalized Lord test*. The Q_j statistic from Equation 14 is then generalized to the following form:

$$Q_j = (Cv_j)'(C\Sigma_j C')^{-1}(Cv_j), \qquad (16)$$

where v_j is obtained by concatenating the vectors of the estimated item parameters in the reference group and in the focal groups, and where Σ_j is the corresponding block diagonal matrix where each diagonal block is the variance–covariance matrix of item parameters in each respective group of subjects. The C matrix is a design matrix indicating the item parameters one is interested in for a comparison between the groups (for further details, see Kim et al., 1995). This generalized Lord statistic also has an asymptotic chi-square distribution with as many degrees of freedom as the rank of the design matrix C. It is important to recall that all parameter estimates in the vector v_j must have a common metric for all groups before the Q_j statistic is computed.

The third method is the *Raju* method (Raju, 1988, 1990), and, in this method, the (signed) area between the item characteristic curves for the focal group and the reference group is computed. The corresponding Z statistic is based on the null hypothesis that the true area is zero. A common metric is required prior to the test. Any item response model can be considered with Raju's (1988) approach. However, an important restriction is that, for each item, the pseudoguessing parameters for both groups of subjects are constrained to be equal.

With the 1PL model, the area between the characteristic curves (in the reference group and in the focal group) of an item is simply given by the difference in item difficulty estimates (Raju, 1988), so that the Z statistic is simply given as follows:

$$Z = \frac{b_{j\mathrm{R}} - b_{j\mathrm{F}}}{\sqrt{\hat{\sigma}^2_{j\mathrm{R}} + \hat{\sigma}^2_{j\mathrm{F}}}}. \qquad (17)$$

The square of this Z statistic is identical to Lord's statistic, as shown in Equation 15. For 2PL and 3PL models, the formula for Z is much more complex and can be found in Raju (1990).

An R Package for DIF

We have developed an R package for nine of the aforementioned methods so they can be used simultaneously

and their results can be compared. The package is called difR and is briefly described below. The interested reader can find more details in the difR manual, which can be obtained by request to the first or second author of the present article.

Installation and software. Working with difR requires the installation of the software R and two working packages: ltm (Rizopoulos, 2006) and lme4 (Bates & Maechler, 2009). Version 2.8.0 of R or a more recent version is required. The latest edition of R can be downloaded from the R Project Web site: www.r-project.org.

The ltm package is required for fitting logistic item response models and provides item parameter estimates and standard errors. The usual 1PL, 2PL, and 3PL item response models can be fitted with this package. The marginal maximum likelihood approach is used for the estimation, with a default of 40 iterations for the expectation maximization (EM) algorithm and 15 quadrature points for the Gauss–Hermite approximation of the required integrals. The R commands of ltm can be used in difR for the Lord and Raju IRT methods.

The lme4 package permits fitting the 1PL model as a generalized linear mixed model, using its lmer function, with fixed item and random person effects, with and without an interaction between the tested item and group membership (for more information, see De Boeck & Wilson, 2004). Such a model is particularly useful for the LRT method, and it is the only one currently available for that method. It can also be used for the Lord and Raju approaches when the 1PL model is considered. For binary data, lme4 makes use of the Laplace approximation of the required integrals. With the current version of lme4 (version 0.999375-32, as of October 20, 2009), it is impossible to fit the 2PL and 3PL models as mixed models. On the other hand, lmer can deal with missing data, whereas ltm cannot.

Both packages can be installed directly from the R Project Web site. When used for model fitting in difR, the item parameter estimates will be extracted from their output and integrated into the DIF detection methods. The difR package itself and its users' manual can be downloaded for free from ppw.kuleuven.be/okp/software/difR. The package can be installed locally from the "Packages" menu of the R console: Select "Install package(s) from local zip files . . ." Finally, the package has to be loaded in R by entering the `require(difR)` command into the R console.

R commands for DIF detection. Basically, all functions for detecting DIF items have the same structure. All of the commands start with "`dif`," followed by the acronym for the specified method. Table 2 lists the nine available methods in the difR package. The first column shows the name of the R command to be called for the requested method, which is displayed in the second column. The third column indicates the names of the required arguments for data input. These arguments are discussed in the next section.

Data input. The user must always provide three pieces of information: (1) the data set, (2) the group membership of the respondents, and (3) the focal group label(s). The data set has the usual structure: one row per subject, one column per item, with 1, 0 entries only. In the current version of the package, complete response patterns must be provided because several methods will fail to provide a result if at least one response pattern is incomplete. The data set can also contain the names of the items to be included as column names. The data set is always passed through the R commands by means of the *data* argument, either as a matrix or as a data frame.

The group membership of the respondents can be provided as a separate vector or as a column of the data set itself. In the latter case, the user has to specify which column of the data set corresponds to the group membership. The *group* argument is used for that. The name of the group membership vector can also be specified as a column name.

The components of the group membership vector can be either numeric strings or character strings, and one is required for specifying the components that refer to the focal group(s). This is achieved by using the *focal.name* arguments if two groups of respondents are considered or the *focal.names* arguments in the multiple groups setting.

If one is interested in the Lord or Raju methods for DIF detection, it is possible to provide the item parameter estimates directly. This is particularly useful if another software tool, such as BILOG (Mislevy & Bock, 1984; Mislevy & Stocking, 1989), is used for item parameter estimation. If the parameter estimates are not given, the user has to specify the model that must be fitted to the data. The package has an internal function (namely,

Table 2
Main R Commands and Related Arguments for Data Input

R Command	Method	Arguments
`difBD`	Breslow–Day	*Data*, *group*, *focal.name*
`difGenLord`	Generalized Lord	*Data*, *group*, *focal.names*, *model*, *c*, *engine*, *irtParam*, *nrFocal*, *same.scale*
`difGMH`	Generalized Mantel–Haenszel	*Data*, *group*, *focal.names*
`difLogistic`	Logistic regression	*Data*, *group*, *focal.name*
`difLord`	Lord's chi-square test	*Data*, *group*, *focal.name*, *model*, *c*, *engine*, *irtParam*, *same.scale*
`difLRT`	Likelihood ratio test	*Data*, *group*, *focal.name*
`difMH`	Mantel–Haenszel	*Data*, *group*, *focal.name*
`difRaju`	Raju's area	*Data*, *group*, *focal.name*, *model*, *c*, *engine*, *irtParam*, *same.scale*
`difStd`	Standardization	*Data*, *group*, *focal.name*

itemParEst) that can fit the selected model to each group, using the commands of the ltm or lme4 packages, according to the user's choice (see below). If preestimated item parameters are used, the computation time may be considerably shorter.

For Lord and Raju methods, the user can provide the estimates of item parameters directly, instead of the full data matrix. These estimates can be passed to the R commands through the *irtParam* argument, in the format of a matrix with one row per item and one column per parameter estimate with standard errors and, possibly, covariances between the parameters. The proper format of this *irtParam* matrix is rather technical, and the interested reader can find more detailed information in the help file of the *itemParEst* function or in the difR documentation.

In addition, the *same.scale* logical argument is used to specify whether the item parameters of the *irtParam* matrix are already placed on a common metric. If they are not, the item parameters of the focal groups are rescaled to the metric of the reference group by equal means anchoring through the `itemRescale` command (see Cook & Eignor, 1991, and the R help file of `itemRescale` for further information). The rescaling is such that the mean difficulty is the same in both groups. Other anchoring methods may be considered, but, currently, only the equal means anchoring approach is implemented in the difR package. Updated versions of the packages will incorporate alternative anchoring methods.

In order to specify the model to be estimated, one makes use of the *model*, *c*, and, possibly, *engine* arguments. The *model* argument must be one of the following three: "1PL," "2PL," or "3PL." The *c* argument is optional and is used to constrain the pseudoguessing parameters, as required by the Raju method, but it can also be applied to other IRT methods. If *c* is a single numeric value, all pseudoguessing parameters (for all groups and all items) are equal to this value. Otherwise, *c* must be a vector of the same length as the number of items, and each entry corresponds to the common value of the pseudoguessing parameters for the considered item in the reference and focal groups. If *c* is left unspecified, the pseudoguessing parameters are estimated separately for each item and each group of subjects.

Finally, the *engine* argument indicates which package will be used for model fitting. The default value is "*ltm*," which refers to the marginal maximum likelihood estimate of the model, but one can also request the Laplace approximation with the value "*lme4*" for the *engine* argument. The default value of *engine* was set because ltm is faster than lme4 for fitting the 1PL model. However, the *engine* argument is not used for the LRT method, because ltm cannot incorporate an interaction between the tested item and group membership (with ltm, the item parameters are estimated separately in each group of subjects). Thus, *engine* is an option only for the Lord, generalized Lord, and Raju methods, whereas, for the LRT method, lme4 is the only option. Moreover, since the 2PL and 3PL models cannot be fitted with lme4, the *engine* argument is actually only useful when the 1PL is considered.

Specific input arguments. Several commands have specific parameters that are intrinsic to the methods. Table 3 displays the full list of specific parameters, providing the names, the precise effects, and the method for which they are designed.

First, the statistical detection threshold must be supplied in the form of an *alpha* argument. For the standardization method, the threshold is not an alpha level, and it must be fully specified through the *thr* argument. An item will be detected as DIF if the absolute value of the corresponding ST-*p*-DIF statistic is larger than *thr*. The default value is .10, but any other value can be considered.

For the MH method, an optional argument is available for obtaining a more continuous distribution and, hence, to better approach the asymptotic normality of that statistic (Holland & Thayer, 1988). The correction of -0.5 is desirable if some of the expected frequencies are very small—especially when they are lower than five (Agresti, 1990). In the DIF framework, this correction is commonly adopted. The *correct* argument is a logical argument and takes the value TRUE by default, in line with the current practice.

The last two specific arguments are related to item purification. The *purify* argument determines whether purification has to be performed. This argument is of the logical type and is FALSE by default, so that item purification is performed only when the argument is used and is given the value TRUE. The second related argument is *nrIter*, and it specifies the maximum number of iterations in the purification process. It may happen that the purification needs a large number of iterations. Because it can lead to an endless loop and would thus fail to stop, it is useful to set a maximum number of iterations (by default, $nrIter = 10$). A warning is given if convergence is not reached after *nrIter* iterations.

Output. There are three kinds of output: (1) the output that is returned by each of the R commands; (2) the

Table 3
Specific Arguments of the Main R Commands

R Argument	Description	Methods
alpha	Numeric: the significance level (default is 0.05)	All methods but Std
thr	Numeric: the threshold (or cut-score) for standardized P-DIF statistic (default is 0.10)	Std
correct	Logical: Should the continuity correction be used? (default is TRUE)	MH
purify	Logical: Should the method be used iteratively to purify the set of anchor items? (default is FALSE)	All methods
nrIter	Numeric: the maximal number of iterations for the purification process (default is 10)	All methods

Note—MH, Mantel–Haenszel; Std, standardization; DIF, differential item functioning.

Table 4
Output Arguments of the Main R Commands

Output Argument	Signification and Value	Methods
Unspecified	Vector of DIF statistic values	All methods
alphaMH	The values of the log-odds ratios α_{MH}	MH
deltaR2	ΔR^2 differences between R^2 coefficients	Logistic
alpha	Significance level	All methods but Std
thr	Threshold for DIF item detection	All methods
df	Degrees of freedom of generalized Lord statistic	GenLord
DIFitems	The column indicators of the items detected as DIF (if any), or "no DIF item detected"	All methods
correct	Logical: Was the continuity correction applied?	MH
purification	Logical: Was item purification applied?	All methods
nrPur	Number of iterations in item purification	All methods
difPur	Matrix of successive classification of the items	All methods
convergence	Logical: Did purification converge?	All methods
model	The fitted item response model	Lord, Raju, GenLord
c	Values of constrained pseudoguessing parameters or NULL	Lord, Raju, GenLord
engine	The engine package for fitting the IRT model	Lord, Raju, GenLord
itemParInit	Initial item parameter estimates	Lord, Raju, GenLord
itemParFinal	Final parameter estimates (after purification)	Lord, Raju, GenLord
estPar	Logical: Were item parameters estimated or provided?	Lord, Raju, GenLord
focal.names	Names of the focal groups	GMH
names	Names of the items	All methods

Note—Std, standardization; MH, Mantel–Haenszel; GMH, generalized Mantel–Haenszel; GenLord, generalized Lord.

output that is displayed into the R console, which is a user-friendly version of the same output in a single output print; and (3) a visual representation of the DIF detection results.

First, each R command for DIF detection returns its own output to be specified through output arguments. The full output varies from method to method, but most of the output elements are common to all methods. Table 4 displays the elements that can be requested for the output list, and it also indicates the methods for which the elements can be requested.

The values of the DIF statistics at the last step of the purification process, if any, are always returned as the first element of the list. Because the names depend on the method, the first element of Table 4 is listed as "Unspecified." If available from the literature, it is also indicated for each method what the cutoff values are for the interpretation of a statistic, with regard to negligible, moderate, or large DIFs. Other common elements of the output are the significance level (except for standardization), the corresponding threshold value of the statistic for flagging an item as DIF, the items, the set of items that are detected as functioning differently (if any), and the names of the items (if provided as column names of the data matrix). These are provided by the *alpha*, *thr*, *DIFitems*, and *names* elements of the output, respectively. For the MH method, the choice of whether or not to apply the continuity correction is also returned (with the *correct* argument), and the number of degrees of freedom, if applicable, is provided by means of the *df* argument.

If the purification process is requested, several additional elements are provided. The *nrPur* argument gives the number of iterations effectively run, and the logical *convergence* element indicates whether the process converged. Finally, *difPur* yields a matrix with one row per iteration and one column per item, with zeros and ones for the items detected as being non-DIF and DIF, respectively. This matrix lists the different detection steps of the purification, and it can be used to determine whether the process shows a loop.

For IRT methods (Lord, Raju, and generalized Lord), the output list also provides the *model* element, which corresponds to the selected item response model, and the *c* argument with the value of the constrained pseudoguessing parameters (if provided). The item parameter estimates are returned in the same format as that of the *irtParam* argument for the data input. The matrix of initial parameter estimates, being either estimated first by the program or provided by the user, is returned through the *itemParInit* element. If item purification is chosen, the *itemParFinal* element returns the final parameter estimates.

The second kind of output is a user-friendly summary of the DIF detection results, possibly of several methods, in a single output printout. This output is provided if the *dichoDif* command is used. Only methods designed for one focal group can be considered, but both IRT and non-IRT methods can be called in this command. The arguments for data input are identical to those previously mentioned (*data*, *group*, *focal.name*, *model*, *c*, *engine*, *irtParam*, *same.scale*). In addition, one has to specify, through the *model* argument, a vector of acronyms for the requested method: "MH" for the Mantel–Haenszel method, "Std" for standardization, "Logistic" for logistic regression, "BD" for the Breslow–Day method, "Lord" for Lord's chi-square test, "Raju" for Raju's area method,

and "LRT" for the likelihood ratio test method. Also, all specific options can be made through the arguments with the same name; for instance, the significance level can be fixed by using the *alpha* argument.

The output of the *dichoDif* command is twofold. First, it lists all specific options chosen. Second, it shows a matrix with one row per item and one column per selected method. Each column displays the final classification of the items with the values "DIF" and "NoDIF." This matrix permits an easy comparison of the methods in terms of the classification of items as DIF or no-DIF.

The third kind of output is a plot of the DIF statistic values for visual inspection of DIF. The plot command is simply called with the R code *plot(result)*, where *result* must be specified by referring to one of the DIF detection methods. The items are displayed on the *x*-axis, and the DIF statistic values are displayed on the *y*-axis; the detection threshold is represented by a horizontal line. Figure 2 shows the visual output for the example to be described next. Several graphical options (such as the color and the type of symbol for item localization) are available. See the help files of the corresponding methods for further information.

difR and other software. One may wonder how well the results of difR would correspond with the results of other, mostly single-method programs. Therefore, we have checked the correspondence between the results returned by the difR commands and those returned by some other software.

For some nonparametric methods (standardization, logistic regression, generalized MH), we did not find any specific DIF software. However, the fitting of the logistic regression models was compared with that of SAS PROC LOGISTIC, and both packages returned identical results. Similarly, the values of the generalized MH statistics were compared with those of SAS PROC FREQ (CMH option), and, again, identical results were returned. Moreover, the MH difR output was compared with that of the DIFAS program (Penfield, 2001), and the results were identical. Because the Breslow–Day method currently implemented in difR is slightly different from that proposed in DIFAS, the latter software was not used for comparisons. Instead, SAS PROC FREQ was used, since it also returns the Breslow–Day statistics, and again, identical results were obtained.

For the parametric methods, the problem is twofold: The item parameters must be estimated adequately, and the methods must be correctly implemented. The difR package relies on the application of estimation routines from the ltm and lme4 packages, and empirical comparisons between these packages and other programs indicate that item parameter estimates are accurate. Moreover, the current implementation of the Lord's and generalized Lord's tests gives similar results to those published in Kim et al. (1995). Also, the results of the Raju method were similar to those from Raju's 1990 article. Note, however, that some differences in DIF statistics occurred, but these were minor and can be attributed to rounding in the published parameter estimates that we used to start from. Finally, no comparison was made for the LRT method using the IRTLRDIF software (Thissen, 2001), since the latter makes use of models (2PL, 3PL, GRM) other than those used for the current implementation of the LRT. Instead, the LRT difR results were compared with those obtained from Multilog (Thissen, Chen, & Bock, 2003), and almost identical results were obtained. For some items, the difference between the LRT statistics was small (≤ 0.1), and this was due to differences in the number of decimal values between Multilog and difR.

In sum, the preliminary checks of the difR package indicate that the current implementation of the DIF detection methods provides accurate and reliable results, although further investigation seems desirable. A full comparison will not be possible because, as mentioned earlier, for some of the methods, there is no standard software to compare.

Example

We illustrate the difR package by analyzing a data set about self-report verbal aggression. This data set stems from a study described in De Boeck and Wilson (2004), Smits, De Boeck, and Vansteelandt (2004), and Vansteelandt (2000), with 316 respondents (243 women and 73 men) and 24 items. The respondents were freshman students in psychology at the K.U. Leuven (Belgium). All items describe a frustrating situation, together with a possible verbal aggression response. The data are binary. The verbal aggression data set is included in both the difR package and the lme4 package and is used in the following to illustrate the commands.

The data set is called "verbal" and consists of 26 columns. The first 24 columns refer to the items, the 25th column (labeled "Anger") corresponds to the *trait anger* score (Spielberger, 1988) of each subject, and the 26th column (labeled "Gender") contains the group membership, with 0 and 1 entries for female and male respondents, respectively.

First, we have to load the *verbal* data set, using the `data(verbal)` R code, and exclude the *anger* variable from the data set, because it is not used here:

```
verbal <- verbal[colnames(verbal)!="Anger"]
```

We specify the *data* argument as the *verbal* full matrix and the *group* argument as *gender*, which is actually the label of the column with the group membership vector. Furthermore, the focal group will correspond to the male respondents, for which *gender* equals one.

The data are analyzed with the MH method as an illustration of uniform DIF detection. Other methods can be used similarly, with an appropriate selection of the options. We set a significance level of .05, and we consider the usual continuity correction. These two are default options, so they do not need to be specified. Furthermore, we request an item purification with no more than 20 iterations. The corresponding R code is given below:

```
difMH(Data=verbal, group="Gender",
focal.name=1, purify=TRUE, nrIter=20)
```

The output is displayed in Figures 1 and 2, exactly as it appears in the R console.

```
Detection of Differential Item Functioning using Mantel-Haenszel method
with continuity correction and with item purification

Convergence reached after 6 iterations

Mantel-Haenszel chi-square statistic:

            Stat.   P-value
S1wantCurse  0.0069  0.9336
S1WantScold  0.0376  0.8462
S1WantShout  0.0087  0.9256
S2WantCurse  0.8881  0.3460
S2WantScold  0.1112  0.7388
S2WantShout  4.2680  0.0388 *
S3WantCurse  0.0987  0.7534
S3WantScold  4.3724  0.0365 *
S3WantShout  0.3454  0.5567
S4WantCurse  0.1406  0.7077
S4WantScold  1.6853  0.1942
S4WantShout  1.0766  0.2995
S1DoCurse    2.0959  0.1477
S1DoScold    6.2736  0.0123 *
S1DoShout    0.0172  0.8957
S2DoCurse    9.6672  0.0019 **
S2DoScold   11.9436  0.0005 ***
S2DoShout    0.6997  0.4029
S3DoCurse    9.4644  0.0021 **
S3DoScold    6.4356  0.0112 *
S3DoShout    1.4190  0.2336
S4DoCurse    3.9323  0.0474 *
S4DoScold    5.7987  0.0160 *
S4DoShout    0.3223  0.5702

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Detection threshold: 3.8415 (significance level: 0.05)

Items detected as DIF items:

 S2WantShout
 S3WantScold
 S1DoScold
 S2DoCurse
 S2DoScold
 S3DoCurse
 S3DoScold
 S4DoCurse
 S4DoScold
```

Figure 1. First part of the output of the `difMH` command with the verbal aggression data set.

The first sentence of the output reports that the MH method is used, that the continuity correction was made, and that an item purification was performed. Next, it is reported that the purification process reached convergence after six iterations. The matrix of successive classifications (not shown in Figure 1) indicates that 18 items are always classified identically across the six iterations. *S2WantShout*, *S2DoCurse*, *S2DoScold*, and *S3DoCurse* were always identified as DIF items; 14 other items were never detected as DIF throughout the purification process. The successive classifications of the remaining six items are displayed in Table 5. Note that Step 0 corresponds to the initial classification of the items, before item purification starts. One can clearly see the slight changes in the successive iterations, until Steps 5 and 6 have identical results, so that the purification process is stopped.

Table 5
Successive Classifications of Items From the Verbal Aggression Data Set During Item Purification

Step	S2WantCurse	S3WantScold	S1DoScold	S3DoScold	S4DoCurse	S4DoScold
0	NoDIF	NoDIF	NoDIF	DIF	NoDIF	NoDIF
1	NoDIF	NoDIF	DIF	NoDIF	NoDIF	NoDIF
2	DIF	NoDIF	DIF	DIF	DIF	NoDIF
3	NoDIF	NoDIF	DIF	DIF	NoDIF	DIF
4	NoDIF	NoDIF	DIF	DIF	DIF	DIF
5	NoDIF	DIF	DIF	DIF	DIF	DIF
6	NoDIF	DIF	DIF	DIF	DIF	DIF

Note—Only items whose DIF or non-DIF status changes over the iterative steps are displayed.

The rest of the output shows the MH chi-square statistic values obtained in the last step of the purification process, when DIF items are discarded from the computation of sum scores. The corresponding *p* values are also displayed, and the significance levels are indicated with one or more asterisks. Nine items (out of 24) are eventually detected as functioning differently, 4 items being always flagged as DIF and 5 coming from the item purification shown in Table 5. They can also be found in the summary table as items with at least one asterisk, and they are listed at the end of the output.

```
Effect size (ETS Delta scale):

Effect size code:
 '*': negligible effect
 '**': moderate effect
 '***': large effect

            alphaMH deltaMH
S1wantCurse  1.1054 -0.2355 *
S1WantScold  1.1357 -0.2990 *
S1WantShout  1.0919 -0.2066 *
S2WantCurse  1.5998 -1.1042 **
S2WantScold  1.1994 -0.4272 *
S2WantShout  2.2088 -1.8623 ***
S3WantCurse  0.8584  0.3588 *
S3WantScold  0.4593  1.8283 ***
S3WantShout  1.3348 -0.6786 *
S4WantCurse  1.2183 -0.4639 *
S4WantScold  0.6322  1.0776 **
S4WantShout  1.5764 -1.0696 **
S1DoCurse    0.5481  1.4130 **
S1DoScold    0.3832  2.2540 ***
S1DoShout    0.9822  0.0421 *
S2DoCurse    0.2658  3.1134 ***
S2DoScold    0.3014  2.8186 ***
S2DoShout    0.6832  0.8954 *
S3DoCurse    0.3713  2.3284 ***
S3DoScold    0.4079  2.1075 ***
S3DoShout    0.4654  1.7974 ***
S4DoCurse    0.4744  1.7526 ***
S4DoScold    0.4148  2.0681 ***
S4DoShout    1.4040 -0.7974 *

Effect size codes: 0 '*' 1.0 '**' 1.5 '***'
 (for absolute values of 'deltaMH')
```

Figure 2. Second part of the output of the `difMH` command with the verbal aggression data set.

The last part of the output (Figure 2) shows the effect sizes, beginning with the three size-interpretation categories. Next follows a table with three columns: the MH common odds ratio estimates (the "alphaMH" column), the effect sizes Δ_{MH} ("deltaMH"), and the ETS Delta scale classification. The classification cutoff values are given at the bottom of Figure 2. Several items exhibit moderate or large DIF effects, but all items flagged as DIF (and listed in the end of Figure 1 and in Table 5) have a large DIF effect. This indicates that all items flagged as DIF on the basis of the significance test can be considered to be largely affected by DIF.

The results in Figure 1 can also be displayed graphically using the following R code:

```
res.MH<- difMH(Data=verbal,
group="Gender", focal.name=1,
purify=TRUE, nrIter=20) plot(res.MH)
```

The first part of the code simply saves the MH results into the so-called *res.MH* variable, which is then plotted following the `plot()` command. The output is given in Figure 3.

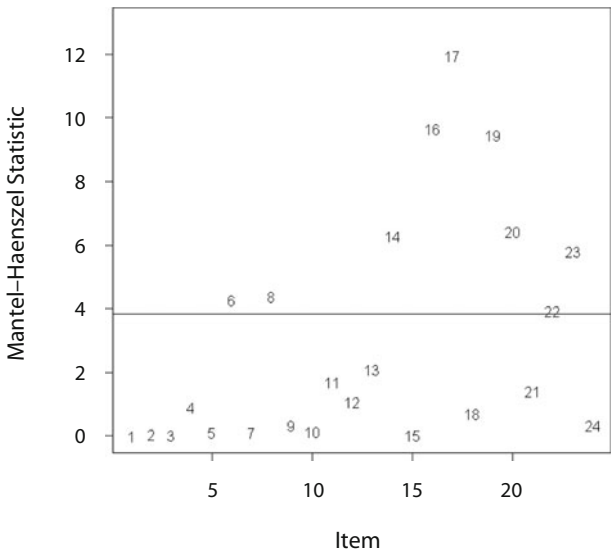

Figure 3. Mantel–Haenszel statistics and detection threshold with the verbal aggression data set.

```
Comparison of DIF detection results using 5 methods

Methods used: Mantel-Haenszel, Standardization, Logistic regression,
Lord's chi-square test, Raju's area

Parameters:
Significance level: 0.05
Standardization threshold: 0.1
Mantel-Haenszel continuity correction: Yes
Item response model: 1PL
Item purification: Yes

Item purification results:

            M-H Stand. Logistic Lord Raju
Convergence Yes    Yes      Yes  Yes  Yes
Iterations   6      5        2    2    2

Comparison of DIF detection results:

            M-H   Stand. Logistic Lord  Raju  #DIF
S1wantCurse NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S1WantScold NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S1WantShout NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S2WantCurse NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S2WantScold NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S2WantShout  DIF   DIF    DIF      DIF   DIF   5/5
S3WantCurse NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S3WantScold  DIF   DIF   NoDIF    NoDIF NoDIF  2/5
S3WantShout NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S4WantCurse NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S4WantScold NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S4WantShout NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S1DoCurse   NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S1DoScold    DIF   DIF    DIF      DIF   DIF   5/5
S1DoShout   NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S2DoCurse    DIF   DIF    DIF      DIF   DIF   5/5
S2DoScold    DIF   DIF    DIF      DIF   DIF   5/5
S2DoShout   NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S3DoCurse    DIF   DIF    DIF      DIF   DIF   5/5
S3DoScold    DIF   DIF    DIF      DIF   DIF   5/5
S3DoShout   NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
S4DoCurse    DIF   DIF   NoDIF    NoDIF NoDIF  2/5
S4DoScold    DIF   DIF   NoDIF    NoDIF NoDIF  2/5
S4DoShout   NoDIF NoDIF  NoDIF    NoDIF NoDIF  0/5
```

Figure 4. Output of the `dichoDif` command with the verbal aggression data set.

Items are represented by integers referring to their rank in the output list of Figure 1 (1 for the *S1wantCurse* item, etc.). Items 6, 8, 14, 16, 17, 19, 20, 22, and 23 are detected as DIF items. Note that Item 22, labeled "*S4DoCurse*," is borderline. Most of these items are of the "do" type, meaning that they refer to a self-report of actual verbal aggression, rather than "wanting" to be verbally aggressive. The obtained positive effect size values mean that men are more inclined than women to actually curse and scold in response to frustrating situations, independent of their degree of inclination to verbal aggression.

The use of the generic `dichoDif` command is also illustrated here. Five methods are being compared: MH, standardization, logistic regression, Lord's test, and Raju's method. The first two methods focus on uniform DIF. For the last two methods, the 1PL model is fitted. The significance level is .05, and the standardization threshold is fixed to .075 (as an average value between .05 and .10, the two thresholds suggested in the literature). Item purification is requested with a maximum of 20 iterations. The R command for these options is given below:

```
dichoDif(Data=verbal, group="Gender",
focal.name=1, method=c("MH", "Std",
"Logistic", "Lord", "Raju"), model="1PL",
purify=TRUE, nrIter=20)
```

The output is presented in Figure 4.

The first part of the output contains the number and the names of the methods under consideration. The list of all options, such as the significance level and the item response model, is also included. Then, since item purification was requested, a summary table is displayed. The first row indicates whether the iterative process converged; in the second row, one can find the actual number of iterations used to purify the set of items. Finally, the summary table of DIF detected items is shown. An additional column counts the number of methods for which the item in question was detected as a DIF item.

In this particular example, one can notice that the five methods mostly agree. Six items are identified as DIF items by all five methods, whereas 15 other items are never identified as such. For 3 items, the MH and standardization methods identify them as DIF, but the other methods do not.

Discussion and Conclusion

The difR package has several advantages with respect to the most common DIF software. First, the different methods can be set up with a similar structure and feature many flexible options for the user. For instance, the IRT methods offer the option of estimating item parameters or providing them as input from a previous calibration. The data input can also be performed in different ways, according to the coding and format of the data set. Second, the package handles several DIF detection methods, so they can be compared in one run. The `dichoDif` command can be used for a comparison of the results, displaying them in an attractive format, so that it is easy to detect the items that are never, sometimes, or always flagged as DIF by the different methods. Third, the package has been developed for the R software, and it can thus be obtained for free. The package requires some knowledge of the R environment; the help manual provides additional information and references for the interested user.

One drawback of the package—and this is inherent to the R software—is that the calculations can be very slow with large data sets. It has been pointed out that the current form of the likelihood ratio test sometimes takes very long to provide the values of the DIF statistics, but this can also be true of the IRT methods when the number of subjects or items is very large. For instance, it took about 2 h to obtain the results of the LRT method with the verbal aggression data set (24 items, 316 subjects); the computational time for the other methods, using the same data set, was less than 2 min (using a PC with a 2.13-GHz processor and 2 GB of RAM).

Two additional issues remain to be investigated further. First, the problem of sparse data is not dealt with any further than discarding the strata of total test scores with fewer than two observations. It would be of interest to consider merging options to improve the adequacy of the null distribution (Agresti, 1990). Because total scores are not used for the IRT approaches reported here, the problem does not occur for these methods.

Second, the impact of missing data on the practical working of the package has not yet been investigated. For parametric methods, the models can be estimated with missing data (assuming they are missing at random), with an impact on the reliability of the estimation results, so that the effect is taken into account. For nonparametric methods, persons with missing data can be omitted from the analysis, or, alternatively, an imputation strategy may be adopted. However, this would require further study, and, for the time being, the recommendation is to consider only full response patterns.

AUTHOR NOTE

The authors thank the editor and two anonymous reviewers for many helpful comments that greatly improved the contents of the article. The authors are also grateful to Kristof Vansteelandt (K.U. Leuven, Belgium), who shared his data and permitted us to incorporate them into the difR package, to Gilles Raîche (UQAM, Canada), who contributed significantly to the compilation of the R package, and to Bruno Facon (Université de Lille, France) for insightful comments on the difR package and its practical usefulness. This research was financially supported by the Belgian Federal Science Policy (Funds IAP/P6/03), the Research Fund GOA/2005/04 of the K.U. Leuven, Belgium, a doctoral grant "Bourse à la mobilité (hors Québec) pour l'intégration à la communauté scientifique en éducation" of the UQAM, Canada, and a postdoctoral grant "Chargé de recherches" of the National Funds for Scientific Research (FNRS), Belgium. Correspondence concerning this article should be addressed to D. Magis, Department of Mathematics, University of Liège, Grande Traverse 12, B-4000 Liège, Belgium (e-mail: david.magis@ulg.ac.be).

REFERENCES

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

AGRESTI, A. (1990). *Categorical data analysis*. New York: Wiley.

AGUERRI, M. E., GALIBERT, M. S., ATTORRESI, H. F., & MARAÑÓN, P. P. (2009). Erroneous detection of nonuniform DIF using the Breslow–Day test in a short test. *Quality & Quantity*, **43**, 35-44.

ANGOFF, W. H., & FORD, S. F. (1973). Item–race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, **10**, 95-106.

BATES, D., & MAECHLER, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package Version 0.999375-32. Available from https://r-forge.r-project.org/R/?group_id=60.

BERK, R. A. (1982). *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

BRESLOW, N. E., & DAY, N. E. (1980). *Statistical methods in cancer research: Vol. 1. The analysis of case–control studies* (Scientific Publication No. 32). Lyon, France: International Agency for Research on Cancer.

BRESLOW, N. E., & LIANG, K. Y. (1982). The variance of the Mantel–Haenszel estimator. *Biometrics*, **38**, 943-952.

CAMILLI, G., & SHEPARD, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

CANDELL, G. L., & DRASGOW, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, **12**, 253-260.

CARDALL, C., & COFFMAN, W. E. (1964). *A method for comparing the performance of different groups on the items in a test* (Research Bulletin 64-61). Princeton, NJ: Educational Testing Service.

CLAUSER, B. E., & MAZOR, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement*, **17**, 31-44.

CLAUSER, B. E., MAZOR, K. M., & HAMBLETON, R. K. (1993). The ef-

fects of purification of matching criterion on the identification of DIF using the Mantel–Haenszel procedure. *Applied Measurement in Education*, **6**, 269-279.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational & Psychological Measurement*, **28**, 61-75.

Cook, L. L., & Eignor, D. R. (1991). NCME instructional module: IRT equating methods. *Educational Measurement*, **10**, 37-45.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning. Standardization and the Mantel–Haenszel method. *Applied Measurement in Education*, **2**, 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, **23**, 355-368.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, **29**, 309-319.

Fidalgo, Á. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel–Haenszel procedures. *Methods of Psychological Research*, **5**, 43-53.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational & Psychological Measurement*, **67**, 565-582.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational & Behavioral Statistics*, **23**, 244-253.

Hauck, W. W. (1979). The large sample variance of the Mantel–Haenszel estimator of a common odds ratio. *Biometrics*, **35**, 817-819.

Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (Research Report RR-85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, **16**, 209-225.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, **14**, 329-349.

Kim, S.-H., & Cohen, A. S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Psychological Measurement*, **16**, 158.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, **32**, 261-276.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, **12**, 365-376.

Li, H.-H., & Stout, W. (1994). SIBTEST: A FORTRAN-V Program for Computing the Simultaneous Item Bias DIF Statistics [Computer program]. Urbana-Champaign, IL: University of Illinois, Department of Statistics.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, **61**, 647-677.

Lord, F. M. (1976). *A study of item bias, using item characteristic curve theory*. Princeton, NJ: Educational Testing Service.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel–Haenszel procedure. *Educational & Psychological Measurement*, **54**, 284-291.

Miller, R. G., Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer.

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

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, **13**, 57-75.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691-692.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, **20**, 257-274.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education*, **14**, 235-259.

Penfield, R. D. (2003). Applying the Breslow–Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. *Alberta Journal of Educational Research*, **49**, 231-243.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, **29**, 150-151.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). Amsterdam: Elsevier.

Philips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel–Haenszel log-odds-ratio estimate. *Biometrics*, **43**, 425-431.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, **53**, 495-502.

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

Raju, N. S. (1995). DFITPU: A FORTRAN program for calculating DIF/DTF [Computer program]. Atlanta: Georgia Institute of Technology.

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, **17**, 1-25.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311-323.

Rogers, H. J., Swaminathan, H., & Hambleton, R. K. (1993). DICHODIF: A FORTRAN program for DIF analysis of dichotomously scored item response data [Computer program]. Amherst, MA: University of Massachusetts.

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

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, **17**, 1-10.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, **16**, 143-152.

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

Shepard, L. [A.], Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational & Behavioral Statistics*, **6**, 317-375.

Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, **18**, 537-555.

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An inte-

grated Bayesian model for DIF analysis. *Journal of Educational & Behavioral Statistics*, **34**, 348-377.

SOMES, G. W. (1986). The generalized Mantel–Haenszel statistic. *American Statistician*, **40**, 106-108.

SPIELBERGER, C. D. (1988). *State–Trait Anger Expression Inventory research edition: Professional manual*. Odessa, FL: Psychological Assessment Resources.

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

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]. Chapel Hill: University of North Carolina, L. L. Thurstone Psychometric Laboratory.

THISSEN, D., CHEN, W.-H., & BOCK, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

THISSEN, D., STEINBERG, L., & WAINER, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-170). Hillsdale, NJ: Erlbaum.

VANSTEELANDT, K. (2000). *Formal models for contextualized personality psychology*. Unpublished doctoral dissertation, K.U. Leuven, Belgium.

WANG, W.-C., & SU, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel–Haenszel method. *Applied Measurement in Education*, **17**, 113-144.

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.

ZUMBO, B. D., & THOMAS, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

(Manuscript received September 25, 2009;
revision accepted for publication March 14, 2010.)