*Journal of Educational and Behavioral Statistics*
*Winter 2005, Vol. 30, No. 4, pp. 443–464*

# Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models

**Wim Van den Noortgate and Paul De Boeck**
*K. U. Leuven, Leuven, Belgium*

*Although differential item functioning (DIF) theory traditionally focuses on the behavior of individual items in two (or a few) specific groups, in educational measurement contexts, it is often plausible to regard the set of items as a random sample from a broader category. This article presents logistic mixed models that can be used to model uniform DIF, treating the item effects and their interaction with groups (DIF) as random. In a similar way, the group effects can be modeled as random instead of fixed, if the groups can be considered a random sample from a population of groups. The models can, furthermore, be adapted easily for modeling DIF over individual persons rather than over groups, or for modeling the differential functioning of groups of items instead of individual items. It is shown that the logistic mixed model approach is not only a comprehensive and economical way to detect these different kinds of DIF, it also encourages us to explore possible explanations of DIF by including group or item covariates in the model.*

Keywords: *differential item functioning, item bias, item response theory, logistic mixed models, random effects*

Differential item functioning (DIF) (e.g., Holland & Wainer, 1993) refers to the phenomenon that, conditionally on the latent ability, the probability of successfully answering a specific item may differ from group to group. The last two decades, differential item functioning has received an increasing attention in educational measurement because DIF may reflect measurement bias (Millsap & Everson, 1993). Items may be biased because they contain sources of difficulty beyond the one(s) of interest, possibly resulting in a discrimination against particular groups (Zumbo, 1999). It is, for example, possible that the results of an intelligence test are systematically lower for a specific minority group, not because the group is less intelligent, but because some items are related to specific knowledge and abilities that are shown more by the majority, while they are not intended to be measured by the test.

Item response theory (IRT) models have been shown to be an interesting tool to understand and model DIF (e.g., Lord, 1980; Thissen, Steinberg, & Wainer, 1993), although the most popular techniques to detect DIF are not IRT based (see Millsap & Everson, 1993, for an overview of these techniques). In IRT models, the probability of a correct response is related to person and item covariates. These covariates often are person and item indicators (dummy covariates), weighted with parameters that are called ability and difficulty, respectively. The popular

two-parameter logistic (2-PL) model, for example, reads as (Hambleton, Swaminathan, & Rogers, 1991):

$$\text{Logit}\left(\pi_{ij}\right) = a_i\theta_j - b_i \quad \text{and} \quad Y_{ij} \sim \text{Bernoulli}\left(\pi_{ij}\right), \tag{1}$$

with

$Y_{ij}$ = the correctness of the response of person $j$ on item $i$, and
$\pi_{ij} = \text{Prob}(Y_{ij} = 1|\theta_j)$ the probability that person $j$ (with ability $\theta_j$) answers item $i$ (with difficulty $b_i$) correctly.

If a specific item shows DIF, the S-shaped curve corresponding to the equation (the *item characteristic curve,* ICC) varies over groups. Not only the location (modeled by the difficulty parameter $b_i$),also the slope (modeled by the discrimination parameter $a_i$) of the curve may vary over groups. In case of a location shift, the probability of a correct answer is always higher for one group than for the other group, while in case of a varying slope, the ICCs may cross, meaning that given a low ability, the probability of a correct answer is higher for a certain group, but given a high ability, it is higher for the other group. Therefore, the former kind of DIF is often called *uniform DIF,* the latter *nonuniform DIF* (Mellenbergh, 1982).

In the following, we discuss the use of logistic mixed models to investigate DIF. We show that the logistic mixed model framework could be regarded as a unifying framework, yielding similar results as traditional DIF methods in situations for which these methods are developed. At the same time, however, the approach suggests extensions of the traditional methods. We show not only that the models we present are very natural tools to detect DIF in a variety of situations, but also that the flexibility of the models has great potential for explaining DIF. We start with discussing logistic mixed DIF models with fixed item and group effects. Next, we discuss DIF models with random item effects and models with random group effects. For simplicity, we focus on uniform DIF, although the extension to models for nonuniform DIF is straightforward. We end with presenting a taxonomy of DIF and a discussion.

## A Logistic Mixed Model for DIF

It has been demonstrated before that IRT models can typically be regarded as logistic mixed models (Adams, Wilson, & Wu, 1997; Kamata, 2001; Mellenbergh, 1994). For example, in the basic IRT model, the Rasch model (Rasch, 1960), the probability of a correct response of person $j$ to an item $i$ is regarded as a function of the person ability ($\theta_j$) and the item difficulty ($b_i$). Persons are commonly regarded as a random sample from a population in which the abilities are independently normally distributed:

$$\text{Logit}\left(\pi_{ij}\right) = \theta_j - b_i \quad \text{with} \quad \theta_j \sim N\left(0, \tau^2\right). \tag{2}$$

The parameters of the model can be estimated using a maximum likelihood procedure, often called the marginal maximum likelihood procedure (MML) (Bock & Aitkin, 1981) because the ability parameters are integrated out. Individual person ability parameters can be estimated afterward, for example using empirical Bayes techniques.

A reformulation of Equation 2 illustrates that the MML formulation of the Rasch model can be regarded as a logistic mixed model, more specifically as a hierarchical two-level logistic model. The model is a repeated measurement model with measurement occasions nested within persons. The logistic mixed model corresponding to the MML formulation of the simple Rasch model, with fixed item parameters and random person parameters, reads as follows:

$$\text{Logit}\left(\pi_{ij}\right) = \sum_{k=1}^{I} \beta_k X_{ki} + u_j, \tag{3}$$

with $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, $X_{ki} = 1$ if $k = i$, 0 otherwise, and $u_j \sim N(0, \sigma_u^2)$.

The logit of the probability of a 1 response is regressed on a set of dummy variables, one for each item. The *i*th item dummy equals 1 if a response is obtained for item *i,* 0 if the response is obtained for another item. Equations 2 and 3 are equivalent. The coefficients of the dummy variables, the $\beta$, are equal to minus the difficulty parameters of Equation 2, the *b,* while the random person effects, indicated by the residual term *u,* equal the person ability parameters, the $\theta$ from Equation 2. Note that although in IRT applications each person usually responds once and only once to each item, the logistic mixed model is also applicable if for some or all person-item combinations there are several scores or if for some combinations there are no observations, assuming that $Y_{ijm}$, the response of person *j* on item *i* on measurement occasion *m* is distributed as $\text{Bernoulli}(\pi_{ij})$.

The logistic mixed model can easily be adapted, for example, by including person or item covariates—other than person or item indicators—as predictor variables in an attempt to describe or explain differences between person abilities or between item difficulties. The framework of the logistic mixed models results in a reformulation of many commonly used IRT models, as well as in several other IRT models that are uncommon but could nevertheless be meaningful in educational measurement (De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). The unknown parameters of logistic mixed models can be estimated using maximum likelihood procedures, as will be discussed later.

DIF analyses are often used to verify if the items in a standardized test do not favor the reference group or a majority group (e.g., males, white people, . . .) in comparison with one or more focal or minority groups (e.g., females, people of color, . . .). To investigate DIF using a logistic mixed model, the main effect of the group membership is modeled by means of one or more additional dummy covariates for the groups. Uniform DIF for a specific item *k* is modeled by defining interaction terms for the item dummy and the group dummy variables. In the following model, *H*-1 dummy variables are included to make a distinction between *H* groups, with the

dummy variables being defined as group specific, and with one of the groups functioning as the reference group (for which all values of the group dummy variables are equal to 0):

$$\text{Logit}\left(\pi_{ijh}\right) = \sum_{k=1}^{I} \beta_k X_{ki} + \sum_{h=2}^{H} \alpha_h G_{hj} + \sum_{h=2}^{H} \gamma_{kh} G_{hj} X_{ki} + u_j, \tag{4}$$

with

$G_{hj}$ = 1 if person $j$ belongs to group h, 0 otherwise;
$\alpha_h$ = the effect of belonging to the focal group compared with the reference group; and
$\gamma_{kh}$ = the $k$th item specific effect of belonging to the focal group $h$ compared with the reference group.

Although in the reference group the item difficulty for item $k$ equals $-\beta_k$, in the focal group $h$, it equals $-(\beta_k + \gamma_{kh})$. Because group differences in overall ability are taken into account by including the group main effects in the model, the deviation of $\gamma_{kh}$ from 0 indicates that there is uniform DIF for item $k$ in focal group $h$ compared to the reference group. To explore or test the DIF of more than one item, interaction terms are included for each of these items. It is also possible to constrain the model to have equal DIF for some items, $\gamma_{kh} = \gamma_{k'h}$, for all $h$ and for $k \neq k'$. If the parameters of the model are estimated using maximum likelihood procedures, the null hypothesis that $\gamma_{kh}$ equals 0 can be tested by comparing the ratio of the coefficient and its standard error of estimation to a standard normal distribution, or to a $t$-distribution with $df = N\text{-}p\text{-}1$, with $N$ equal to the total number of scores, and $p$ equal to the number of explanatory variables (Snijders & Bosker, 1999). Alternatively, the existence of DIF may be evaluated using a likelihood ratio test. This test compares the likelihood of the model with DIF for item $k$ to the likelihood of the model without DIF for item $k$. Because the latter model is a special case of the former, the difference in the deviances, defined as minus twice the natural logarithm of the likelihood, is approximately distributed as a chi-square distribution with 1 degree of freedom (Snijders & Bosker). In their simulation study, Cohen, Kim, and Wollack (1996) confirm the validity of the test for evaluating the group-item interactions in the 2-PL model.

From Equation 4, it can be deduced that the difference in the logit between the reference group and a focal group $h$ for item $k$ equals $\alpha_h + \gamma_{kh}$. Conditional on the latent ability, this difference in the logits, or otherwise stated, the logarithm of the ratio of the odds for the reference group and the focal group $h$, thus equals $\gamma_{kh}$. The DIF statistics from the logistic mixed model, thus, can be interpreted in much the same way as the Mantel–Haenszel (MH) statistic (see Holland and Thayer, 1988, for a description of the close connection between both procedures). In the MH procedure, one estimates the ratio of the odds for the two groups, for each level of the matching variable (often operationalized as the sum score) separately. These odds ratios then are

combined to an overall odds ratio. The MH statistic then is the logarithm of this odds ratio. Under the null hypothesis that there is no DIF, the sampling distribution of the MH statistic is approximately normal, with zero mean and a sampling variance that was approximated by Phillips and Holland (1987). Note that the presence of items showing DIF deteriorates the value of the sum score as a proxy of the latent ability. Therefore, one could use an iterative approach to find a set of "anchor items" that are assumed to be free of DIF and that are used to obtain a better proxy of the latent ability. Similarly, in the mixed model approach, the estimate of the DIF coefficient may be biased if the group main effect $\alpha_h$ is not correctly estimated. This may be the case if the model does not include a DIF parameter for items that actually show DIF. Therefore, one could follow an iterative approach to find a set of anchor items for which no DIF parameters are included in the model.

### *Example 1*

The Flemish Community in Belgium issued a set of attainment targets that specify the basic competencies that are expected from pupils leaving primary education. De Boeck, Daems, Meulders, & Rymenans (1997) explored the assessment of the attainment targets of reading comprehension in Dutch. In the example, we use the data from one of the tests that were developed by the authors. The data consist of the scores of a group of 539 (male and female) pupils from 29 classes out of 15 schools, who answered 57 items assumed to measure 9 attainment targets.

We used the logistic mixed model as well as the MH-procedure to explore if items function differently for male and female pupils. The mixed model includes an interaction term for each item, thus, estimating for each item the item difficulty for male pupils, as well as the additional item difficulty for female pupils. Because the difficulty of each item is estimated for both groups, we dropped the group indicator from the model to make the model identified. To test the DIF parameter for item *k,* we contrasted the parameter with the mean of the other DIF parameters, assuming that this mean approximates the true group differences in ability. Parameter estimation and testing was done using the GLIMMIX macro from SAS (see the Estimation section).

As expected, the estimated DIF parameters from both approaches are very similar. The correlation between the MH statistics and the estimates for the interaction term from the logistic mixed model is equal to .97. Moreover, we found that the standard errors of both kinds of DIF statistics are comparable. As a result, both approaches yield comparable *p* values when using the Wald test for evaluating the DIF statistics. The mean of the *p* values equals .37 and .35 for the mixed model and the MH approach, respectively. If the DIF statistic is significant in one approach, it is also significant or very close to significance in the other approach. In the mixed model approach, the DIF coefficients of 9 items (5 positive, 4 negative) were found statistically significantly different from 0, using an alpha level of .05. According to the MH procedure, 12 items show DIF, including 8 of the items that show DIF according to the mixed model approach. There is a strong correspondence between both approaches in the results of the statistical tests (Cohen's kappa = .71). When using an alpha level of .01, the mixed model approach results in significant DIF

parameters for 5 items, including the only 4 items that show significance in the MH approach (Cohen's kappa = .88).

### Logistic Mixed DIF Models With Random Item Effects

In most common procedures for DIF analysis, statistical tests for DIF are carried out separately for each item and each pair of focal and reference group. The large number of tests results in the problem of false positives (Longford, Holland, & Thayer, 1993). Although in the mixed model approach based on Equation 4 the DIF parameters can be estimated and tested in one analysis (as in the example), the approach is also hampered by the problem of capitalization of chance. Moreover, Longford et al. (1993) argue that "the procedures involving a large number of statistical tests reflect the optimistic view that most of the items contain no DIF, and the few aberrant items should be easy to identify" (p. 176), but that there is evidence that the true DIF parameters may be distributed continuously. This is supported by Figure 1, representing the distribution of the DIF-parameters ($\gamma_k$) from Example 1.

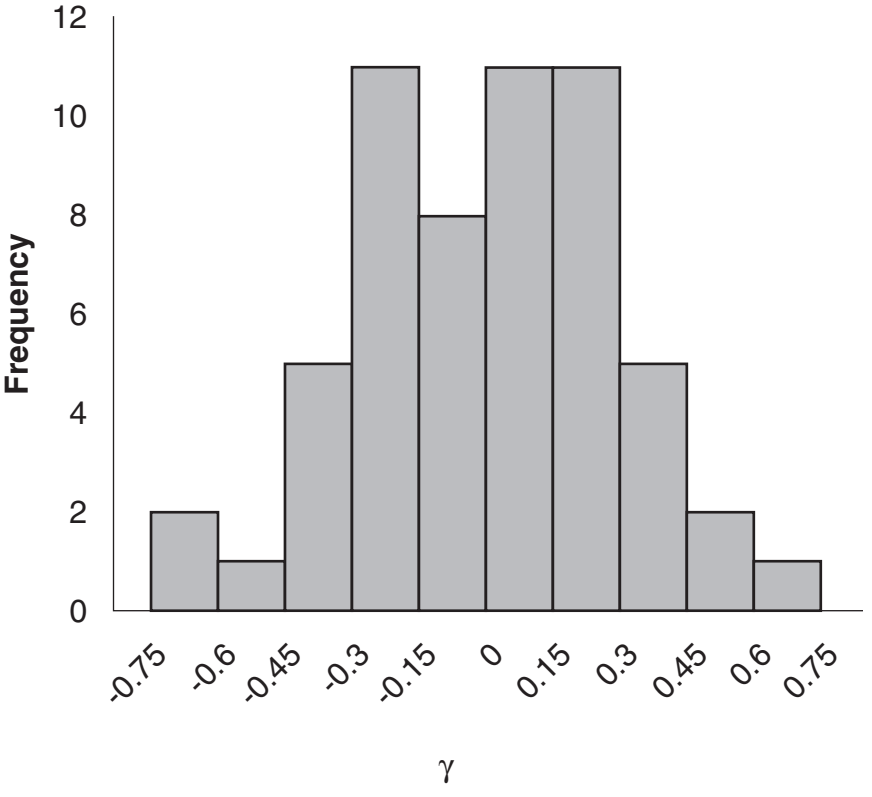


FIGURE 1. *The distribution of the DIF parameters for the attainment target data (Example 1).*

There is also another problem: to model item difficulties and DIF, Equation 4 includes an item dummy variable for each item and an interaction term for each item with potential DIF. This means that the model cannot be extended by including additional item predictors in order to explain differences between item difficulties or to explain DIF. Of course, one could do a second analysis to explain the DIF. For example, Smith and Reise (1998) performed IRT analyses to calculate for each item the difficulty for males and females. They found that the items with positive differences in difficulty between males and females and items with negative differences generally belong to specific factors resulting from a factor analysis on the set of items. Although very informative, this kind of procedure is laborious and rests on a second analysis, independent of the first. Alternatively, one could replace the set of item dummy variables by a smaller set of real item properties. If this is done in Equation 3, the resulting model is equivalent to the linear logistic test model (Fisher, 1973, 1983). Similarly, the interaction terms used to model DIF (interaction of group and item dummy variables) can be replaced by interaction terms of group x item property. These models, however, are very restrictive and often unrealistic because there is no residual term for the item effects, and they consequently assume that the item difficulties or the DIF parameters are exactly the same for all items sharing the same item properties.

Longford et al. (1993) propose to use a random effects model to model the MH statistics, partly to overcome these shortcomings. The statistics are assumed to vary randomly over items, around a mean value that is usually close to 0:

$$\mathrm{MH}_i = \mu + r_i + e_i, \tag{5}$$

with $\mathrm{MH}_i$ as the MH statistic for item *i,* $\mu$ as an overall mean (usually close to 0), and $r_i$ as the random item effect, following a normal distribution with zero mean. The residual $e_i$ is assumed to follow a normal distribution with zero mean and variance $s_i^2$, the sampling variance of $\mathrm{MH}_i$ that is estimated using the Phillips and Holland (1987) formula. If the variance of the random item effects ($\sigma_r^2$) is larger than zero, there is DIF for at least one item. If one wants to know which specific items show DIF, the individual item parameters can be estimated afterward using empirical Bayes techniques. A well-known property of empirical Bayes estimates is that although they are biased, they reduce the *MSE* of estimation. Furthermore, the model can be extended by including person covariates to explore or test possible explanations of DIF. The variance of the random item effects ($\sigma_r^2$) then indicates the residual variance.

An alternative approach is the logistic mixed model approach, based on a model with random item effects. Logistic mixed models with random item effects beside the random person effects were proposed by Van den Noortgate, De Boeck, and Meulders (2003) to explain differences between item difficulties using item properties, without assuming that these properties explain all differences. Such models with crossed random effects are appropriate if items can be considered to be a random sample, and the primary interest of the researcher is not in the particular items,

but in the category they belong to. In a similar way, the logistic mixed model from Equation 4 can be adapted for random item effects. The approach results in more economical models: instead of estimating the individual main and interaction effects for each item separately, only the parameters of the distribution of these random effects are estimated, thus leaving room for the inclusion of additional item properties.

Although fixed item effects are modeled using item dummy variables (Equation 4), random item effects are modeled using parameters that vary randomly over items. Main group effects are again modeled by using dummy variables indicating the group membership:

$$\text{Logit}\left(\pi_{ijh}\right) = \beta_0 + r_{0i} + \sum_{h=2}^{H} \alpha_h G_{hj} + \sum_{h=1}^{H} r_{hi} G_{hj} + u_i, \tag{6}$$

with

$\beta_0$ = the expected logit in the reference group for an "average item";
$r_{0i}$ = the random main effect of item *i;*
$G_{hj}$ = 1 if person *j* belongs to group *h,* 0 otherwise;
$\alpha_h$ = the overall effect of belonging to group *h;*
$r_{hi}$ = the random *i*-th item-specific effect of belonging to group *h;* and
$r_{0i}$ and $r_{hi}$ = following a multivariate normal distribution with means 0.

The item difficulty of item *i* in group *h* equals $-(\beta_0 + r_{0i} + r_{hi})$. If for a specific group *h* the variance of $r_{hi}$ over the items differs from 0, then one may conclude that there is uniform DIF for the random set of items over the groups. The random effects $r_{0i}$ and $r_{hi}$ may be correlated. A positive correlation means that conditionally on the overall performance of the groups, the most difficult items are especially difficult for group *h.*

Note that for each group, there is a set of random item effects. Because for each group the variance of the random effects can be estimated separately, the model allows heterogeneity of variances over groups. To simplify the model and to reduce the computational demands, one could constrain the variances of these sets of random effects to be equal, and covariances to be zero. In case only two groups are compared and homogeneous variances are assumed—a common situation in traditional DIF analyses—the model can be simplified as follows:

$$\text{Logit}\left(\pi_{ij}\right) = \beta_0 + r_{0i} + \alpha G_j + r_{1i} G_j + u_j, \tag{7}$$

with $G_j$ equal to –.5 if person *j* belongs to the reference group, or .5 if person *j* belongs to the focal group.

We chose for an effect coding now, (–.5, .5) because with a dummy coding (0, 1), the item variances are not constrained to be equal. The item difficulty of item *i* equals $-(\beta_0 + r_{0i} - .5\ r_{1i})$ if person *j* belongs to the reference group, $-(\beta_0 + r_{0i} + .5\ r_{1i})$ if per-

son *j* belongs to the focal group. The model implies that the difference in difficulty between the two groups, $r_{1i}$, is normally distributed over the items.

The model with a DIF parameter that varies randomly over items, thus, can be used to detect DIF in a test. If there is evidence for DIF, one could estimate the DIF parameters of the individual items using empirical Bayes techniques. Alternatively, the model can be extended by additional item predictors to explain the DIF. Suppose a certain item covariate W indicates a specific source of difficulty. If conditional on the overall ability, groups differ in the ability to cope with this kind of difficulty, DIF will appear: items that require the additional latent ability will be more difficult to groups that show less of this additional ability. To explore if DIF can be explained by a certain item covariate W, Equation 6 with fixed group effects and random item effects is extended as follows:

$$\text{Logit}(\pi_{ijh}) = \beta_0 + r_{0i} + \sum_{h=2}^{H} \alpha_h G_{hj} + \sum_{h=2}^{H} r_{hi} G_{hj} + \beta_1 W_i + \sum_{h=2}^{H} \beta_h W_i G_{hj} + u_j, \quad (8)$$

with $G_{hj}$ equal to 1 if person *j* belongs to group *h,* 0 otherwise; or, simplified for comparing two groups assuming homogeneous variances, as

$$\text{Logit}(\pi_{ij}) = \beta_0 + r_{0i} + \alpha G_j + r_{1i} G_j + \beta_1 W_i + \beta_2 W_i G_j + u_j, \quad (9)$$

with $G_j$ equal to −.5 if person *j* belongs to the reference group, or .5 if person *j* belongs to the focal group.

Although $\beta_1$ indicates the effect of the item covariate on the item difficulty in the reference group, $\beta_h$ indicates the additional effect of the item covariate in focal group *h.* The random coefficient $r_{hi}$ now represents the residual DIF. If the item covariate explains DIF in group *h,* we expect that the coefficient $\beta_h$ differs from zero and that the variance of $r_{hi}$ over the items will decrease.

One application of the random DIF models with item covariates is to explore a possible differential testlet functioning. A testlet is "an interrelated and integrated group of items, always presented as a single unit" (Wainer & Kiely, 1987; Wainer, Sireci, & Thissen, 1991, p. 197). Wainer et al. (1991) propose to detect differential testlet functioning by considering testlets as polytomous items, and using IRT-based likelihood ratio procedures for DIF. Differential testlet functioning, however, can also be investigated by using a logistic mixed model for random DIF, with testlet dummy variables as item covariates with varying effects over groups. In this way, one can not only investigate differential testlet functioning, but also the possible within-testlet DIF because beside the coefficients for the testlet dummy variables, the unexplained variance of the DIF parameters also is estimated.

### *Example 2*

In the preceding example, we estimated the DIF parameters for the 57 items measuring different attainment targets for reading comprehension. Our analyses

described earlier showed that the DIF parameters seem to be continuously distributed and to follow an approximately normal distribution (Figure 1). Therefore, we reanalyzed the data, using Equation 7, with random effects for persons and items. The fixed group effects are modeled using *G,* indicating the gender (coded –.5 for male, .5 for female pupils), with a fixed and a random coefficient. For simplicity, we assumed that the random main and interaction item effects are independent. Parameters are again estimated and tested using the GLIMMIX-macro from SAS. The results are given in the third column of Table 1.

As can be seen in the table, the effect of gender is 0.18. The logit for male pupils equals 0.51, for female pupils 0.69, corresponding to a probability of a correct response of .62 and .67 respectively. Performing a two-sided Wald test with an alpha level of .05, dividing the coefficient estimate by the corresponding standard error with a standard normal distribution, reveals that the difference between male and female pupils, $\alpha$, is statistically significant ($z = 2.43$, $p = .015$). Moreover, it can be seen that the probability of success varies over pupils and especially over items ($\sigma^2_u = 0.53$ and $\sigma^2_{r0} = 1.26$). Finally, the difference between male and female seems to vary slightly over items. Although the estimate is small, it is statistically significant on a .05 alpha level ($\sigma^2_{r1} = 0.045$, $z = 2.81$, $p = .002$[1]), indicating that, in general, the items show DIF.

This DIF is possibly associated with the type of item. De Boeck et al. (1997) distinguished three types of items depending on the highest level of hypothesized processing. The three level types are: retrieving, structuring, and evaluating. To explore the plausibility of this explanation, we used Equation 9, including the level of processing as an item covariate, coded as 1, 2, 3, for the three levels, respectively. The parameter estimates are found in the last column of Table 1. There is a tendency that the higher the hypothesized level of processing, the lower the probability of a correct response, although this observed relation is not statistically significant on an alpha

TABLE 1
*Estimates of the Parameters of Logistic Mixed Models With Fixed Groups and Random Items to Detect and Explain DIF*

| Parameter | Notation | Equation 7 | Equation 9 |
|---|---|---|---|
| Fixed coefficients | | | |
| Base level | $\beta_0$ | 0.60 (0.15) | 0.95 (0.29) |
| Effect of G (Gender) | $\alpha$ | 0.18 (0.074) | 0.17 (0.10) |
| Effect of W (Level of processing) | $\beta_1$ | | –0.34 (0.24) |
| Interaction of W * G | $\beta_2$ | | 0.0072 (0.065) |
| Variances | | | |
| Random pupil effects | $\sigma^2_u$ | 0.53 (0.040) | 0.53 (0.040) |
| Random item effects | $\sigma^2_{r0}$ | 1.26 (0.24) | 1.24 (0.24) |
| Random effects of G over items | $\sigma^2_{r1}$ | 0.045 (0.016) | 0.046 (0.017) |

*Note:* Standard errors are given within parentheses.

level of .05 ($\beta_1 = -0.34$, $z = -1.42$, $p = .16$). Moreover, it is clear from the table that the level of processing must not be considered as a source of DIF: not only is the interaction term between the gender and the level of processing not statistically significant ($\beta_2 = 0.0072$, $z = 0.11$, $p = .91$), the variance of $r_{1i}$ over items, furthermore, does not decrease (0.045 vs. 0.046).

The Longford et al. (1993) approach gives similar results: the variance of the random item effects (Equation 5) equals 0.056 with a standard error of 0.020 (to be compared with the variance estimate of 0.045 and the standard error of 0.016, Table 1). Moreover, empirical Bayes estimates of the DIF parameter for the individual items are also comparable for both procedures, with a correlation between both equal to .98. Standard errors are comparable as well. If the item predictor Level of Processing is inserted in the model, the variance becomes 0.058 with a standard error of 0.020. The weight of the item predictor is 0.028 (with a standard error of 0.072), which is different from the corresponding coefficient in the logistic mixed model approach (0.0072), but is also very small and statistically far from significant.

## Logistic Mixed DIF Models With Random Group Effects

In social research, hierarchical data structures are rather common (Bryk & Raudenbush, 1992). Individual study participants usually belong to different sorts of groups (e.g., schools, households, countries, . . .). Often the researcher is not interested in the specific groups found in the study sample, and other groups could have been obtained if another sample was drawn. In the increasingly popular multilevel models, groups are considered to be random, and group effects are modeled by means of additional random parameters. Especially when the number of groups is large, and the groups could be regarded as a random sample from a certain population of groups, rather than as a fixed set, the researcher may focus on the characteristics of the population of groups, rather than on the individual groups. To detect DIF in such a multilevel setting, the logistic mixed model with fixed item effects (Equation 4) can be adapted by defining random parameters for the main group effects and for the interaction effects:

$$\text{Logit}\left(\pi_{ijh}\right) = \sum_{k=1}^{I} \beta_k X_{ki} + v_{0h} + v_{kh} X_{ki} + u_j \tag{10}$$

with

$v_{0h}$ = the random effect of group $h$ on the overall performance,
$v_{kh}$ = the random effect of group $h$ on the difficulty of item $k$, and
$v_{0h}$ and $v_{kh}$ = following a multivariate normal distribution with means 0.

In Equation 10, the difficulty of item $k$ equals $-(\beta_k + v_{kh})$. If the variance of $v_{kh}$ over the groups differs from 0, then the difficulty of item $k$ depends on the group, or in

other words, there is uniform DIF for item $k$ over the random set of groups. Note that the random effects $v_{0h}$ and $v_{kh}$ may be correlated. A positive correlation means that, in general, the estimated difficulty of item $k$ is higher for the less able groups than for the most able groups, even after correction for the overall group ability.

If, in addition, items are also considered to be random, no dummy variables are used to indicate the item (no $X$, as in Equation 10), but parameters are included that vary randomly over items ($r_{0i}$ instead of $\beta_k X_{ki}$) and over pairs of groups and items ($t_{ih}$ instead of $v_{kh}X_{ki}$):

$$\text{Logit}\left(\pi_{ijh}\right) = \beta_0 + r_{0i} + v_{0h} + t_{ih} + u_j. \tag{11}$$

The difficulty of item $i$ in group $h$ equals $-(\beta_0 + r_{0i} + t_{ih})$. The DIF parameter, $t_{ih}$, is normally distributed over groups and items with mean 0, and variance $\sigma_t^2$.

A consequence of considering the group effects as random, is that one can evaluate to which degree DIF is affected by a group characteristic. When using models with fixed group effects on DIF, this cannot be done without assuming that the differential effect of the group characteristics included in the model explains all DIF. With group effects on DIF defined as random over groups, residual DIF terms are incorporated in the model (see $v_{kh}X_{ki}$ in Equation 12 and $t_{ih}$ in Equation 13). The logistic mixed model with a group characteristic $Z$ with an effect that depends on the item, whereas, the items have fixed effects reads as:

$$\text{Logit}\left(\pi_{ijh}\right) = \sum_{k=1}^{I} \beta_k X_{ki} + v_{0h} + v_{kh}X_{ki} + \alpha_1 Z_h + \alpha_2 X_{ki} Z_h + u_j \tag{12}$$

to explain DIF for a specific item $k$, or if items have random effects as

$$\text{Logit}\left(\pi_{ijh}\right) = \beta_0 + r_{0i} + v_{0h} + t_{ih} + \alpha_1 Z_h + r_{1i} Z_h + u_j, \tag{13}$$

extending Equations 10 and 11, respectively.

Equation 13 shows that the effect on the group ability because of an increase of the group covariate $Z$ with one unit equals $(\beta_1 + r_{1i})$. If this sensitivity for the additional dimension is stronger for some items than for other items, $r_{1i}$ will vary over items, explaining (part) of the DIF. If the group covariate $Z$ induces DIF, this is indicated by a statistically significant variance of $r_{1i}$ and a decrease of the variance of $t_{ih}$ over items and groups. Note that because in Equation 13, item effects are considered to be random, the model can further be extended by the inclusion of item covariate(s) with effects that vary over groups.

*Example 3*

To illustrate the DIF models with random group effects, we analyze part of the data that stem from the "*Longitudinaal Onderzoek Secundair Onderwijs*" (*LOSO*) (Longitudinal Research in Secondary Education), which was performed in Belgium by the research team of Van Damme (Van Damme, De Fraine, Van Landeghem,

TABLE 2
*Estimates of the Parameters of Logistic Mixed Models With Random Groups and Random Items to Detect and Explain DIF*

| Parameter | Notation | Equation 11 | Equation 13 |
|---|---|---|---|
| Fixed coefficients | | | |
| Base level | $\beta_0$ | –0.92 (0.19) | –0.93 (0.18) |
| Effect of Z (School type) | $\alpha_1$ | | 0.19 (0.18) |
| Variances | | | |
| Random pupil effects | $\sigma^2_u$ | 0.28 (0.027) | 0.28 (0.027) |
| Random school effects | $\sigma^2_{v0}$ | 0.15 (0.054) | 0.15 (0.051) |
| Random item effects | $\sigma^2_{r0}$ | 1.23 (0.27) | 1.18 (0.26) |
| Random interaction of school*items | $\sigma^2_t$ | 0.27 (0.026) | 0.22 (0.023) |
| Random effects of Z over items | $\sigma^2_{r1}$ | | 0.27 (0.080) |

*Note:* Standard errors are given within parentheses.

Opdenakker, & Onghena, 2002) and funded by the Department of Education of the Ministry of the Flemish community. We used data regarding the mathematics achievement at the end of the 2nd year of the general track of secondary education, measured using 44 items in a sample of 524 pupils, nested in 33 secondary schools. We first explored if the item difficulty varies over schools. To evaluate this kind of DIF, we used Equation 11, regarding items as well as groups as random. Parameter estimates obtained with the GLIMMIX macro of SAS are found in the third column of Table 2.

The intercept $\beta_0$ equals –0.92. Taking the antilogit of this value results in .28, indicating that an average person has a probability of .28 for a correct answer on an average item. There are, however, substantial differences between items ($\sigma^2_{v0} = 1.23$) and between pupils ($\sigma^2_u = 0.28$). The variance over schools ($\sigma^2_{v0} = 0.15$) is smaller than the pupil variance, but is not negligible. In school effectiveness studies (e.g., Hill & Rowe, 1996; Scheerens & Bosker, 1997), it is a common finding that only a relatively small part of the variance between pupil scores is attributable to differences between classes or schools. More important for our application here, is that the item difficulties seem to vary over schools. The random interaction term between schools and items is statistically highly significant ($\sigma^2_t = 0.27$, $z = 10.38$, $p < .001$). To obtain an idea of the size of DIF, we look at the effect of minus and plus one standard deviation of $t_{ih}\left(\sqrt{0.27} = 0.52\right)$ on the probability of success for an item of an average difficulty in a class of an average ability (for which the probability without DIF would be .28). The resulting probabilities are .19 and .40, respectively. Thus, the effect of DIF seems to be substantial, or in general items difficulties seem to differ substantially in different classes.

How can this DIF be explained? The data set includes data from two kinds of schools, Catholic and public, with a slightly different curriculum with respect to mathematics. Possibly, some items deal with mathematical topics that have received less attention in one of the schools than in the other, resulting in a DIF. To investigate if the DIF results from item-varying differences between schools,

we used Equation 13, including an indicator for the kind of school (Catholic = 1, public = 0). For simplicity, we assumed in the example that the random main and interaction item effects ($r_0$ and $r_1$ respectively) are independent. The results are found in the last column of Table 2. We conclude that the DIF is explained partly by the type of school: not only has the school type an effect that varies in a statistically significant way over items ($\sigma^2_{r1} = 0.27$, $z = 3.38$, $p < .001$), also the variation over items and schools ($\sigma^2_t$), indicating DIF, is reduced from 0.27 to 0.22, a difference that is relatively large compared to the standard error of the estimates. The proportion of DIF that is explained by the school type is .20 (= 0.27 − 0.22/0.27), but the residual DIF is still statistically highly significant ($z = 9.57$, $p < .001$).

## Classification of DIF

DIF refers to the differential functioning of items over groups of persons. As discussed above, both the items and groups may be regarded as fixed or as random. The distinction between fixed and random groups and between fixed and random items results in four kinds of DIF (Table 3). We discussed logistic mixed models for each of the cells of Table 3. The applications concerned cells, a, c, and d from Table 3 (Examples 1, 2, and 3, respectively).

Note that there are other kinds of interaction effects that could be regarded as DIF. For example, it is possible that there is a differential functioning of *groups* of items, instead of, or in addition to, a differential item functioning of individual items. Moreover, it is not uncommon that items function differently over persons belonging to the same group. This means that there are also four types of interaction depending upon the level one is considering (person or person groups, items or item groups) (Table 4).

All models we discussed before describe DIF that is situated in cell b from Table 4. In an analogous way as we did for this cell, we can model DIF for the other three cells. For example, if the differential functioning of groups of items over groups of persons is considered (cell d from Table 4), the groups of items as well as the groups of persons can be regarded as random or as fixed. Of course, different kinds of DIF can be modeled simultaneously, by including several interaction terms corresponding to different cells in Table 4. Note that the models we proposed to explain DIF based on an item property (Equations 8 and 9) could be regarded as simultaneously modeling two kinds of DIF because the item covariate W, in fact, defines a (fixed) set of groups of items: in addition to the interaction term between the items

TABLE 3
*Four Kinds of DIF Models Based on the Status of the Effects of the Groups of Persons and of the Items*

| | Groups of Persons | |
|---|---|---|
| Items | Fixed | Random |
| Fixed | (a) | (b) |
| Random | (c) | (d) |

TABLE 4
*Four Kinds of DIF, Distinguished Based on the Components of the Interaction Term*

| | Persons | |
|---|---|---|
| Items | Individual | Group |
| Individual | (a) | (b) |
| Group | (c) | (d) |

and groups of persons, an interaction term is included between groups of items and groups of persons. Similarly, the models including a person covariate (Equations 12 and 13) describe two kinds of DIF.

Finally, we note that sometimes a further grouping of groups of items and/or of groups of persons is possible, and IRT models can be adapted to model variation at these additional levels (Van den Noortgate & Paek, 2004). For instance, in the examples described above, pupils are grouped in classes, which in turn are grouped in schools, but only school effects were studied. The logistic mixed models can further be adapted to investigate if items function differently over persons, over classes, and over schools, by including three kinds of interaction terms in one model.

## Estimation

Parameters of mixed models are commonly estimated using maximum likelihood procedures. To obtain the likelihood function of the complete data, the random effects are integrated out. Unfortunately, for the logistic mixed model, unlike for the linear mixed model, this likelihood cannot be written in a closed form. One solution is to approximate the likelihood with linearization techniques. This is done in the iterative quasi-likelihood procedures (Breslow & Clayton, 1993). Given starting values for the unknown parameters, the first one or two terms of a Taylor series expansion are used to linearize the logistic function, resulting in a standard linear mixed model. After linearizing the logistic function, the parameters are estimated using estimation procedures for linear mixed models, and the estimates are used for a new Taylor expansion of the logistic function, and so on. Parameter estimates are improved in each successive iteration. In the marginal quasi-likelihood (MQL) procedure, the Taylor expansion uses only the parameter estimates for the fixed part. In the penalized quasi-likelihood (PQL) procedure, current estimates of the random effects are used as well, and, therefore, usually give more accurate results (Rodriguez & Goldman, 1995; Goldstein & Rasbash, 1996). Quasi-likelihood procedures are used in the SAS-macro GLIMMIX (Wolfinger & O'Connell, 1993), and in popular specialized software for hierarchical mixed models, such as VARCL (Longford, 1993), HLM (Bryk, Raudenbush, Cheong, & Congdon, 2001), and MLwiN (Goldstein et al., 1998). Because the models with random item effects and random person effects are not strictly hierarchical in nature, they must be reformulated before software for hierarchical models can be used (Van den Noortgate et al., 2003). An important disadvantage of the quasi-likelihood procedures is that not the "real" likelihood, but

only an approximate likelihood is used, so that model fit statistics based on the likelihood are also only approximate and may not be used for testing the model fit (Hox, 2002). That is the reason why in the examples, we did not use a fit statistic (e.g., deviance, AIC, . . .) to evaluate the absolute or relative fit of a model. Instead of using fit statistics to select the model with the best model fit, one can build the logistic mixed model using a stepwise procedure, starting from a simple descriptive model. In each step, additional parameters are added. The significance of the parameters is evaluated using the Wald test, comparing the estimates divided by their standard errors with a standard normal distribution. Parameters that do not seem to be significantly different from zero possibly are omitted in a next step. Model building and specification is more extensively discussed by, for example, Snijders and Bosker (1999). A second disadvantage of the quasilikelihood procedures is that if the number of items is small, estimates of the fixed parameters and covariance parameters may be negatively biased.

To avoid both problems with the quasi-likelihood procedures, one can try to approximate the marginal density by numerical integration, for example using the Gauss-Hermite Quadrature, as implemented in proc NLMIXED of SAS (SAS Institute, 2000). A drawback of numerical integration is the computational intensity, making this procedure much slower than the quasi-likelihood procedures. Moreover, for reasons of computational demands, (to date) only one kind of random effects can be defined in NLMIXED. One can, for example, define one or more random effects over persons, or one or more random effects over items, but not both kinds of random effects at the same time. The random effects may be multiple if the effects are of one kind, but in practice sometimes problems arise when many random effects are defined or when the random effects are highly correlated (Kachman, 2001). The limitation of NLMIXED to only one type of random effect means it can be used for estimating the parameters of logistic mixed models for detecting DIF if groups and items are regarded as fixed (cell a from Table 3), but not if besides the persons, also groups or items are considered to be random (other three cells from Table 3).

A flexible but relatively complex approach for estimating the unknown parameters of logistic mixed models, possibly with crossed random effects, is the use of Bayesian techniques (Van den Noortgate et al., 2003). For details about the Bayesian approach, we refer to Gelman, Carlin, Stern, and Rubin (1995); for an educational measurement application with a logistic mixed model with random item and random person effects, see Janssen, Tuerlinckx, Meulders, and De Boeck (2000). Although the Bayesian approach has its strengths, we have concentrated here on estimation procedures implemented in widely available software that is easily accessible for practitioners.

## Discussion and Conclusions

Since the 1980s, the popularity of mixed effects or multilevel models has increased exponentially in several research domains, for example, in education, psychology, and biomedical sciences. Also in IRT applications, mixed models set in

(see, e.g., Adams, Wilson, & Wu, 1997; De Boeck & Wilson, 2004; Kamata, 2001; Mellenbergh, 1994). The mixed model perspective suggests using random group effects in item response models if persons belong to groups that can be regarded as a random sample of groups, and the researcher is not primarily interested in the specific group effects, but rather in the distribution of these effects. Furthermore, item response models can include random item effects (crossed with the random person effects), as discussed by Van den Noortgate et al. (2003).

These important evolutions in item response modeling suggest new models and approaches for DIF. Traditionally, an item is said to show DIF if conditionally on the ability, the probability of correctly answering the item depends on the group the person belongs to and models and techniques for DIF treat both the items and the groups as fixed. In this article, we argued that it is sometimes plausible to consider items and/or groups to be a random sample and that doing so provides us with a flexible modeling tool. We suggested a taxonomy of DIF in terms of random and fixed effects and in terms of the level at which it is studied: items or groups of items, persons or groups of persons. We showed that for all resulting kinds of DIF, which all can have a substantive meaning in educational measurement, specific instantiations of the general logistic mixed model can be formulated.

Although in traditional DIF analyses DIF is considered specific and limited, this is not true if items or groups are considered random, a possibility that is explicitly mentioned in the taxonomy. For example, the effect of fixed groups may be modeled as varying at random over items, following a normal distribution. Although the assumption of random variation applies to all items, this assumption is not really restrictive in practice. If there is DIF for just one item, this would still show in an empirical Bayes estimate of the DIF, even while a normal distribution assumption is made. The assumption functions as a prior distribution, but the posterior distribution (given the data) as derived from empirical Bayes estimates can turn out to deviate from the normal distribution. If this deviation from a normal distribution is quite serious, it is recommended to use estimation procedures that are not based on the normality assumption, for example, a Bayesian procedure. A description of checking the assumptions of mixed models and possible remedies for violation of the assumptions are given, for example, by Snijders and Bosker (1999).

Although still uncommon in DIF analysis, the idea of regarding DIF as random over items is not new: it was proposed before by Longford et al. (1993). The Longford et al. approach is extended by using the logistic mixed model approach in the following ways: the logistic mixed model approach allows the investigation of (a) DIF for a set of subpopulations (e.g., schools) by using random group parameters, (b) the differential functioning of groups of items over groups of persons, or of items over categories of groups of persons, (c) several kinds of DIF in a single analysis, and (d) nonuniform DIF. As to the latter, logistic mixed models could include a discrimination and/or a guessing parameter (Van den Noortgate et al., 2003). To model nonuniform DIF, the discrimination parameter is allowed to vary over groups. The parameter may be estimated for specific items and groups, but may also be assumed to vary randomly over items and/or groups, possibly conditional on item

and group covariates. Finally, we note that the approach of Longford et al. consists of two steps: in a first step, the MH statistics are calculated together with their estimated standard errors. In a second step, the MH statistics and corresponding standard errors are modeled using a random effects model, assuming that the standard errors are known. In the logistic mixed model approach, the raw data are modeled directly and all parameters are estimated in one single analysis.

We note that in the MH procedure, an observed variable (usually the total test score) is used as a stratifying variable for examining the relation between group and outcome, in order to distinguish DIF from ordinary group differences. Also in the logistic regression procedure (Swaminathan & Rogers, 1990), one conditions on an observed variable that serves as a proxy to the latent ability. The logit of the probability of a correct answer on a specific item is regressed on the observed variable, the group membership, and their interaction. A significant group effect and interaction effect are indications for uniform and nonuniform DIF, respectively. In these procedures, the observed total test score is usually used as a proxy, although alternative matching criteria may be more appropriate. Clauser, Nungester, Mazor, & Ripkey (1996), for example, discuss extensions of the basic logistic regression DIF model, to match on subtest scores or on multiple subtest scores, which may be appropriate in case of multidimensional tests. Clauser, Nungester, and Swaminathan (1996) show that matching on a person background variable besides the total test score could reduce substantially the items identified as showing DIF, in case the background variable and the group are confounded. In the logistic mixed models or IRT models, conditioning on the ability is done by relating directly the observed responses to the latent trait. Similar alternative types of matching are possibly. The models we proposed assume unidimensionality: only the intercept of the models has a random effect over persons. It is, however, also possible to define an additional random person effect for an item covariate that indicates a specific kinds of difficulty, resulting in multidimensional logistic mixed models (Rijmen & Briggs, 2004). One then controls for group differences on this additional dimension by including the fixed interaction effect between the group and the item covariate. We also showed above how one can control for the effect of person or group covariates on observed DIF, by modeling their main effects and their interaction effects with the items being investigated.

In general, we see three major strengths of the logistic mixed models approach for modeling DIF. First, logistic mixed models are easy to understand and very flexible. As demonstrated above, the logistic mixed models can be adapted to the kind of DIF that is to be modeled. Several extensions of the basic DIF models are possible, for example, by including covariates for each kind of unit or by defining additional levels, allowing the researcher to adapt the model to her or his research interests. A result of the amazing flexibility of the logistic mixed model is that the approach is applicable in specific situations for which valuable approaches, such as the MH approach, exist (yielding similar results), but also in situations for which DIF methods are not well developed. We believe that extending approaches and formulating unifying frameworks can be of theoretical and practical value when it comes to selecting a

method. It also gives us the comfort of staying within the same kind of global model and software.

Second, the logistic mixed model framework allows considering items and/or groups to be random, resulting in more economical models: under the assumption of normality, only the mean(s) and variance(s) of the population distribution(s) of the random effects are estimated, instead of all individual effects separately. The variance of the random DIF parameter summarizes possible DIF in the test, and the statistical test of this variance can be considered as an overall test for DIF. The parameters for individual groups or items could be estimated and tested afterward using empirical Bayes techniques, efficiently borrowing strength from other groups or items. By using an overall test, one avoids the problem of capitalization of chance. At the same time, the test can be more powerful if several items tend to show DIF, but the individual DIF parameters are statistically not significant.

Third, using random DIF effects, a hypothesized explanation for DIF can be included in the model through the effects of covariates, without requiring that the explanation is perfect. The covariates can relate to items, persons, or (higher-level) groups of items or persons.

Although user-friendly and commonly available statistical software can be used, the estimation of the unknown parameters is a point of concern. The Gaussian Quadrature procedure to obtain maximum likelihood estimates is not feasible if there is more than one kind of random effect. Even for the much faster quasi-likelihood procedures, estimating the unknown parameters of the models with several kinds of random effects is a hard job: each of the analyses performed for the example based on a model with crossed random (item and person) effects took several hours on a Pentium III 1.5 GHz, using the GLIMMIX macro from SAS. Analyses of larger data sets (possibly with hundreds of items) are likely to be problematic. Because of the increasing computing power of personal computers, however, this problem can be expected to become less onerous in the future. A final problem associated with using the multilevel software, as well as the GLIMMIX macro, is that they fail to estimate the parameters of a model, including a guessing parameter. Estimation for these models can be done by more complex Bayesian estimation procedures.

## Note

[1]Because negative variance estimates are truncated at zero, the null hypothesis of no variance is tested against the one-sided alternative hypothesis that the variance is larger than zero (Verbeke & Molenberghs, 2000). Therefore, two-sided *p* values are conservative, and one-sided *p* values for the variance parameters are used instead.

## References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25.

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

Bryk, A. S., Raudenbush, S. W., Cheong, Y. F., & Congdon, R. T. (2001). *HLM5: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International, Inc.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 203–215.

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., & Wollack, J. A. (1996). An investigation of the Likelihood Ratio Test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26.

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

De Boeck, P., Daems, F., Meulders, M., & Rymenams, R. (1997). *Ontwikkeling van een toets voor de eindtermen begrijpend lezen* [*Construction of a test for the educational targets of reading comprehension*]. Leuven/Antwerpen (Belgium): Katholieke Universiteit Leuven/Universiteit Antwerpen.

Fisher, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fisher, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3–26.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 159,* 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., et al. (1998). *A user's guide to MLwiN.* Multilevel Models Project, University of London.

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

Hill, P. W., & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7,* 1–34.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Hox, J. (2002). *Multilevel analysis. Techniques and applications.* Mahwah, NJ: Erlbaum.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Kachman, S. D. (2001). *Generalized linear mixed models* [Web Page]. URL Nebraska Institute of Agriculture & Natural Resources, Department of Biometrics Web Site: http://biometry.unl.edu/faculty/Steve/GLMM/2001 [2002, June 25].

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79–93.

Longford, N. T. (1993). *Random coefficient models.* Oxford: Clarendon Press.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–118.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research.

Rijmen, F., & Briggs, D. (2004). Multidimensional person variance and latent item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized and nonlinear approach* (pp. 247–265). New York: Springer.

Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 158,* 73–90.

SAS Institute. (2000). *SAS/STAT User's Guide, Version 8.* Cary, NC: SAS Institute, Inc.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford: Elsevier.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75,* 1350–1362.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage.

Swaminathan, H., & Rogers, H. J. (1990). *Detecting differential item functioning using logistic regression procedures, 27,* 361–370.

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: Erlbaum

Van Damme, J., De Fraine, B., Van Landeghem, G., Opdenakker, M.-C., & Onghena, P. (2002). A new study on educational effectiveness in secondary schools in Flanders: An introduction. *School Effectiveness and School Improvement, 13,* 383–397.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28,* 369–386.

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized and nonlinear approach.* New York: Springer.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48,* 233–243.

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.

## Authors

WIM VAN DEN NOORTGATE is Assistant Professor, Department of Educational Sciences, Katholieke Universiteit Leuven, Vesaliusstraat 2, B-3000 Leuven, Belgium; Wim.VandenNoortgate@ped.kuleuven.ac.be. His areas of specialization are multilevel analysis, meta-analysis, and item response theory.

PAUL DE BOECK is Professor of Psychology, Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102, B-3000 Leuven, Belgium; Paul.DeBoeck@psy.kuleuven.be. His areas of specialization are psychometrics, tests, and personality and intelligence.

Manuscript received October 21, 2002
Revision received May 10, 2004
Accepted May 12, 2004