RIM: A Random Item Mixture Model to Detect Differential Item Functioning

Sofie Frederickx and Francis Tuerlinckx

K.U. Leuven

Paul De Boeck

University of Amsterdam

David Magis

University of Liège

In this paper we present a new methodology for detecting differential item functioning (DIF). We introduce a DIF model, called the random item mixture (RIM), that is based on a Rasch model with random item difficulties (besides the common random person abilities). In addition, a mixture model is assumed for the item difficulties such that the items may belong to one of two classes: a DIF or a non-DIF class. The crucial difference between the DIF class and the non-DIF class is that the item difficulties in the DIF class may differ according to the observed person groups while they are equal across the person groups for the items from the non-DIF class. Statistical inference for the RIM is carried out in a Bayesian framework. The performance of the RIM is evaluated using a simulation study in which it is compared with traditional procedures, like the likelihood ratio test, the Mantel-Haenszel procedure and the standardized p -DIF procedure. In this comparison, the RIM performs better than the other methods. Finally, the usefulness of the model is also demonstrated on a real life data set.

When assembling a test, an important objective is to draw valid inferences about the construct that one intends to measure. An obstacle on this road to valid inferences is differential item functioning (DIF). DIF is present when examinees from different groups with the same ability have a different probability of answering an item correctly. If an item is diagnosed with DIF, it does not measure the concept that it is intended to measure adequately, but instead it measures some additional nuisance dimensions (Ackerman, 1992). Hence, it is important to detect DIF items so that appropriate actions can be taken to solve the problem (e.g., excluding the item from the test, rewriting the item).

Several methods for detecting DIF have been proposed in the literature. In this study we compared a method we propose to three commonly-used ones: the likelihood ratio test (LRT) procedure (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988), the Mantel-Haenszel (MH) approach (Holland & Thayer, 1988), and the standardized p -DIF (ST p -DIF) procedure (Dorans & Kulick, 1986). For a more thorough overview and more techniques, we refer to Holland and Wainer (1993). The aforementioned approaches will be reviewed briefly in the following paragraphs (for a more general overview, see, e.g., Zumbo, 2007).

The MH approach (Holland & Thayer, 1988) is a test of the null hypothesis that there is no relationship between two categorical variables given a stratifying variable. More specifically, it tests whether the odds of getting an item correct are the same across all ability levels of the examinees. If the odds for the two groups differ, it is concluded that the item contains DIF. As a proxy to ability level, one often works with the sum score (i.e., the total test score). A related method is the ST *p*-DIF method (Dorans & Kulick, 1986) where the weighted difference in item performance between the focal group (focal refers to the group of interest) and the reference group (the group with whom the focal group is compared) matched on the underlying ability is calculated. An item is deemed to contain DIF when performance on it by examinees from different groups differs while controlling for ability. The test sum score (or a limited sum score range) is used as a proxy for the ability of an examinee. The ST *p*-DIF is an index that can take values between -1 and $+1$ because it is based on a difference between two proportions. Positive values indicate that the item favors the focal group, whereas negative values indicate that the focal group has a disadvantage on the item. An advantage of both the MH and the ST *p*-DIF technique is that these methods do not make use of a parametric model for the item responses and therefore rely on a minimal set of underlying assumptions. On the other hand, the LRT procedure (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988) makes use of an IRT model. All items are tested one at a time in this procedure: apart from the item under investigation, the other items are considered as anchor or DIF-free items. Then, a likelihood ratio test is performed in which two models are compared: the augmented model in which the investigated item is allowed to vary between the groups, and the compact model in which the item is restricted to be the same in all groups. If the likelihood ratio test turns out to be statistically significant, it is concluded that the studied item shows DIF.

The outlined methods are well designed, they are easy to implement, they have been studied extensively in the literature, and they have been applied widely for DIF detection. However, there are two aspects common to all of them that may be improved upon: the necessary selection of anchor items, and the iterative nature of the procedure. First, all the methods described above require the selection of a set of anchor items to detect DIF. In the LRT, we have already made this explicitly clear, but anchor items are also used for the two other methods. Both in the MH method as in the ST *p*-DIF method, one matches the groups with respect to ability, which is approximated by a summary measure based on the non-DIF items also called anchor items. When there are items having DIF among these anchor items this may affect the results. Finch (2005) demonstrated that the type I error of the MH procedure and the LRT increases and the power decreases when the anchor items are not completely DIF-free. Also, Navas-Ara and Gomez-Benito (2002) showed that when there are items exhibiting DIF in the anchor set, the estimates of the underlying abilities in the IRT model are biased. Hence, defining items as DIF-free and hence as anchor items when they actually contain DIF is not without consequences. Furthermore, it has been shown that experts are generally poor at predicting which items would function differently across groups (Engelhard, Hansche, & Rutledge, 1990; Plake, 1980; Sandoval & Miille, 1980).

To overcome the issue of contamination of the matching variable, several authors (Candell & Drasgow, 1988; Clauser, Mazor, & Hambleton, 1993; Fidalgo, Mellenbergh, & Muniz, 2000; Wang & Su, 2004) suggested performing an iterative procedure to purify the matching variable, which is now commonly called item purification. In the case of non-IRT methods, this implies that one computes the total test score of the examinees (i.e., the matching variable) iteratively by discarding the previously detected DIF items. For IRT methods, item purification comes down to an iterative rescaling of the item parameters (in which DIF items are removed in the rescaling step). When two successive runs of the method provide the same classification of the items (as DIF or non-DIF) the process stops. Making use of an iterative procedure to identify a set of items is frequently used in the DIF research and this brings us to a second issue that could be improved upon.

There are several iterative procedures in DIF detection. For instance, analyses are typically performed for each item separately or for a small subset of items. Moreover, finding a DIF-free anchor item set is also done in a stepwise fashion. A problem with these iterative approaches is that the order in which the steps are taken can have an influence on the result and that it may become very difficult to study the underlying causes of the DIF because an encompassing model is lacking.

Building on ideas first put forward in De Boeck (2008), we propose in this paper an alternative methodology to detect DIF that does not require the prespecification of a set of anchor items (but instead automatically provides a set of anchor items) and that aims at an automatic classification of DIF and non-DIF items. Unlike the above-mentioned methods, it is not necessary to prespecify a set of anchor items. The items belonging to the non-DIF class will act as anchor items. So we have an automatic specification of a set of anchor items. This way, the drawbacks associated with the more standard methods are avoided. Also, our method can be used as a way to obtain a purified set of anchor items. This set can then be a starting point from which more traditional methods can be used. Contrary to the MH and ST p -DIF approaches (but similar to the LRT), our method is based on an underlying IRT model, in this case the Rasch model. In our model it is assumed that the item difficulties are random effects but instead of restricting all item difficulties to be equal across the person groups, some of them are allowed to vary across the groups. To decide which items have an equal difficulty in both groups and which ones have different values, a mixture model for the item difficulties is used. We discuss the two main innovative aspects of our model, the random items and the item mixture, in greater detail in the following paragraphs.

First, the item difficulties are random variables. This may seem unusual because in most IRT models the item difficulties are taken to be fixed. However, random items have been proposed already in the framework of generalizability theory (Brennan, 2001) or, more generally, random factors appear in random or mixed effects ANOVA (Maxwell & Delaney, 2004). In these cases, the items or factor levels are taken to be a sample from a population and one aims at generalization of the results to this wider population or universe of items. Moreover, considering items to be random is closely related to the idea of exchangeability among the items (see also Lindley & Novick, 1981; Snijders, 2005). In this context, it means that we have no specific interest in the sampled set of items out of a broader category of items

(e.g., items concerning spatial ability). We would also be satisfied with any other sample from the same universe of items; only items that are diagnosed with DIF are set apart from the rest. Note that some authors even advise always using random effect for categorical predictors (see Gelman & Hill, 2007, p. 246). Random items have been previously introduced into IRT by Janssen, Tuerlinckx, Meulders, and De Boeck (2000), Glas and van der Linden (2001), Van den Noortgate, De Boeck, and Meulders (2003), Chaimongkol, Huffer, and Kamata (2006), Chaimongkol, Huffer, and Kamata (2007), De Boeck (2008), Gonzalez, De Boeck, and Tuerlinckx (2008), and Kamata, Bauer, and Miyazaki (2008). Assuming that both the item and the person parameters are random leads to a so-called crossed random effects model.

Second, instead of using the same standard population distribution for all items, we model the item random effects with a normal mixture distribution with two components. One component refers to the DIF class and the other component represents the non-DIF class. In the non-DIF class, the item parameters are univariate and therefore the same across the groups of examinees. For the DIF class distribution on the other hand, a G -dimensional distribution (where G equals the number of person groups) is used such that each person group can have its own item parameter. In line with De Boeck (2008), we will call this model the random item mixture (RIM).

While this work was under review, Soares, Goncalves, and Gamerman (2009) published a closely related idea to the RIM, a model which was first formulated by De Boeck (2008). Soares et al. identify very similar concerns with traditional DIF analyses and proposed an integrated Bayesian model to deal with DIF. Although there are many similarities with the model from this paper, there are also a number of differences. The idea of item mixtures is embedded in a more complex Bayesian model; for example, there is even DIF for the guessing parameters and there are random mixing probabilities, so this makes it not always straightforward to see why a particular result is obtained. In line with the complexity of the model, custom-made software for Bayesian estimation is used, while we use publicly available code (which can be shared easily).

It should also be noted that mixture IRT models have been used in the context of DIF detection previously as a useful way to detect latent groups of examinees (Bolt & Cohen, 2005; Bolt, Cohen, & Wollack, 2001; Rost, 1990). For instance, Bolt and Cohen proposed a mixture IRT model to define latent groups of examinees based on their response patterns. In this way, the actual causes of DIF are captured more adequately than when one focuses on manifest person characteristics to form examinee groups. However, the mixture distribution is always applied to classify the persons, but not, as in our case, to classify the items. In this paper, the person groups are always manifest, whereas the item classes are latent.

RIM: The Random Item Mixture DIF Model

The RIM is introduced in this section in a step-by-step fashion. Our starting point is the Rasch model (Fischer & Molenaar, 1995; Rasch, 1960). Let us define a binary random variable Y_{ij} for person i ($i = 1, \dots, I$) responding to item j ($j = 1, \dots, J$).

The random variable Y_{ij} takes the value 1 if person i answers item j correctly and 0 otherwise. The distribution of Y_{ij} is then the Bernoulli:

$$Y_{ij} \sim \text{Bern}(\pi_{ij}) \quad (1)$$

such that

$$\text{logit}(\pi_{ij}) = \text{logit}[\Pr(Y_{ij} = 1)] = \theta_i - \beta_j \quad (2)$$

and θ_i is the ability of person i and β_j the difficulty of item j . Commonly, the persons are considered to be a random sample from a population. Common practice is to assume that the θ_i 's are sampled independently from a normal distribution with mean μ_θ and standard deviation σ_θ : $\theta_i \stackrel{iid}{\sim} N(\mu_\theta, \sigma_\theta^2)$. Most typically, the mean is set to zero to identify the model (otherwise the mean could be considered as a general intercept and this would create a trade-off with the item difficulties). A crossed random effects version of the Rasch model (see, e.g., De Boeck, 2008) is obtained by additionally assuming that the item difficulties are a random sample from a population, most commonly also the normal: $\beta_j \stackrel{iid}{\sim} N(\mu_\beta, \sigma_\beta^2)$.

In a next step, we will assume that there are several groups of examinees. The index g will be used to distinguish between groups ($g = 1, \dots, G$). Possibly, the distributions of the ability differs across the G groups and this is incorporated in the model as follows:

$$\text{logit}[\Pr(Y_{ijg} = 1)] = \theta_{ig} - \beta_j, \quad (3)$$

where θ_{ig} is the latent ability of person i in group g and with $\theta_{ig} \sim N(\mu_{\theta_g}, \sigma_{\theta_g}^2)$. In this model, the ability follows a normal distribution in all G groups, but with a group-specific mean and variance. To identify the model in Equation 3, a constraint is needed because we can always add a constant to the β_j 's (or to μ_β in the random item case) and subtract it from μ_{θ_g} without changing the likelihood of the model. A natural constraint is to set μ_{θ_1} equal to 0. This means that the mean latent ability for groups 2 to G can be seen as deviations from the mean ability of the first group.

In the mixture DIF model there are two subgroups or latent classes of items: one class contains the DIF items and the other class the non-DIF items. To represent the class membership of an item, we introduce a latent indicator C_j defined as follows:

$$C_j = \begin{cases} 0 & \text{if item } j \text{ shows no DIF} \\ 1 & \text{if item } j \text{ shows DIF.} \end{cases}$$

It is assumed that the distribution of this latent indicator C_j is Bernoulli with probability π_{DIF} : $C_j \stackrel{iid}{\sim} \text{Bern}(\pi_{\text{DIF}})$. In this model, we should look at the conditional probabilities of responding correctly to item j in the two latent classes and the conditional item difficulty distribution.

For an item j of the non-DIF class, the model for the conditional probability of responding correctly to the item can be written as follows:

$$\text{logit}[\Pr(Y_{ijg} = 1 | C_j = 0)] = \theta_{ig} - \beta_j \quad (4)$$

with the item difficulty sampled randomly from a univariate normal distribution:

$$\beta_j | C_j = 0 \sim N(\mu_\beta, \sigma_\beta^2). \quad (5)$$

For an item j of the DIF class, the model becomes:

$$\text{logit}[\Pr(Y_{ijg} = 1 | C_j = 1)] = \theta_{ig} - \beta_{jg} \quad (6)$$

with the item difficulty sampled randomly from a G -variate normal distribution:

$$\begin{bmatrix} \beta_{j1} \\ \vdots \\ \beta_{jG} \end{bmatrix} \bigg| C_j = 1 \sim N \left[\begin{pmatrix} \mu_{\beta_1} \\ \vdots \\ \mu_{\beta_G} \end{pmatrix}, \Sigma_\beta \right] \quad (7)$$

with covariance matrix Σ_β of the item difficulties across the G groups equal to

$$\Sigma_\beta = \begin{pmatrix} \sigma_{\beta_1}^2 & \cdots & \sigma_{\beta_1\beta_G} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta_G\beta_1} & \cdots & \sigma_{\beta_G}^2 \end{pmatrix}. \quad (8)$$

For both latent classes, it is assumed that $\theta_{ig} \sim N(\mu_{\theta_g}, \sigma_{\theta_g}^2)$. Equations 4 and 6 indicate that when items are classified as belonging to the non-DIF class the item difficulty, β_j , is the same for every group of examinees. However, if an item is classified as belonging to the DIF class, the item has a different difficulty level for each group (i.e., β_{jg} for group g).

To identify this model it is necessary to set some constraints. First of all, we constrain μ_{θ_1} to be zero as was also the case in the previous models. Second, we constrain the mean item difficulty of the non-DIF items (μ_β) to be equal to the mean item difficulty of the DIF items for all G groups (i.e., $\mu_\beta = \mu_{\beta_1} = \cdots = \mu_{\beta_G}$). We put the constraints on the mean item difficulties because we consider the items to be random.

For the most common case of two groups ($G = 2$), the model then becomes:

$$\begin{aligned} \text{logit}[\Pr(Y_{ijg} = 1 | C_j = 0)] &= \theta_{ig} - \beta_j \\ \text{logit}[\Pr(Y_{ijg} = 1 | C_j = 1)] &= \theta_{ig} - \beta_{jg} \end{aligned} \quad (9)$$

with the following set of distribution assumptions:

$$\begin{aligned}\theta_{i1} &\sim N(0, \sigma_{\theta_1}^2) \\ \theta_{i2} &\sim N(\mu_{\theta_2}, \sigma_{\theta_2}^2) \\ \beta_j | C_j = 0 &\sim N(\mu_\beta, \sigma_\beta^2)\end{aligned}\tag{10}$$

and

$$\begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} \Big| C_j = 1 \sim N \left[\begin{pmatrix} \mu_\beta \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_1\beta_2} \\ \sigma_{\beta_1\beta_2} & \sigma_{\beta_2}^2 \end{pmatrix} \right].\tag{11}$$

The covariance parameter $\sigma_{\beta_1\beta_2}$ (i.e., the covariance of the item difficulties of the DIF items between the two person groups) will not be estimated directly. Instead, the model was reparametrized such that the correlation ρ enters as a parameter in the model (the off-diagonal elements of Σ_β are then equal to $\rho\sigma_{\beta_1}\sigma_{\beta_2}$).

In what follows, we will consider further only the common case with two groups of examinees (e.g., men and women). So when an item is said to exhibit DIF, there are two different item difficulties, namely one for the first group of examinees and one for the second group of examinees. When an item is free of DIF the two groups share the same item difficulty.

Our goal is to classify items, and therefore we will use the conditional or posterior probability $\Pr(C_j = 1 | \mathbf{y})$ (with $\mathbf{y}$ being the vector of all observed data points). Because the conditional probability can lie anywhere in the unit interval, we have to use a discretized rule for allocating an item to the DIF or non-DIF class. The rule we propose is the following: If an item j has a conditional or posterior probability of belonging to the DIF class exceeding .5, the item is classified as a DIF item in the RIM. More complex allocation rules can be used (e.g., with an additional intermediate zone of indifference), but they will not be discussed further in this paper.

Statistical Inference

Let us first consider the crossed random effects Rasch model in Equation 2. The fixed parameters of the model (i.e., σ_θ , μ_β , and σ_β) can be estimated by maximizing the following marginal likelihood:

$$\begin{aligned}L(\sigma_\theta, \mu_\beta, \sigma_\beta | \mathbf{y}) &= \int_{\theta_1} \cdots \int_{\theta_I} \int_{\beta_1} \cdots \int_{\beta_J} \prod_{i=1}^I \prod_{j=1}^J \Pr(Y_{ij} = y_{ij}) \\ &\quad \times \prod_{i=1}^I \phi(\theta_i; 0, \sigma_\theta^2) \prod_{j=1}^J \phi(\beta_j; \mu_\beta, \sigma_\beta^2) d\theta_1 \cdots d\theta_I d\beta_1 \cdots d\beta_J,\end{aligned}\tag{12}$$

where $\mathbf{y}$ is again the vector of all observed data points and $\phi(x; \mu, \sigma)$ is the normal density function evaluated at x with mean μ and variance σ^2 . It can be seen that the integral to marginalize over the random effects is $(I + J)$ -dimensional. Because the

integral is intractable and because it is high-dimensional, the traditional numerical approach (e.g., using Gauss-Hermite quadrature or some other numerical integration method)¹ becomes prohibitive (moreover, adding a person or an item increases the dimensionality of the integral.). For this reason, we will rely on Bayesian methods for statistical inference.

The possibility of fitting high-dimensional models like the one outlined in this paper is one of the major assets of Bayesian analysis (Gelman, Carlin, Stern, & Rubin, 2004). The powerful computational methods applied in Bayesian analysis are known as Markov chain Monte Carlo (MCMC) procedures, with most notably the Gibbs sampler and the Metropolis-Hastings sampler (Gelman et al., 2004; Tanner, 1996). The development of user-friendly software like WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) has further fueled the popularity of Bayesian methods. In this paper, we make use of WinBUGS as a tool for Bayesian statistical inference.

In a Bayesian approach, one considers all parameters of the model to be random variables (note that this conceptually matches nicely with our assumption of random effects for both items and persons). Let the vector ξ contain all parameters of the model: not only the fixed effects parameters but also the random effects are part of ξ . Moreover, it is beneficial to consider the latent indicators C_j as latent or missing data and augment the observed data with these latent data (see Tanner, 1996). To simplify matters, we will put both the parameters and the latent data in the vector ξ . The posterior is then proportional to the product of the likelihood and priors:

$$p(\xi | y) \propto L(\xi | y)p(\xi), \quad (13)$$

where the choice of prior distributions $p(\xi)$ will be discussed later. Because of the fact that the random effects and latent indicators are considered to be parameters of the model, the likelihood is defined as follows:

$$\begin{aligned} L(\xi | y) = & \prod_{i=1}^I \prod_{j=1}^J \prod_{c_j=0}^1 \Pr(Y_{gij} = y_{gij} | C_j = c_j) \prod_{j=1}^J \left[(1 - \pi_{\text{DIF}})^{1-c_j} \pi_{\text{DIF}}^{c_j} \right] \\ & \times \prod_{i=1}^I \phi(\theta_{ig}; \mu_{\theta_g}, \sigma_{\theta_g}^2) \prod_{j=1}^J \left[\phi(\beta_j; \mu_{\beta}, \sigma_{\beta}^2)^{1-c_j} \phi(\beta_j; \mu_{\beta}, \Sigma_{\beta})^{c_j} \right]. \end{aligned} \quad (14)$$

With respect to the choice of the prior distributions there are both theoretical and practical considerations we have taken into account, namely vagueness and practical implementation. Initially we wanted to have priors as vague as possible (vague priors are typically used for parameters about which not much is known beforehand). However, this did not always work out very well in practice for the means of the distributions, because it resulted in technical WinBUGS problems. For instance, if we used the following vague priors: $\mu_{\theta_2} \sim N(0, 1000)$ and $\mu_{\beta} \sim N(0, 1000)$, WinBUGS was not able to sample these parameters and hence the program did not work. To avoid such problems, we made the priors for the means step by step less vague, until there were no more technical WinBUGS problems (i.e., first we tried to run the

model using $N(0, 1000)$ as a prior, then because this did not work we considered $N(0, 100)$ as a prior, and so on).

Also, the following standard normal priors for the mean parameters were chosen:

$$\begin{aligned}\mu_{\theta_2} &\sim N(0, 1) \\ \mu_{\beta} &\sim N(0, 1).\end{aligned}\tag{15}$$

Because the standard normal is a quite informative prior, in a simulation study we examined whether or not this has an effect on the results. In addition, we used a Beta prior for $\pi_{\text{DIF}} : \pi_{\text{DIF}} \sim \text{Beta}(1, 1)$.

For the standard deviation parameters σ_{θ} , σ_{β} , and σ_{β_2} it is not clear which priors are the best ones. There is not yet a standard solution to the problem of choosing a prior for the standard deviation components in random-effects models. Various prior choices for modeling standard deviations have been suggested in Bayesian literature and software, including uniform, inverse gamma and truncated normal priors (see, e.g., Gelman et al., 2004). Gelman and Hill (2007) recommend starting with a noninformative uniform prior density for the standard deviations. They do not recommend the use of the inverse-gamma(ϵ, ϵ) family because, in cases when the variance is estimated to be near zero, the resulting inferences will depend strongly on ϵ . From preliminary analyses (not reported here), it turns out that there is not much difference between these three different priors and therefore we opted for a uniform prior on the standard deviation between 0 and 3. Once again, we balanced between vagueness on the one hand and practical WinBUGS constraints on the other hand when constructing this prior by restricting the range step by step until WinBUGS stopped giving problems (first, we considered the interval $[0, 1000]$, because this did not work we tried the $[0, 100]$ interval and so on). As a prior for ρ , we choose a uniform distribution between -1 and 1 .

Another constraint we imposed is that the variance of the item difficulties of the non-DIF items is restricted to be equal to the variance of the DIF items of the first group ($\sigma_{\beta}^2 = \sigma_{\beta_1}^2$) and also the ability variance is restricted to be equal across the two groups ($\sigma_{\theta_1}^2 = \sigma_{\theta_2}^2$). These two restrictions are not strictly necessary but we imposed them for the sake of simplicity.

To explore the posterior distribution, we make use of the MCMC procedures provided by WinBUGS. In all further applications in this paper five chains of 8,000 iterations were run, starting from randomly generated starting points, and 4,000 iterations were discarded as burn-in. For the remaining iterations, convergence of the Markov chain was assessed using the $\hat{R}$ criterion of Brooks and Gelman (1998). The basic WinBUGS code of the RIM is given in the Appendix (the code is only for two groups of examinees, but it can easily be generalized to deal with more than two groups of examinees).

Simulation Studies

Implementation and Decision Rules

In all three simulation studies, the RIM is estimated using the WinBUGS software. A critical issue in applying MCMC techniques is whether the Markov chain(s)

have converged to the true posterior distribution. As stated previously, we use the convergence measure $\hat{R}$, which is approximately the square root of the ratio of the between-chain variance to the within-chain variance (Brooks & Gelman, 1998; Gelman & Rubin, 1992). Gelman and Hill (2007) advise to wait until $\hat{R} \leq 1.1$ for all parameters, although, if sampling from the posterior is proceeding slowly, one may also work with chains that have still not completely mixed, for example with $\hat{R} \approx 1.5$ for some parameters. If we calculate the $\hat{R}$ for every estimated parameter across all 320 data sets in the first simulation study we see that $\hat{R}$ is in approximately 94% of all cases below 1.1. This indicates that, in general, all chains seem to converge very well. In the cases that $\hat{R}$ exceeded 1.1, $\hat{R}$ was always smaller than 1.5, so we can conclude that there are also no parameters with an extremely poor convergence.

The methods used to decide on the DIF status of an item also deserve some attention. First, in the section on the RIM, it has already been explained that we base our decision on the posterior probability $\Pr(C_j = 1 | y)$. An estimate of this posterior probability is relatively easy to obtain from the MCMC output. For item j , it is simply the posterior mean of C_j (and the latent indicators are sampled at each iteration). If the posterior probability of belonging to the DIF class for item j exceeds .5 (meaning that in more than half of the iterations the item was allocated to the DIF class), the item is classified as a DIF item in the RIM. In the first simulation study, the RIM is compared to the MH and LRT procedures. In both methods, the test statistic is chi-square distributed with one degree of freedom. For both methods, we use .05 as the p -value threshold to decide whether an item exhibits DIF or not. Traditionally, ST p -DIF values outside the $-.10$ and $.10$ interval are considered to be unusual and should be examined very carefully (Holland & Wainer, 1993) and we will adhere to this rule of thumb.

For the LRT, the MH, and the ST p -DIF procedure we used all items as anchor items. In the case of the MH and the ST p -DIF procedure, we used the total test score as a proxy for the ability of an examinee. In the case of the LRT we constrained all the test item difficulties to be equal in the compact model. In the extended model, these equality constraints for the studied item are dropped.

Simulation Study 1

Design. In our first simulation study, our objective is to compare the performance of the RIM with more traditional methods like the LRT, the MH, and the ST p -DIF procedure. We have a fully crossed and balanced four-factorial design with the following factors and levels:

1. the number of examinees ($I = 500$ and $I = 1,000$);
2. the number of items ($J = 20$ and $J = 50$);
3. the number of items that contain DIF ($D = 0$ and $D = 5$); and
4. the difference in latent ability between the two groups ($\mu_{\theta_2} = 0$ and $\mu_{\theta_2} = .5$).

Each group consists of 50% of the examinees (e.g., if there are 500 examinees, each group consists of 250 examinees). The person abilities of the reference group and item difficulties are always drawn from a standard normal distribution. In cases where there is a difference in latent ability between the two groups, the person

abilities from the focal group will be drawn from a normal distribution with mean .5 and standard deviation 1.

When items were designated to exhibit DIF, the size of the DIF was manipulated systematically as well. In case of DIF, the item difficulties differ between the two groups in the following possible amounts: .4, .6, .8, $-.8$, -1 . Hence, there is always one item with the difference in item difficulty for the two groups being .4, one item with a difference of .6, and so on. Note that this systematic manipulation of the DIF size is actually not in correspondence with the underlying theoretical model, because in the model it is assumed that for the DIF items the item difficulties for both groups are randomly drawn from a bivariate normal distribution.

In each cell of the design, there are 20 replications, yielding 2 (number of examinees) $\times 2$ (DIF items) $\times 2$ (latent ability) $\times 2$ (number of items) $\times 20$ (replicates) = 320 simulated data sets. Note in this respect that the MCMC procedure is rather time consuming (e.g., when there are 500 examinees and 20 items, it takes approximately 2 hours to estimate all parameters when using five chains that each consist of 8,000 iterations on a computer with an Intel Core2 Duo 3.16 GHz CPU).

Results. Table 1 gives an overview of the number of misclassifications, the Type 1 error, and the power of the RIM, the LRT, the MH and the ST p -DIF procedure. It should be noted that there are two possible misclassifications : a DIF-free item is wrongly allocated to the DIF class (a Type I error), and a DIF item is erroneously assumed to be DIF-free (a Type II error). We will address this distinction later.

An analysis of variance was carried out as an approximate method to determine the proportion of variance explained (η^2) by each of the factors (number of examinees, number of items, number of DIF items and difference in mean ability level) on the average number of errors of the RIM. This shows that the factor “number of DIF items” explains 52% of the variance in the errors made (with more mistakes being made when there are DIF items as opposed to when there are none), the factor “number of items” 10% (with more mistakes being made when there are 20 items as opposed to when there are 50 items) and the number of examinees 11% (with more mistakes being made when there are only 500 examinees as opposed to 1,000). The effect of “difference in latent ability” was not significant.

In a next step, we looked at the Type I errors and the power of each method. It can be seen in Table 1 that there is a difference between the LRT, the MH procedure, the ST p -DIF procedure, and the RIM in this respect. The RIM and ST p -DIF methods seem to be rather conservative; the Type I error is rather small ($<5\%$). Both the LRT and the MH on the other hand have a Type I error around 5%. Also, the MH and LRT seem to have more power than the RIM and ST p -DIF method.

From Table 1, one may conclude therefore that the LRT and the MH procedures tend to flag more items than the mixture DIF model and the ST p -DIF method. Moreover, a similar trend (smaller Type I errors, smaller power) for both the ST p -DIF procedure and the RIM suggests that the flagging rule is possibly too strict. For the ST p -DIF procedure, the thresholds were fixed to $-.10$ and $.10$, that is, items with a ST p -DIF statistic larger (in absolute value) than $.10$ were flagged as DIF. Selecting smaller thresholds, for instance $-.05$ and $.05$, would reduce the conservativeness of the procedure. Indeed, more items would be flagged as DIF,

Table 1

The Type I Error, the Power, and the Average Misclassification Rate (a.m.r.) of the RIM, the LRT, the MH, and the ST *p*-DIF Procedure

<i>I</i>	<i>J</i>	<i>D</i>	μ_2	RIM			LRT			MH			ST- <i>p</i> -DIF		
				α	1 - β	a.m.r.	α	1 - β	a.m.r.	α	1 - β	a.m.r.	α	1 - β	a.m.r.
500	20	0	0	0	-	0	4.5	-	4.5	3.3	-	3.3	.8	-	7.5
500	50	0	0	0	-	0	4.3	-	4.3	3.7	-	3.7	1.6	-	1.6
500	20	5	0	2.7	68	10	5	81	8.5	4	80	8	1.7	61	11
500	50	5	0	.6	59	4.6	4.9	81	6.3	3.6	80	5.2	1.5	70	4.3
500	20	0	.5	.3	-	.3	3.8	-	3.8	3.8	-	3.8	1.3	-	1.3
500	50	0	.5	0	-	0	4.8	-	4.8	4.6	-	4.6	2.1	-	2.1
500	20	5	.5	4.3	60	13.3	5.3	85	7.8	4.3	75	9.5	1.7	61	11
500	50	5	.5	.1	51	5	5.7	88	6.3	5.2	80	6.7	2.6	67	5.6
1000	20	0	0	.3	-	.3	7	-	7	2.8	-	2.8	0	-	0
1000	50	0	0	0	-	0	5.4	-	5.4	4.8	-	4.8	0	-	0
1000	20	5	0	1.3	85	4.7	6.7	90	7.5	4.7	94	5	0	66	8.5
1000	50	5	0	0	83	1.7	3.9	93	4.2	3.9	94	4.1	.1	71	3
1000	20	0	.5	.3	-	.3	3.3	-	3.3	3	-	3	0	-	0
1000	50	0	.5	.2	-	.2	4.6	-	4.6	4.6	-	4.6	0	-	0
1000	20	5	.5	1	87	4	5	96	4.8	9.6	96	3.8	.3	70	7.8
1000	50	5	.5	.3	85	18	4.7	97	4.5	4.9	93	5.1	0	66	3.4

Note. *I* refers to the number of examinees, *J* to the number of items and *D* to the number of DIF items. The errors are calculated as the average percentage of misclassifications over the 20 replications. For example, 4.5% in the first row is obtained by averaging the percentage of misclassifications across the 20 data sets with the following characteristics: 500 examinees, 20 items, 0 DIF items and no difference in mean latent ability between the groups. For each replication the percentage of misclassifications was calculated (e.g., if there is 1 item out of 20 misclassified, the percentage of misclassifications is 1/20 = 5%). The mean of these percentages are presented in the table. α refers to the Type I error (i.e., the percentage of cases in which an item that does not contain DIF is wrongly classified as a DIF item) and 1 - β to the power (i.e., the percentage of cases in which an item that contains DIF is correctly flagged as a DIF item); a dash (-) indicates that the power cannot be computed.

Table 2

Recovery of the Parameters (the Posterior Mean and Standard Deviation of the Estimates)

Parameter	True Value	Posterior Mean	Posterior Standard Deviation
μ_2	0	.006	.087
μ_2	.5	.511	.082
μ_β	0	.004	.201
ρ	0	.164	.179
σ^2	1	1.002	.037
σ_β	1	1.051	.144
σ_{β_2}	1	1.080	.363
DIF size	0	-.001	.004
DIF size	.4	.27	.22
DIF size	.6	.51	.25
DIF size	.8	.76	.21
DIF size	-.8	-.74	.23
DIF size	-1	-1.02	.18

increasing both the Type I error and the power. The choice of accurate thresholds for the ST p -DIF procedure is unfortunately not clearly stated in the literature. For the RIM, items with a posterior DIF probability larger than .5 are flagged as DIF. Here also, reducing this threshold would increase both the Type I error and the power.

In Table 2, we assess the recovery of some parameters. To do so, we made use of the estimated posterior means of the parameters of interest across all data sets. To summarize the recovery, we calculated both the mean and the standard deviation of these posterior means (note that no parameter has to be equated back to another scale of measurement, because the RIM is an integrative item response model for all data such that all item difficulties are on the same scale). Table 2 shows us that most parameters are recovered rather well. For the lower DIF sizes, the estimates seem rather low, but notice that the 95% credibility intervals do contain the true values.

Simulation Study 2

Design. In a second simulation study, we investigated the sensitivity of the results to the constraint that the mean of the non-DIF item difficulties equals the mean of the DIF item difficulties (i.e., $\mu_\beta = \mu_{\beta_1} = \mu_{\beta_2}$). For this reason, we manipulated systematically the difference between the means. The mean difficulty of the non-DIF items (i.e., μ_β) and the DIF-items of the first group of examinees (i.e., μ_{β_1}) is always equal to zero but we generated the mean difficulty of the DIF items of the second group of examinees (i.e., μ_{β_2}) with three different values: .4, .8, and 1. The size of the DIF was manipulated systematically: in the case when the difference in mean difficulty was .4, we added .4 for all DIF items (so the difference in item difficulty between the two groups was exactly .4 for all DIF items) and so on. Note that this is also a sensitivity test of the influence of the standard normal prior on μ_β . In addition, in this second simulation study, we varied the number of examinees ($I = 500$ and $I = 1,000$), the number of items ($J = 20$ and $J = 50$), and the difference in latent

ability ($\mu_{\theta_2} = 0$ and $\mu_{\theta_2} = .5$). The number of DIF items remained constant ($D = 5$). All factors are fully crossed and for each cell there were 20 replications.

It is important to notice that the results of simulation studies 1 and 2 cannot be compared directly, for at least two reasons. First, in simulation study 1 the average difference in item difficulty is set to zero by selecting appropriate DIF effects, some being positive, some being negative, but summing to zero. Study 1 focuses therefore on symmetric DIF, some DIF items in favor of the reference group, others in favor of the focal group. In simulation study 2, the DIF effect is asymmetric because all DIF items are more difficult for the focal group, their difficulty levels are increased by a constant value (either .4, .8, or 1). Notice that this simulation study is thus also a test on whether our model performs well in case of asymmetric DIF. A second reason is that the size of DIF is different for each DIF item in the first study: the difference in item difficulties range from .4 to 1 (in absolute values) with an average value of .72. In the second study, all DIF items have the same size of DIF.

Results. The results of the second simulation study can be found in Table 3. In Table 3, the Type I error, the power, and the average misclassification rate of the RIM are tabulated as a function of the difference in mean difficulty between the DIF and non-DIF items (and the other factors).

The results indicate that a violation of the restrictions does not result in an increase in errors. In contrast, an increase in item difficulty differences yields an important decrease in average misclassification rates. This is mainly due to the important increase of power, that is, the DIF items are more often identified as the differences in mean item difficulty become very large. With differences of one unit, the power is usually equal to or larger than .9. In parallel, the Type I errors increase slightly (except in a few situations leading to a nonmonotonic trend). The Type I errors, however, remain smaller than the nominal significance level of 5% in almost all situations. The large misclassification rates, observed when the difference in mean difficulty equals .4, are due to an important lack of power. This is because the size of DIF is rather small and many DIF items are not identified as such.

Recall that these results cannot be directly compared to those from Table 1, because (a) the size of DIF is constant across the DIF items in this study, while it changes in the first simulation study; (b) the DIF effect is asymmetric in the present simulation, while it was symmetric in the previous one. It is nevertheless interesting to notice that the best correspondence between Tables 1 and 3 is obtained when the DIF size equals .8. This DIF size is similar to the average absolute DIF effect of .72 in the first simulation study.

In sum, the model restrictions do not interfere with the correct identification of DIF. With large differences in mean difficulty levels, the model performs even better than with smaller DIF effects. Notice that our model also performs well in case of asymmetric DIF. Hence, it is concluded that our model is robust against this kind of violation of the model assumptions.

Simulation Study 3

Design. In a third simulation study, we investigated how sensitive the conclusions are with respect to the distributional assumptions regarding the item populations (i.e.,

Table 3
The Type I Error, the Power, and the Average Misclassification Rate (a.m.r.) of the RIM with the Difference in Mean Difficulty of the DIF Items and the Non-DIF Items Manipulated

<i>I</i>	<i>J</i>	<i>D</i>	μ_2	Difference in Mean Difficulty	RIM		
					α	$1 - \beta$	<i>a.m.r.</i>
500	20	5	0	.4	0.3	8	23.3
500	20	5	0	.8	4	75	9.3
500	20	5	0	1	5	93	5.5
500	50	5	0	.4	0	7	9.3
500	50	5	0	.8	0.6	75	3
500	50	5	0	1	1.1	87	2.3
500	20	5	.5	.4	0	5	23.8
500	20	5	.5	.8	3.7	77	8.5
500	20	5	.5	1	7	95	6.5
500	50	5	.5	.4	0.1	7	9.4
500	50	5	.5	.8	0.3	82	2.1
500	50	5	.5	1	0.7	89	1.7
1000	20	5	0	.4	1	22	2.3
1000	20	5	0	.8	2	98	2
1000	20	5	0	1	1.7	99	1.5
1000	50	5	0	.4	0.1	19	8.2
1000	50	5	0	.8	0	99	0.1
1000	50	5	0	1	0.6	99	0.6
1000	20	5	.5	.4	1	28	18.8
1000	20	5	.5	.8	0.7	97	1.3
1000	20	5	.5	1	3.7	99	3
1000	50	5	.5	.4	0.1	16	8.5
1000	50	5	.5	.8	0.2	92	1
1000	50	5	.5	1	0.3	100	0.3

Note. *I* refers to the number of examinees, *J* to the number of items and *D* to the number of DIF items. α refers to the Type I error (i.e., the percentage of cases in which an item that does not contain DIF is wrongly classified as a DIF item) and $1 - \beta$ to the power (i.e., the percentage of cases in which an item that contains DIF is correctly flagged as a DIF item).

normal distributions). Therefore, a small simulation study was carried out in which the item difficulties were drawn from of a uniform distribution. Two additional factors were manipulated in this study: the number of examinees at two levels (*I* = 500 and *I* = 1, 000) and the number of DIF items at two levels (*D* = 0, *D* = 5). The total number of items remained constant at 20 and the difference in mean latent ability between the two groups was 0. For each cell we used 20 replications.

Results. Table 4 gives the results of the performance of the RIM. Again it can be concluded that the RIM does a good job at identifying DIF items when the item difficulties are drawn out of a uniform distribution because the number of errors made is limited. Comparing these results with the ones obtained when the item difficulties were drawn from a normal distribution (Table 1) shows a close similarity : When the

Table 4
The Type I Error, the Power, and the Average Misclassification Rate (a.m.r.) of the RIM Using a Uniform Distribution to Generate the Item Difficulties

<i>I</i>	<i>J</i>	<i>D</i>	μ_2	RIM		
				α	$1 - \beta$	a.m.r.
500	20	0	0	0	–	0
500	20	5	0	2.7	67	10
1000	20	0	0	0	–	0
1000	20	5	0	1	86	4.3

Note. *I* refers to the number of examinees, *J* to the number of items and *D* to the number of DIF items. α refers to the Type I error (i.e., the percentage of cases in which an item that does not contain DIF is wrongly classified as a DIF item) and $1 - \beta$ to the power (i.e., the percentage of cases in which an item that contains DIF is correctly flagged as a DIF item); a dash (–) indicates that the power cannot be computed.

item difficulties were drawn from a normal distribution, the two cases with no DIF items also had a very low Type I error (i.e., 0 and .3). Also in the case when there were DIF items, the results are consistent (highly similar Type I error and power).

Application

As an application of the method we analyzed data concerning required attainment targets for high school students (of different grades) set by the Ministry of Education of the Flemish government (regional government of the northern part of Belgium). Such attainment targets are the minimum goals and competency levels students have to reach at the end of a certain educational level. In this study, the attainment targets under study were about information acquiring and processing in graphical material (plans, maps, and drawings). A test of 36 items was constructed and administered and we will use the responses of the 1,905 examinees to these items to illustrate the detection of DIF.

Previous research has indicated that men and women differ with regard to their spatial skills (Geary & DeSoto, 2001; Voyer, Voyer, & Bryden, 1995). Men generally outperform women on tests of spatial abilities. The most robust gender difference in this area is found for mental rotation (Voyer et al., 1995). Since men seem to have better spatial skills than women, we expect there to be a difference in the abilities of men and women in the domain of information acquiring and processing in graphical material. To account for this difference, we consider men and women as the two groups in our study. The main goal of this study is to examine if there are items that function differently for men and women.

To estimate the RIM, five chains were run for 8,000 iterations starting from different starting points after a burn-in period of 4,000 iterations. Because $\hat{R} \leq 1.1$ for all parameters, it may be concluded that all chains converged. Figure 1 shows the sample chains for μ_{θ_2} , μ_{β} , σ_{θ}^2 , and σ_{β}^2 . Here we can see that all five chains converged rapidly for these parameters (although there is a difference between the parameters with regard to number of iterations necessary before there is convergence).

Figure 1. Sample chains for μ_{θ_2} , μ_{β} , σ_{θ}^2 , and σ_{β}^2 . For all five chains, the first 2,000 out of 8,000 iterations are shown.

Nine items were classified as containing DIF by the RIM. The difference in mean latent ability between men and women is .33 (in favor of men) with a standard error of .05. The 95% credibility interval for μ_{θ_2} (i.e., the difference in mean latent ability between men and women) ranges from .22 to .43. This indicates, in line with previous research, that men are on average more capable of solving items related to consulting maps, drawings, and plans than women. In Figure 2, the posterior distributions of μ_{θ_2} , μ_{β} , σ_{θ}^2 , and σ_{β}^2 are shown.

We also compared the RIM with the LRT, the MH, and the ST p -DIF procedure; 18 items were identified as containing DIF by the LRT, 14 by the MH procedure, and three by the ST p -DIF.² The fact that the ST p -DIF procedure only identified three items as containing DIF is consistent with the earlier finding that the ST p -DIF procedure has lower power to detect DIF items (Table 1). The LRT and the MH procedures, on the other hand, flagged more items as DIF than the RIM and the ST p -DIF procedure, which was also expected from the simulation results of Table 1.

The different methods all give more or less the same classification with regard to the items (Table 5). Of course, the number of items classified as DIF differs among the methods, but the judgments with respect to the items are consistent. For example, the ST p -DIF method classified the fewest as containing DIF, but these items were

Figure 2. Posterior distribution of μ_{θ_2} , μ_{β} , σ_{θ}^2 , and σ_{β}^2 .

also categorized as DIF items by the other methods, showing that the methods do not randomly point to items as containing DIF, but point to the same items. More specifically, the RIM has made the same classification (containing DIF or not) for 27 items as the LRT. For the ST p -DIF and the MH procedure, there is a higher correspondence with the RIM, respectively 30 and 29 items. So, although the methods differ with respect to the number of items they classify as DIF, there is a correspondence with regard to the specific items they classify as DIF. As stated earlier, the three items that were classified as DIF items by the ST p -DIF procedure were also classified as DIF items by all other methods. The RIM classified six items more than the ST p -DIF procedure as DIF and again these items were also classified as DIF by the MH procedure and the LRT. From the five extra items the MH procedure picked out as DIF items, the LRT also classified four as DIF items. Hence, we can conclude that the methods differ mainly in severity but that they do point to the same items as containing DIF. To sum up, three items were classified as DIF by all four methods, five items were categorized as DIF items by three methods, six items were classified as DIF by two methods, and five items were diagnosed as DIF items by only one method. Note in this respect that we are dealing with real data and so the true pattern of DIF and non-DIF items is unknown. Therefore, we cannot simply state which method yields the most correct classification of the items.

Table 5

Classification of the Detection Status for the Items According to the Different Methods ("n" Refers to a Non-DIF Item, "y" to a DIF Item)

Item	RIM	ST <i>p</i> -DIF Procedure	LRT	MH Procedure
1	n	n	n	n
2	n	n	n	n
3	n	n	n	n
4	n	n	n	n
5	n	n	n	n
6	n	n	n	n
7	n	n	n	n
8	n	n	n	n
9	n	n	n	n
10	n	n	n	n
11	n	n	n	n
12	n	n	n	n
13	n	n	n	n
14	n	n	n	n
15	n	n	n	n
16	n	n	n	n
17	n	n	n	n
18	y	y	y	y
19	y	y	y	y
20	y	y	y	y
21	y	n	y	y
22	y	n	y	y
23	y	n	y	y
24	y	n	y	y
25	y	n	y	n
26	y	n	y	y
27	n	n	y	y
28	n	n	y	y
29	n	n	y	y
30	n	n	y	y
31	n	n	y	y
32	n	n	y	n
33	n	n	y	n
34	n	n	y	n
35	n	n	y	n
36	n	n	n	y

An important feature of the test under study is that the item responses are not independent of each other. Instead there are groups of items based on a common stimulus (e.g., a picture). Such items groups are called testlets (Wainer & Kiely, 1987) and they are primarily used because they are time-efficient. The basic idea is that reading and comprehending a stimulus requires time. But combining different items with just one stimulus (instead of having a separate stimulus for every item)

decreases the total amount of test time. It has been shown that when this dependence structure between items is ignored in an IRT context, this results in an overstatement of the precision of the underlying abilities of the examinees as well as a bias in item difficulty estimates (Tuerlinckx & De Boeck, 2001; Wainer & Thissen, 1996). Hence, it is important to extend the RIM in such a way that it can account for these item dependences. Being able to include testlets in our model (and so having a flexible model that can easily be adjusted) is a huge advantage of our approach. One way to do this is to follow the proposal of Bradlow, Wainer, and Wang (1999) to include an additional random effect for items nested within the same testlet. We extended our RIM with a random effect for items nested within the same testlet, hence our model can be written as follows:

$$\begin{aligned}\text{logit} [\Pr(Y_{ijg} = 1 \mid C_j = 0)] &= \theta_{ig} - \beta_j - \gamma_{id(j)} \\ \text{logit} [\Pr(Y_{ijg} = 1 \mid C_j = 1)] &= \theta_{ig} - \beta_{jg} - \gamma_{id(j)}\end{aligned}\quad (16)$$

with the following priors:

$$\gamma_{id(j)} \sim N(0, \sigma_\gamma^2) \text{ and } \sigma_\gamma^2 \sim \text{Unif}(0, 5). \quad (17)$$

Note that the effect for items nested within the same testlet can vary over persons (as can be seen from the i subscript).

We also used the testlet RIM as a way to identify DIF items. To estimate this model, five chains were run for 8,000 iterations starting from different starting points after a burn-in period of 4,000 iterations. There are only two items for which the RIM gives a different classification, namely items 2 and 5. These items are classified as non-DIF items by the RIM and as DIF items by the testlet RIM. Extending the RIM with a person specific testlet effect thus changes the item classification to a small extent. This would seem to indicate that it is important to take the dependence structure between the items into account. Other aspects of the inference also remain largely unaffected. We also looked at the 95% credibility interval for μ_{θ_2} : this ranges from .22 to .40. This credibility interval largely overlaps with the one found when estimating the RIM. Hence, when one takes the testlet structure into account, men seem to outperform women on tasks that involve spatial ability.

Discussion, Limitations, and Further Research

The goal of this paper is to introduce a new method to detect DIF items. A unique feature of the RIM presented in this paper is that it does not require that one specify a set of anchor items beforehand to classify items (instead the model automatically selects anchor items). Also, both the abilities of the examinees and the item difficulties are considered to be random. Defining the difficulty of the items as random is an efficient and realistic way to model the variation in item difficulty. One advantage of the model is that it provides an automatic, one-step classification of DIF and non-DIF items. This is in contrast with previous methods which are based primarily on a stepwise strategy to detect DIF. A second advantage is that the method does not require the prespecification of a set of anchor items. This has the advantage that

possible problems associated with identifying the wrong items as anchor items are bypassed. In particular, it is not necessary to perform any kind of item purification with this model (although it could still be done if one wanted to). A third advantage is that it is possible to change certain details of the model to account for specific aspects of the design that may violate certain assumptions. We showed this for instance with the testlet design. The fact that the RIM with the testlet extension gives more or less the same results as the normal RIM shows the potential of the model. Also, the model gives a clear-cut classification of the items as containing DIF or not, even when there is a violation of the model's assumptions. Moreover, the RIM seems to perform better than the MH and the ST p -DIF procedures and slightly better than the LRT. Note that the RIM has a smaller Type 1 error rate than the other methods, but also appears to be less powerful. We suggest additional research to test this further in a larger more extensive simulation study.

The major drawback of the RIM is that it cannot be run currently with software other than WinBUGS. This obviously limits the range of application of the model, because one needs to be familiar with Bayesian statistics (including the specification of priors and the Monte Carlo computation techniques) and the inference process can be rather time-consuming. Further research should be aimed at developing alternative methods for fitting the RIM in an efficient way (this could be done for instance using an EM algorithm).

There are also some possible extensions of the current model. First of all, it is possible to introduce more than two latent item classes. Here we limited ourselves to classifying items as DIF or non-DIF because we only consider the case when there are two groups of examinees. However, when we would consider more than two groups of examinees, we could induce several latent classes of items based on the performance of the items among the groups of examinees. It is, for instance, possible that one item functions differently for groups one and two, but that a second item functions differently for groups one and three. Also, it would be interesting to extend the RIM by including external covariates (this can also be done with only two groups of examinees). Taking this specification into the model would provide a good basis on which the underlying causes of DIF could be inferred.

Second, the current RIM could also be extended to model nonuniform DIF. Nonuniform DIF exists when the difference in probability of success between two groups is not constant across all ability levels, hence there is an interaction between group membership and ability level. We have limited ourselves in this paper to model only uniform DIF, in which there is a constant difference between group-item performance over ability levels and no interaction between group membership and ability level (Mellenbergh, 1982).

Third, it is also possible to introduce a mixture component on the examinee side. This would mean one deals with a double mixture model, namely a mixture on the item side and a mixture on the examinee side. Bolt and Cohen (2005) already proposed a model with a mixture component on the examinee side and concluded that this could be a way to enhance the understanding of the causes underlying DIF in test items.

Fourth, it is also possible to include anchor items in the model. If one could know beforehand that some items are DIF-free, this information could easily be

incorporated into the model by specifying some items as having the same parameter values across latent classes. However, a priori identifying items as DIF-free is very difficult. Note in this respect that the RIM can be used as a way to identify anchor items. The model provides posterior DIF probabilities that can be used to select anchor items; for instance, with very small estimated DIF probabilities.

This mixture approach may also be an excellent tool for identifying a purified set of items that could serve as the anchors in one of the other DIF analyses. In this way, the researcher could take advantage of the mixture model not needing an anchor and still use a method which may be more powerful and would provide the well-known effect sizes that are commonly used to interpret DIF.

Appendix

WinBUGS Code

```
{
#N1: number of examinees in the first group
#N2: total number of examinees
#T: number of items
#theta: the ability of person i
#sigmath: the standard deviation of the abilities of the
  examinees
#gamma: the difference in mean ability for the two groups of
  examinees
#beta1: the item difficulties for the first group of examinees
#beta2: the item difficulties for the second group of examinees
#mu1: the average item difficulty of the items
#p.var1: the variance of the item difficulties for the non-DIF
  items
  and the DIF items of the first group of examinees
#p.var2: the variance of the item difficulties for the DIF items
  of the
  second group of examinees
#rho: the correlation between the two groups in the DIF-class

# fitting a rasch model for the first group of examinees
{
for (i in 1:N1)
{
for (k in 1:T)
{
resp[i,k]~dbern(p[i,k]) #the responses are sampled from a
Bernoulli distribution
logit(p[i,k])<-sigmath*theta[i]-beta1[k]
}
}
theta[i]~dnorm(0,1)
}
```

```
# fitting a rasch model for the second group of examinees
for (i in N1+1:N2)
{
  for (k in 1:T)
  {
    resp[i,k]~dbern(p[i,k]) #the responses are sampled from a
      Bernoulli distribution
    logit(p[i,k])<- sigmath*theta[i]-beta2[k]+gamma
  }
  theta[i]~dnorm(0,1)
}

p.prec1<-pow(p.var1,-1)
# if the latent indicator Z equals 2 for a certain item k (i.e.,
no dif), then the same item
difficulty is estimated for both person groups
(beta1[k]=beta2[k]). If the latent
#indicator Z equals 1 for a certain item k (i.e., dif), then the
conditional probabiltiy for
that item is being estimated.

for (k in 1:T)
{
  beta1[k]~dnorm(mu_beta1,p.prec1)
  mu.a[k]<-(beta1[k])*D[k]+(mu_beta1+rho*(beta1[k]-mu_beta1)
    *sqrt(p.var2/p.var1))*(1-D[k])
  var.a[k]<-1000*D[k]+(1/p.var2*(1-pow(rho,2)))*(1-D[k])
  beta2[k]~dnorm(mu.a[k],var.a[k])

  D[k]<-step(Z[k]-1.5)
  Z[k] dcat(P[1:2])
}

# specifying priors
p.var1<-pow(std1,2)
p.var2<-pow(std2,2)
mu_beta1~dnorm(0,1)
rho~dunif(-1,1)
std1~dunif(0,3)
std2~dunif(0,3)
sigmath~dunif(0,3)
P[1:2]~ddirch(kappa[])
kappa[1]<-1
kappa[2]<-1
gamma dnorm(0,0.01)
}
```

Acknowledgments

The research reported in this paper was partly supported by the Research Fund of the University of Leuven (GOA/2010/02) and the IAP research network (P6/03) of the Belgian Government. We are grateful to Rianne Janssen for providing us with a data set.

Notes

¹For more information, consult Gonzalez, Tuerlinckx, De Boeck, and Cools (2006) in which a quasi-Monte Carlo method is proposed for high-dimensional problems.

²The number of items that the ST p -DIF method identifies as DIF items depends of the interval one uses to decide whether an item has DIF or not. In this paper, we used the $-.10$ to $.10$ interval. Another interval gives other results, for instance using the $-.05$ to $.05$ interval there are 11 items classified as DIF items.

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.
- Bolt, D. M., & Cohen, A. S. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133–148.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26, 381–410.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.
- 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.
- Chaimongkol, S., Huffer, F. W., & Kamata, A. (2006). A Bayesian approach for fitting a random effect differential item functioning across group units. *Thailand Statistician*, 4, 27–41.
- Chaimongkol, S., Huffer, F. W., & Kamata, A. (2007). An explanatory differential item functioning (DIF) model by the WinBUGS 1.4. *Songklanakarin Journal of Science and Technology*, 29, 449–458.
- Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6, 269–279.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533–559.
- 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.
- Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 3, 347–360.
- Fidalgo, A. M., Mellenbergh, G. J., & Muniz, 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, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, Sibtest, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29, 278–295.
- Fischer, G., & Molenaar, I. (1995). *Rasch models: Foundations and recent developments*. New York, NY: Springer-Verlag.
- Geary, D. C., & DeSoto, M. C. (2001). Sex differences in spatial abilities among adults from the United States and China: Implications for evolutionary theory. *Evolution and Cognition*, 7, 172–176.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, England: Cambridge University Press.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–472.
- Glas, C. A. W., & van der Linden, W. J. (2001). *Modelling variability in item parameters in item response models* (Technical Report). Enschede, the Netherlands: University of Twente.
- Gonzalez, J., De Boeck, P., & Tuerlinckx, F. (2008). A double-structure equation model for three-mode data. *Psychological Methods*, 13, 337–353.
- Gonzalez, J., Tuerlinckx, F., De Boeck, P., & Cools, R. (2006). Numerical integration in logistic-normal models. *Computational Statistics and Data Analysis*, 51, 1535–1548.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, 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.
- Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. O'Connell & D. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345–388). Charlotte, NC: Information Age.
- Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of Statistics*, 9, 45–58.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105–118.
- Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, 19, 9–15.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, 40, 397–404.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: The Danish Institute for Educational Research.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.

- Sandoval, J., & Mülle, M. E. (1980). Accuracy of judgments on WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology*, 48, 249–253.
- Snijders, T. A. B. (2005). Models for longitudinal network data. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 215–215). New York, NY: Cambridge University Press.
- Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, 34, 348–377.
- Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York, NY: Springer-Verlag.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118–128.
- 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. 147–169). Hillsdale, NJ: Erlbaum.
- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6, 181–195.
- 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.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117, 250–270.
- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185–201.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22–29.
- 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.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.

Authors

- SOFIE FREDERICKX is a Researcher at the Fund for Scientific Research-Flanders (FWO), Department of Psychology, K.U. Leuven, Tiensestraat 102, 3000 Leuven, Belgium; sofie.frederickx@psy.kuleuven.be. Her primary research interests include statistical methods and personality.
- FRANCIS TUERLINCKX is Professor of Psychology, Department of Psychology, K.U. Leuven, Tiensestraat 102, 3000 Leuven, Belgium; francis.tuerlinckx@psy.kuleuven.be. His research interests include the mathematical modeling of behavior (item responses, reaction times, dynamical aspects of behavior) and statistics.
- PAUL DE BOECK is Emeritus Professor at K.U. Leuven and Full Professor at the University of Amsterdam, Department of Psychology, Roeterstraat 15, NL-1018WB Amsterdam; paul.deboeck@uva.nl. His primary research interests include item response models.
- DAVID MAGIS is Post-Doctoral Researcher for the National Funds for Scientific Research (FNRS), Department of Mathematics, University of Liège, Grande Traverse 12, B-4000 Liège, Belgium; david.magis@ulg.ac.be. His primary research interests include statistical methods in psychometrics.