Simultaneous Multifactor DIF Analysis and Detection in Item Response Theory

Flávio B. Gonçalves[ab], Dani Gamerman[b] and Tufi M. Soares[a]

[a] Departamento de Estatística e Centro de Políticas Públicas e Avaliação da Educação, Universidade Federal de Juiz de Fora

[b] Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro

Abstract

In this paper, two integrated Bayesian models for differential item functioning (DIF) analysis in item response theory (IRT) models are proposed and compared. The model is integrated in the sense of modelling the responses along with the DIF determination, thus allowing DIF detection and DIF explanation in a simultaneous setup. DIF occurs because item characteristics may change according to grouping factors and its explanation is provided in the form of mixed models. Practical situations may lead to the simultaneous consideration of a number of factors. This gives rise to an extended class of models also introduced in this paper. The hypothesis of multifactorial DIF with possibly different explanations for each factor or combination of factors is also introduced. Important practical issues concerning identifiability of the models and the convergence of MCMC algorithms are discussed and illustrated in simulated examples. A real data set analysis of a Mathematics exam applied nationally to Brazilian elementary school students is performed considering two DIF factors: geographical region and type of school. The results highlight the relevance of the proposed methodology and of each of its model components to address important issues in educational studies and testing.

Key Words: Bayesian analysis, Item Response Theory, Differential Item Functioning, MCMC, multifactor.

1 Introduction

Standard item response theory (IRT) is based on models that assume that items behave equally to all individuals. Differential Item Functioning (DIF) allows an item to function differently among groups in its usual characteristics: discrimination, difficulty and guessing. Namely, if an item has DIF, students from different groups and with same proficiency may have different probability of correctly answering it. It is important to have items without DIF (usually called anchor items) in order to identify the proficiencies and the DIF's correctly.

Differential Item Functioning (DIF) analysis has been very important in educational research since the 1960's. Its importance is shown in a large number of applied studies. For instance, studies on performance differences in items of educational assessment tests (like GRE, SAT, GMAT, etc.) in groups defined by different ethnic characteristics, gender and socioeconomic status have been frequently shown in literature (*eg.* O'Neil & McPeck (1993), Shimitt & Bleinstein (1987), Berberoglu (1995), Gierl et al. (2003) and cited references). Therefore, it is not surprising that several statistical methods have been developed in order to support empirical analysis. There is a profusion of methods for the detection of items with DIF. Some of the most

used ones are: Mantel-Haenszel statistic based procedures (particularly the MH D-DIF statistic, *cf.* Dorans & Holland (1993)), the logistic regression method (Swaminathan & Rogers (1990)), the Simultaneous Item Bias Test - SIBTEST (Shealy & Stout (1993)) and methods that use the item parameters from Item Response Theory (IRT) models (Lord (1980); Thissen, Steinberg & Wainer, (1993)). Other methods are found, for example, in Clauser & Mazor (1998).

The first three methods cited above depend, directly or indirectly, on a previous estimation of the ability or from alternative criteria to match the individuals, which, in general, can not be dissociated from the DIF existence. Therefore, the ability purification in successive stages is recommended where the DIF item detected (in each stage) are eliminated from the ability calculation for the next analysis (Holland & Thayer (1988); Wang & Su (2004)). On the other hand, the methods based on IRT models construction postulate two approaches for model identification. The first one consists in setting the mean of the DIF parameters equal to a constant, typically zero. In the second one, a subset of anchor items, for which the non existence of DIF is assumed, is defined *a priori*, like the IRT-LR and IRT-D^2 (Thissen, Steinberg & Wainer (1993)). In other approaches, the anchor items can remain constant or vary in different stages depending on the results of the tests (Wang & Yeh (2003)). In any case, all these proposals involve different stages of DIF parameters detection and decision about which items have DIF. May (2006) proposes an extension of the standard graded response model (Samejima, 1997) which allows the overall threshold (difficulty) and discrimination parameter to vary across different groups, except for a set of anchor itens.

The detection of items with DIF is an important step in DIF analysis, but a complete analysis also requires satisfactory classification of the DIF found, identification of the factors associated to DIF along with the respective hypotheses formulation related to the DIF causes and perhaps, a hypotheses confirmatory analysis. For example, Schmitt, Holland & Dorans (1993) suggest that specially planned studies should be used to confirm the hypotheses formulated from the DIF factors study. In this context, it is natural to construct regression models that associate covariates, related to the items, to the DIF's magnitude. The covariates would represent the DIF factors in such a way that the results of the regression analyses would confirm or not the formulated hypotheses.

Longford, Holland & Thayer (1993) proposed a random effects regression model where the parameters of the model vary according to different forms of the test administration and, in addition, the DIF's magnitude is explained by variables correlated to the items (in particular, a difficulty measure of the item). Rogers & Swaminathan (2000) used characteristics of the individuals on the second level of their two levels model to improve the equalization between the members of the reference and focal groups. Swanson et al. (2002) propose an extension of the logistic regression method of Swaminathan & Rogers (1990) imposing an hierarchical structure where characteristics related to the items via covariates are included on the second level of the regression model allowing the confirmation or rejection of the hypotheses about DIF. Once again, all the three approaches above require the previous determination of the ability or of a DIF measure. In addition, the detection and the explanation (through associated factors) steps are performed separately.

The importance of the Bayesian approach have been steadily growing in IRT (see, for example, Albert (1992), Patz & Junker (1999a), Patz & Junker (1999b), Fox & Glas (2001), Béguin & Glas (2001)). Janssen, Schepers & Peres (2004) propose a model with item group predictors using a Bayesian approach. DIF analysis is a particularly appropriate environment for a genuine Bayesian formulation due to the complex structure of the models and the subjective decision features involved, which can be naturally formulated through the Bayesian argument. For ex-

ample, Zwick, Thayer & Lewis (1999, 2000) and Zwick & Thayer (2002) consider a formulation where the MH D-DIF statistic is represented by a normal model where the mean is equal to the "real DIF parameter" for which a normal prior distribution is explicitly considered. These authors use the Empirical Bayes (EB) approach for the posterior estimation of the parameters. Sinharay et al. (2006) consider the same formulation and propose informative prior distributions based on past information and show that the "Full Bayes" (FB) method leads to improvements if compared to the two other approaches, specially in small samples. More recently, Soares, Gonçalves & Gamerman (2006) proposed a methodology for incorporating DIF detection along with parameter estimation from a Bayesian perspective.

This paper presents an integrated Bayesian approach for DIF detection and analysis. A model based on Gonçalves (2006) is introduced and compared with the model proposed by Soares, Gonçalves & Gamerman (2006). Both models simultaneously detect items with DIF in difficulty and discrimination and estimate all the other parameters of the model. They additionally provide explanation of DIF through covariates.

All models above consider the existence of DIF in the item characteristics and/or allow for changes in the population model for the proficiencies according to a single classifying factor. Examples of factors include: gender, ethnicity, geographical region, type of school. However, these factors may jointly intervene in the analysis. The different forms this can take place range from no interactions between them to the saturated model with all possible forms of interaction.

In this paper, practical considerations of the data set lead to the consideration of multiple factors. Wang (2000) presents a model for multifactor analysis of DIF considering only the Rasch Model. In his model, at least one item has to be anchored (i.e., believed to have no DIF) for model identification. Besides that, the DIF detection is not made simultaneously with the estimation of the other parameter and the DIF parameters do not have a structure of explanation through covariates.

The models introduced in this paper consider the 3 parameters logistic model (Birnbaum (1968)), and do not required neither any previously set anchor item nor setting the mean of the DIF parameters equal to zero. Besides that, the DIF detection is made simultaneously with the estimation of the other parameter and a regression structure is used to explain the DIF parameters using covariates. The model identification is achieved by imposing prior distributions either to the DIF parameters and/or to the DIF probability parameters. In the second situation, prior distributions may be asymmetric or not.

Given the large dimensionality of the parameter space, it becomes important to understand the implications of the model assumptions. Some issues of practical relevance concerning prior distributions, the identifiability of the model and the convergence of the MCMC algorithm will also be discussed. Simulated data sets are used to illustrate the capabilities of the model in a variety of settings. The models are then used to analyse a data set concerning a Mathematics Exam applied to students of the 4^{th} grade of elementary school in Brazil. The students are firstly separated by a factor given by the country's main regions. Important cultural differences between the regions are shown to provide a substantial impact to the study. Later, another grouping factor associated with the type of school is also incorporated to the analysis. Different multifactor DIF models are then formed and compared.

Section 2 shows the proposed models for DIF Analysis. Section 3 presents aspects of the Bayesian inference procedure. In Section 4 simulated studies are applied to the proposed models and Section 5 shows a real data analysis where the proposed models are applied.

2 Models for DIF Analysis

Typically, in educational assessment, a test is formed by I items, but student j only answers a subset $I(j)$ of these items. Let $Y_{ij},\ j = 1, \ldots, J$, be the score attributed to the answer given by the student j to the item $i \in I(j) \subset \{1, \ldots, I\}$. Only the dichotomic case, where one of the scores in $\{0, 1\}$ is attributed to the item will be considered. This way, $Y_{ij} = 1$ if the answer is correct and $Y_{ij} = 0$ if the answer is wrong. In general, there can be different types of DIF (see Hanson (1998) for a wider characterization), but restricted to the characteristics which are made explicit by the three parameters model - 3PL (Birnbaum (1968)), the types of DIF can be immediately characterized according to the difficulty, discrimination and guessing.

2.1 3PL Model with DIF

Define $P(Y_{ij} = 1) = p_{ij}$, and $\Delta_{ij} = logit(p_{ij}) = \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right)$. Then, $p_{ij} = logit^{-1}(\Delta_{ij}) = \frac{1}{1 + e^{-\Delta_{ij}}}$. The main structure of the models used in this paper to associate the student's answer to his/her ability is:

$$\begin{aligned} &P\left(Y_{ij} = 1 | \theta_j, a_{ig}, b_{ig}, c_{ig}\right) = c_{ig} + (1 - c_{ig}) logit^{-1}(\Delta_{ij}) \\ &\text{where} \quad \Delta_{ij} = D a_{ig}(\theta_j - b_{ig}) \end{aligned} \tag{1}$$

for $i = 1, \ldots, I\ ,\ j = 1, \ldots, J$ and g denotes the group the individual belongs to, where $g = 1, ..., G$ is assumed to be known for all individuals, with $G < J$.

DIF is explicited in the model by setting the item's discrimination parameter as $a_{ig} = e^{-d^a_{ig}} a_i$, the item's difficulty parameter as $b_{ig} = b_i - d^b_{ig}$ and the item's guessing parameter as $c_{ig} = c_i\ (\in [0, 1]), \forall g$ where d^h_{ig} is the DIF parameter for discrimination and difficulty respectively, for $h = a, b$. For identification, $d^h_{ig} = 0$ is set for the reference group $g = 1$ and $h = a, b$. Thus, the DIF related to the discrimination or difficulty of item i in group g is $d^h_{ig} \neq 0$ whenever $g = 2, \ldots, G$, for $h = a, b$, respectively. It represents the DIF related to the difficulty or discrimination of the item in group g with respect to group 1. Although conceptually possible, DIF is not incorporated to the guessing parameter due to the known difficulties in the estimation of this parameter and by practical restrictions.

It can also be assumed that the students proficiencies are affected by the grouping factor and thus, they are modelled as $\theta_j | \lambda_g \sim N(\mu_g, \sigma^2_g)$, where $\lambda_g = (\mu_g, \sigma^2_g)$, where $g = 1, ..., G$. Some constraints are also required for identification of the mean and variance parameters for the proficiencies in this model. In line with the restrictions for the DIF parameters, restrictions are set by taking $\mu_g = 0$ and $\sigma^2_g = 1$ when $g = 1$. Thus, $\lambda_g = (\mu_g, \sigma^2_g)$, for $g = 2, \ldots, G$, is unknown and must be estimated along with the other parameters.

2.2 DIF Detection

It is possible to structure DIF parameters in a variety of ways. An interesting approach is provided by mixed model through the inclusion of covariates and random effects to explain them. This mixed model can be placed as one component in a mixture while the other one accounts for the absence of DIF. This way, DIF can be detected and, if detected, explained simultaneously. One simple model accommodating for that assumes that

$$d^h_{ig} \sim \left(1 - \pi^h_{ig}\right) F_0 + \pi^h_{ig} N\left(w^h_{ig}\gamma^h_g, (\tau^h_g)^2\right), \tag{2}$$

where w_{ig}^h is a row vector containing the values of the covariates that are relevant to explain the DIF behaviour, γ_g^h is the vector of their coefficients and $\left(\tau_g^h\right)^2$ is the item's specific random factor variance in group g for parameter h.

Equation (2) defines a normal regression model where a design matrix W_g^h is formed with rows w_{ig}^h, $i = 1, \ldots, I$. The explanatory variables can be different for each group. When they are the same, the design matrix can be denoted by W^h. If the covariates to explain DIF for difficulty and discrimination are the same, the notation for the design matrix simplifies further to W.

Two classes of model can be proposed under (2). Model 1 assumes that $F_0 = N\left(0, s_i^2(\tau_g^h)^2\right)$ and $\pi_{ig}^h \in \{0, 1\}$, where s_i^2 is a suitably small number. This kind of mixture was proposed by George & McCulloch (1993) for variable selection in a regression model and was applied for DIF analysis by Soares, Gonçalves and Gamerman (2006). Model 2 assumes that $F_0 = \delta_0$, where δ_0 is a point-mass at zero and $\pi_{ig}^h \in [0, 1]$. Point-mass mixture ideas go back to Jeffreys (1939) and are the basis of point null versus composite alternative hypothesis testing from a Bayesian viewpoint. This kind of mixture is also used nowadays, for example, in gene expression genomics modeling (West *et al.* 2006).

The distribution in (2) can be rewritten with the use of latent variables z_{ig}^h as

$$d_{ig}^h | z_{ig}^h \sim N\left(\left(w_i^h \gamma_g^h\right) z_{ig}^h \ , \ [s_i^2]^{1-z_{ig}^h} (\tau_g^h)^2\right) \text{ and } z_{ig}^h | \pi_{ig}^h \sim Ber(\pi_{ig}^h), \tag{3}$$

where $s_i^2 = 0$ for Model 2, defining $0^0 = 0$. Despite this similarity, the range of possible values and consequently the prior distribution for the weights π is substantially different between the two models presented above and results below illustrate some of the consequences of this difference.

2.3 Multifactor Models

It is possible to consider more than one group division for the DIF analysis. This possibility is very relevant in practical situations when a number of factors such as gender, ethnicity, region and type of school affect the item´s operating characteristics. The existence of more than one grouping factor opens up a host of possible models to describe the DIF parameters d_{ig}^h.

Models entertained by categorical data analysis (Agresti, 2002) can be applied. On one extreme there is the saturated model where all interactions between factors are allowed and on the other extreme, the factors may not interact leading to models having only main factor effects. Intermediate models may allow for second or higher order interactions as in standard categorical models. These ideas can be applied for both proficiencies and DIF parameters as described below.

2.3.1 Multifactor DIF

The K-factor model for the DIF parameters is obtained by generalizing the 1-factor model as follows. Let $g = (g_1, \ldots, g_K)$ denote the group an individual belongs to according to its classification in each of the K grouping factors, where g_k is the group the individual belongs to in the k-th factor for $g_k = 1, ..., G_k$ and $k = 1, ..., K$. (For example, if sex and region are the factors of the model then $K = 2$, $G_1 = 2$ and $G_2 = 5$ for a country with 5 regions.) From this formulation, DIF parameters become denoted by $d_{i(g_1(j),\ldots,g_K(j))}^h$, for $i = 1, ..., I$ and $h = a, b$.

It is possible to have several models for both item parameters a and b by including or not each possible interaction between factors. A general form of the model for $h = a, b$ is

$$d^h_{i(g_1,...,g_K)} = \sum_{k=1}^{K} d^h_{i(g_k)} + \sum_{k=1}^{K}\sum_{l=1}^{K} d^h_{i(g_k,g_l)} + ... \tag{4}$$

The first summation in the right hand side considers only the main effects while the second considers 2nd order interactions between pairs of factors. Wang (2000) introduced these models for the DIF difficulty. Here, they are extended to account for DIF in discrimination but also to incorporate detection and explanation of DIF, as detailed below.

The multifactor modelling of the DIF can be combined with the mixed model structure (2) of the previous subsection. Thus, each main effect $d^h_{i(g_k)}$ or p-th order interaction parameter $d^h_{i(g_{k_1},...,g_{k_p})}$ $(2 \leq p \leq K)$ can be explained by a mixture allowing for simultaneous detection and explanation. For the main effects, the multi-factor model with detection and explanation is given by

$$d^h_{ig_k}|z^h_{ig_k} \sim N\left(\left(w^h_{ig_k}\gamma^h_{g_k}\right) z^h_{ig_k}\ ,\ [s^2_i]^{1-z^h_{ig_k}}(\tau^h_{g_k})^2\right) \text{ and } z^h_{ig_k}|\pi^h_{ig_k} \sim Ber(\pi^h_{ig_k}), \tag{5}$$

where $w^h_{ig_k}$ is the vector of the DIF explanatory variables for item i, $\gamma^h_{g_k}$ is the vector of their coefficients and $\left(\tau^h_{g_k}\right)^2$ is the item's specific random factor variance, for group g_k of factor k for parameter h, for $g_k = 1, ..., G_k$, $k = 1, ..., K$ and $h = a, b$. Similar mixed models can be built for detection and explanation of interaction effects.

The saturated model is obtained when all possible interactions are considered and is equivalent to the model with a single grouping factor with $\prod_{k=1}^{K} G_k$ groups. The DIF model having only main effects and no interaction is given by the first terms in the right hand side of expression (4). In any case, some identification restrictions must be imposed. For simplicity, we assumed for main effects that $d^h_{ig_k} = 0$ when $g_k = 1$ and $d^h_{ig_k} \neq 0$ whenever $g_k = 2, ..., G_k$, for $k = 1, ..., K$ and $h = a, b$. Under this parametrization, $d^h_{ig_k}$ represents the DIF parameter of item i in group g_k compared to group 1 for factor k, $k = 1, ..., K$.

2.3.2 Multifactor Proficiencies

Similar ideas may be applied for the proficiencies as a number of factors may intervene and induce heterogeneity in the characterization of the student population, possibly the same factors that affect the items characteristics. For the multifactor modelling of the proficiencies it is assumed that, *a priori*, $\theta_j|\lambda_{(g_1,...,g_K)} \sim N(\mu_{(g_1,...,g_K)}, \sigma^2_{(g_1,...,g_K)})$, where $\lambda_{(g_1,...,g_K)} = (\mu_{(g_1,...,g_K)}, \sigma^2_{(g_1,...,g_K)})$.

Same comments made above are applied here and a number of models may be considered according to the each interactions between factors is allowed. If the factors do not interact in the specification of both population parameters for the proficiencies, then $\mu_{(g_1,...,g_K)} = \sum_{k=1}^{K} \mu_{g_k}$ and $\sigma^2_{(g_1,...,g_K)} = \prod_{k=1}^{K} \sigma^2_{g_k}$, for $i = 1, \ldots, I$, $j = 1, \ldots, J$. The usual additive form is used for variances in the logarithmic scale. This is not compulsory but makes interpretation and identification easier.

Some constraints are required for identification of the mean and variance parameters for the proficiencies in this model. These restrictions are set for the main effects by taking $\mu_{g_k} = 0$ and $\sigma^2_{g_k} = 1$ when $g_k = 1$, for $k = 1, ..., K$. Similar constraints are applied to interaction effects. Thus, $\lambda_{g_k} = (\mu_{g_k}, \sigma^2_{g_k})$, for $g_k = 2, \ldots, G_k$ and $k = 1, \ldots, K$, is unknown and must be estimated along with the other parameters.

In this paper, special attention is given to the model only with main effects (no interaction) and the saturated model (all interactions) for both DIF parameters and proficiencies. They have a strong and meaningful interpretation in the light of the data under analysis.

3 Bayesian Inference

This section explains the Bayesian inference process and discusses some relevant issues concerning the prior distribution of the parameters $\pi^h_{ig_k}$ in Model 2, the identifiability of the models and the convergence of the MCMC algorithms.

3.1 Inference

Define the parameter vector $\Psi = \{\Psi_1, \Psi_2, \Psi_3\}$, divided into components $\Psi_1 = \{\theta, \lambda\}$, $\Psi_2 = \{a, b, c\}$ e $\Psi_3 = \{d, \pi, \gamma, \tau\}$, and each element in the components represents the corresponding parameter with all possible indexes, for example $\theta = \{\theta_1, \ldots, \theta_J\}$. The main aim of the inference process is to obtain the joint posterior $p(\Psi|Y)$ distribution given by:

$$p(\Psi|Y) = p(\Psi|Y, W) = \frac{p(Y|\Psi)p(\Psi|W)}{\int \ldots \int p(Y|\Psi)p(\Psi|W)d\Psi}, \tag{6}$$

where the vector Y contains all data values Y_{ij} that are observed.

Since it is not possible to obtain the complete analytic expression for the density above, MCMC (Markov chain Monte Carlo) methods are used to draw samples of this distribution and approximate it (see Gamerman & Lopes (2006) for details).

It is assumed, *a priori*, that Ψ_1, Ψ_2 and Ψ_3 are independent, that is: $p(\Psi) = \prod_{i=1}^{3} p(\Psi_i)$. The prior below is detailed for the multifactor model for DIF and proficiencies only with main effects. The prior for other DIF and proficiencies models with interaction between effects differ only in a few details and are therefore omitted for conciseness.

The prior for the first group of parameters is given by

$$p(\Psi_1) = \prod_{j=1}^{J} p(\theta_j|\mu_{(g_1,\ldots,g_K)}, \sigma^2_{(g_1,\ldots,g_K)}) \prod_{k=1}^{K} \prod_{g_k=2}^{G_k} p(\mu_{g_k})p(\sigma^2_{g_k})$$

where:

$$(\theta_j|\mu_{(g_1,\ldots,g_K)}, \sigma^2_{(g_1,\ldots,g_K)}) \sim N\left(\sum_{k=1}^{K} \mu_{g_k}, \prod_{k=1}^{K} \sigma^2_{g_k}\right),\ \mu_{g_k} \sim N\left(m_0, s_0^2\right) \quad \text{and} \quad \sigma^2_{g_k} \sim IG\left(\alpha_0, \beta_0\right),$$

for $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $g = 2, \ldots, G$, where IG denotes the Inverse Gama distribution.

The prior for the second group of parameters is given by

$$p(\Psi_2) = \prod_{i=1}^{I} p(a_i)p(b_i)p(c_i)$$

where $a_i \sim LN\left(m_{a_i}, s^2_{a_i}\right)$, $b_i \sim N\left(m_{b_i}, s^2_{b_i}\right)$ and $c_i \sim beta\left(\alpha_{c_i}, \beta_{c_i}\right)$, for $i = 1, \ldots, I$, where LN denotes the Log-Normal distribution.

The prior distribution of Ψ_3 is

$$p(\Psi_3|W) = \prod_{h=a,b} \prod_{k=1}^{K} \prod_{g=2}^{G} \left(\prod_{i=1}^{I} p\left(d^h_{ig_k} | \pi^h_{ig_k}, W^h_{g_k}, \gamma^h_{g_k}, (\tau^h_{g_k})^2\right) p(\pi^h_{ig_k}) \right) p(\gamma^h_{g_k}) p\left((\tau^h_{g_k})^2\right)$$

where $p\left(d^h_{ig_k} | \pi^h_{ig_k}, W^h, \gamma^h_{g_k}, (\tau^h_{g_k})^2\right)$ is given in (3), $\pi^h_{ig_k} \sim Ber\left(\xi^h_{ig_k}\right)$ in Model 1, $\pi^h_{ig_k} \sim beta\left(\alpha_{\pi^h_{ig_k}}, \beta_{\pi^h_{ig_k}}\right)$ in Model 2, $\gamma^h_{g_k} \sim N\left(m_{\gamma^h_{g_k}}, s^2_{\gamma^h_{g_k}} I_{L^h_{g_k}}\right)$ and $(\tau^h_{g_k})^2 \sim GI(\alpha_{\tau^h_{g_k}}, \beta_{\tau^h_{g_k}})$ for $h = a, b$, $i = 1, \ldots, I$, $k = 1, \ldots, K$ and $g = 2, \ldots, G$. Note that Model 2 has one hierarchical level more than Model 1. This is equivalent to setting $\pi^h_{ig} | \xi^h_{ig} \sim Ber(\xi^h_{ig})$ and $\xi^h_{ig} \sim beta(\alpha_{\pi^h_{ig}}, \beta_{\pi^h_{ig}})$.

3.2 Prior distribution of $\pi^h_{ig_k}$

Simulated studies in Gonçalves (2006) show that the prior distribution of the parameters $\pi^h_{ig_k}$ has a substantial influence on the joint posterior distribution, specially on the estimation of these parameters. The following Lemma helps to explain why. The proof is left for the Appendix.

Lemma 1 *If a prior distribution $beta(\alpha, \beta)$ is assumed for $\pi^h_{ig_k}$ in Model 2, then, its posterior mean is restricted to the interval* $\left[\dfrac{\alpha}{\alpha+\beta+1}, \dfrac{\alpha+1}{\alpha+\beta+1}\right]$.

If, for example, a beta(1,1) prior distribution is used then, from Lemma 1, the posterior mean of $\pi^h_{ig_k}$ will be restricted to the interval $[1/3, 2/3]$. Larger values of the parameters will make this interval even smaller. This is bound to make the classification of an item as having or not having DIF more difficult and, consequently, lead to a more difficult estimation of the other parameters of the model. For that reason, beta distributions with parameters smaller than 1 are recommended whenever the analyst lacks knowledge about them. This will be assumed as the prior distribution for the parameters $\pi^h_{ig_k}$ in this paper. This kind of distribution has a "bath tub" shape, that is, it is bimodal and concentrates most of its mass in the extremes of the interval $(0, 1)$.

Simulation studies presented in Gonçalves (2006) compare several prior distributions and concludes that the $beta(0.01,0.01)$ leads to better estimation of the other parameters of the model. This distribution will be used when there is little or no information about the item. If there is information only about the location of π, asymmetric distributions with small but unequal parameters such as $beta(0.01, 0.05)$ or $beta(0.07, 0.01)$, may be a possibility. The use of "bath tub" shape beta prior distributions makes the marginal posterior distribution of $\pi^h_{ig_k}$ bimodal, with the modes located in the extremes of the interval $(0, 1)$, as illustrated in the example shown in Figure 1.

Although $\pi^h_{ig_k}$ represents a probability, and, in principle, can vary freely in the unit interval, the kind of posterior distributions obtained indicates that bernoulli prior distributions, restricting the distribution of the $\pi^h_{ig_k}$'s to $\{0, 1\}$, may yield basically the same results. Figure 2 shows the posterior mean of the $\pi^h_{ig_k}$'s obtained in a simulated study using different prior distributions. The results of the $beta(0.01,0.01)$ and the bernoulli(0.5) (both symmetric with mean 0.5) are very close. When using asymmetric distributions, more difference may be found. In any case, bernoulli distributions seem to provide more discrimination. This feature may be helpful when analysing real datasets.

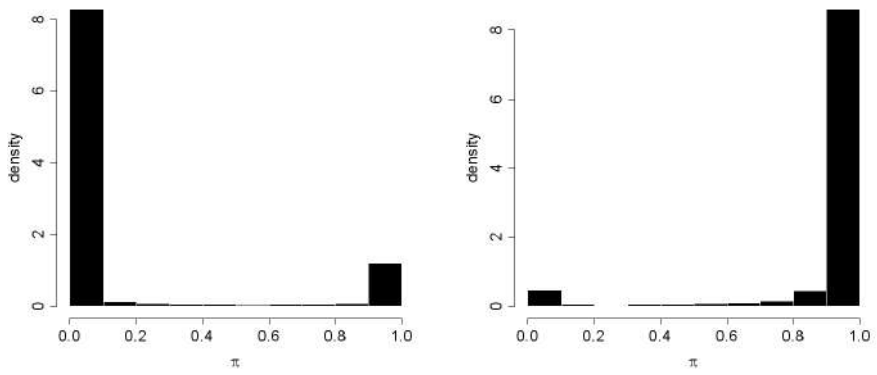

Figure 1: *Histograms of the marginalized posterior distributions of $\pi^a_{4,2}$ (left) and $\pi^b_{4,2}$ (right) for item 4 in a simulated example with one factor ($K = 1$) and 2 groups ($G = 2$). The prior distribution of these parameters is a beta(0.01,0.01). This item only has DIF in the difficulty and this is correctly estimated as the histograms show.*

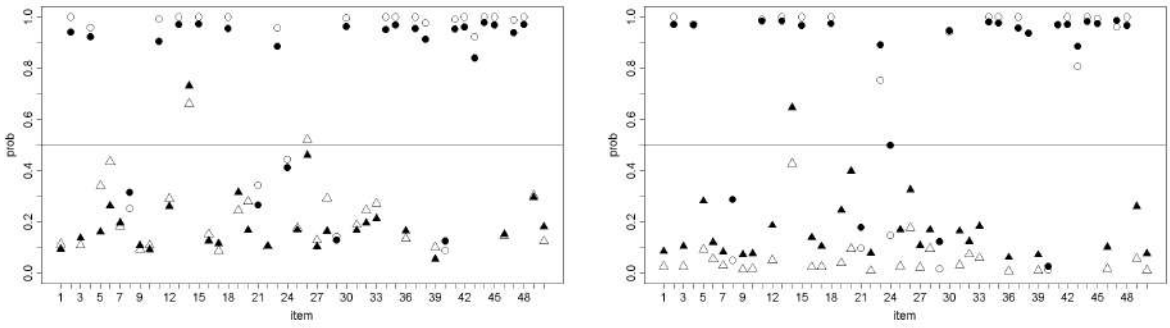

Figure 2: *Posterior mean of $\pi^b_{i,2}$ in the same simulation data set of Figure 1 using different prior distributions. Left panel shows symmetric priors with mean = 0.5: beta$(0.01, 0.01)$ - solid; bernoulli(0.5) - hollow. Right panel shows asymmetric priors with mean = 0.2: beta$(0.01, 0.04)$ - solid; bernoulli(0.2) - hollow. Circles represent items with DIF and triangles represent items without DIF.*

3.3 Model identifiability and convergence issues

Note that the likelihood function of both models is not identifiable, since any transformation $\theta_j{}^* = \frac{\theta_j + r_2}{e^{-r_1}} + \frac{b_i(e^{-r_1} - 1)}{e^{-r_1}}$, $\left(d^a_{ig_k(j)}\right)^* = d^a_{ig_k(j)} + r_1$ and $\left(d^b_{ig_k(j)}\right)^* = \frac{d^b_{ig_k(j)} - r_2}{e^{-r_1}}$, $\forall\ j$ for which $g_k \geq 2$, $k = 1, \ldots, K$, $g = 2, \ldots, G$, $i = 1, \ldots, I$ and $r_1, r_2 \in \mathbb{R}$, gives the same probability for the individual j to correctly answer item i. Simulated studies presented in the next section show that the parameters of the model are well estimated despite that. The reason can only be attributed to the influence of the prior distribution. It identifies the parameters and thus makes the posterior distribution identifiable.

This is a situation encountered in other statistical models; the likelihood is not identifiable or is not efficient to estimate the parameters, and the prior distribution "corrects" the likelihood misbehavior by forming a posterior distribution that estimates the parameters effectively (see,

for example, Steel & Fernández (1999) and Liseo & Loperfido (2006) for different instances of this phenomenon).

Although the posterior distribution corrects the likelihood problems, it is still possible to have problems when using MCMC algorithms; and they may converge to secondary, unimportant modes. The choice of the initial values of the chains has a strong influence on the convergence. Simulated examples presented in Gonçalves (2006) show that it is crucial to set the initial values of the parameters $d^h_{ig_k}$ as zero to obtain good estimates for both models' parameters. Different initial values may lead to convergence to secondary modes.

3.4 Inference Procedures

The decision on the classification of an item as having or not DIF is based on the parameters $\pi^h_{ig_k}$. A simple classification rule decides based on the posterior mean of $\pi^h_{ig_k}$. If it is greater than a threshold value p^*, item i is classified as having DIF in parameter h, in group g_k with respect to group 1 in factor k and vice-versa if it is smaller than p^*. Special attention must be paid to items for which $\pi^h_{ig_k}$ is close to p^*. In this paper it is assumed $p^* = 0.5$, but more elaborated criteria may be used to choose the value of p^*, for example, based on other scoring rules (see Migon & Gamerman (1999)). Once the item is classified, it is only necessary to analyse the DIF parameters of the items classified as DIF ones. For the other items, the DIF parameter is assumed to be 0.

Considering the scale of the model, DIF parameters with absolute value smaller than 0.1 may be considered insignificant, between 0.1 and 0.2 very small, between 0.2 and 0.3 small, between 0.3 and 0.5 medium and greater than 0.5 as a high DIF.

An important issue concerning DIF analysis is how to proceed in a situation with more than 2 groups in a given factor k. Since DIF in group g_k is detected and estimated with reference to group 1 (reference group), it seems sensible to perform the analysis as follows:
a) if a given item is detected as a non DIF one in L groups, then, there is no DIF between these L groups and the item behaves the same way in these groups;
b) if a certain item is detected as having DIF in a group and as not having DIF in another group, then, the item behaves differently with respect to the reference group only for the first group;
c) if a certain item has DIF in two groups, the magnitudes of the DIF´s can be compared by the estimates of the DIF parameters of this item in both groups since they have the same scale.

4 Simulated examples

This section briefly presents results of two simulated studies. The first one designed to compare the results obtained with Models 1 and 2 described in Section 2. The second one analyzes a two factor data set to show the model's ability to estimate the parameters.

In all examples, the parameters of the items and the proficiencies are randomly drawn. The discrimination parameters a_i are drawn from a $LN(-0.1, 0.25)$ distribution, the difficulty parameters b_i are drawn from a standard normal distribution and the guessing parameters c_i from a beta distribution with parameters $(25, 85)$. These values are chosen in the range typically encountered in our practical experience.

The data sets are composed of $I = 40$ dichotomic items and $J = 1000$ individuals in each group. The items with DIF are also randomly chosen. Around 30% of them have DIF in

discrimination and 40% in difficulty. A binary covariate is used to explain DIF in discrimination and difficulty.

The convergence of the Markov chains drawn from the MCMC was tested using the R criteria of Gelman & Rubin (RGR) (see Gamerman & Lopes (2006) for details), calculated from three parallel chains with 10000 iterations (after a burn-in of 5000) and different initial values. The test showed that all the chains converged with $RGR < 1.1$ for all of them, when parameters d_{ig}^h are initialized as zero in all the chains. The algorithm was implemented in Ox language.

The following prior distributions are used in all examples: $\mu_{g_k} \sim N(0,1)$, $\sigma_{g_k}^2 \sim IG(0.1, 0.1)$, $\gamma_{g_k}^h \sim N_2((0,0)', 10I_2)$ and $(\tau_{g_k}^h)^2 \sim GI(0.1, 0.1)$, for $h = a, b$, $g_k = 2, \ldots, G_k$ and $k = 1, 2$, $a_i \sim LN(0, 2)$, $b_i \sim N(0, 1)$ and $c_i \sim beta(5, 17)$ for $i = 1, \ldots, 40$, $\pi_{ig_k}^h \sim Ber(0.5)$ in Model 1 and $\pi_{ig_k}^h \sim beta(0.01, 0.01)$ in Model 2, for $h = a, b$, $i = 1 \ldots, 40$, $g_k = 2 \ldots, G_k$ and $k = 1, 2$.

4.1 Comparison of Models 1 and 2

Comparison between both models defined in (3) is based on the posterior mean of the π_{ig}^h's and for the d_{ig}^h's it is also possible to compare means against the actual values used in the simulation. The deviance information criterion (DIC) (Spiegelhalter et al., 2002) is used to compare the models in a more comprehensive form. Posterior means of the parameters of the proficiencies' distributions and the parameters of the DIF regression are also presented.

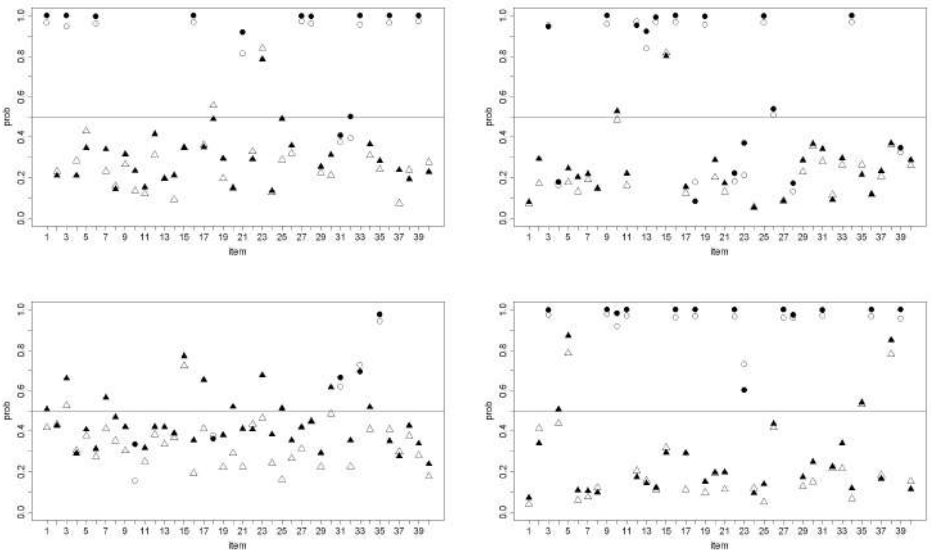

Figure 3: *Graphs of the posterior mean of parameters π_{ig}^h. Solid symbols represent Model 1 and empty symbols represent Model 2. Circles represent items with DIF and triangles items without DIF: π^a (left) and π^b (right), group 2 (top) and group 3 (bottom).*

Figure 3 shows that the results for the DIF probabilities π_{ig}^h are similar for most of the items for the two models for group 2 (when $g = 2$). Their posterior mean differs by more than 0.10 for eight items in the discrimination and for only two items in the difficulty. The results are also similar concerning difficulty in group 3 with only three items having their posterior mean differing more than 0.10. The results for group 3 are quite different for the discrimination; Model 1 presents twelve misclassified items (taking 0.5 as the cut-off point), Model 2 presents only four,

and one of them (item 10), for which the posterior mean is smaller than 0.5, has almost no DIF with $E(d^a_{10,3}|Y) = -0.035$.

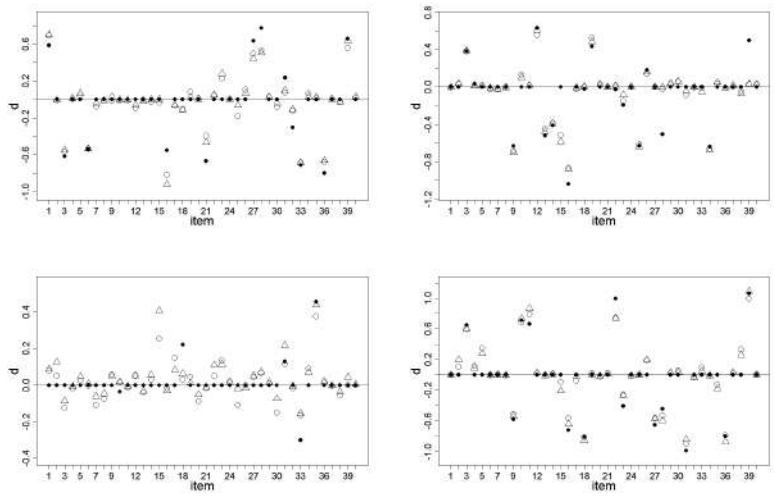

Figure 4: *Plots of the real values and posterior means of parameters d^h_{ig}. Solid circle represents the real value, empty circle represents the posterior mean in Model 1 and empty triangle in Model 2. d^a (left) and d^b (right), group 2 (top) and group 3 (bottom).*

	Discrimination	Difficulty
Model 1	82.5%	85%
Model 2	90%	87.5%

Table 1: *Percentage of correct classification in each model for discrimination and difficulty (considering even the items with insignificant DIF as DIF items).*

The results for the parameters d^h_{ig} are similar for all groups in the discrimination and in the difficulty (see Figure 4). Little difference is found in the posterior means of the parameters presented in Table 1 but Model 2 is a little more effective on the parameters estimation, with a smaller MSE.

	$\bar{D}$	p_D	DIC
Model 1	115834	-91213	24621
Model 2	115835	-90693	25142

Table 2: *$\bar{D}$, p_D and DIC in both Models.*

The DIC criterion is the sum of of two parts. The first one, $\bar{D}$, is the posterior expectation of the deviance and is a goodness of fit measure; the smaller it is, the best the model fits the data. The second part, p_D, measures the complexity of the model through the effective number of parameters; the smaller it is, the less complex is the model. It can be seen, from Table 2, that the value of $\bar{D}$ is essentially the same in both models, that is, the models are equally good to fit the data. Besides that, Model 1 has a p_D value smaller than Model 2, reflecting the greater complexity of Model 2 due to its extra hierarchical level.

Parameter	Real value	Model 1	Model 2
μ_2	0.12	0.114	0.098
μ_3	-0.03	-0.063	-0.070
σ_2	1	1.068	1.078
σ_3	2	2.124	2.134
γ^a_{02}	0.7	0.280	0.309
γ^a_{12}	-1.4	-0.734	-0.801
$\left(\tau^a_2\right)^2$	0.04	0.139	0.141
γ^b_{02}	-0.4	-0.532	-0.542
γ^b_{12}	0.8	0.780	0.834
$\left(\tau^b_2\right)^2$	0.0625	0.109	0.121
γ^a_{03}	0.4	0.189	0.222
γ^a_{13}	-0.8	-0.286	-0.334
$\left(\tau^a_3\right)^2$	0.0625	0.050	0.069
γ^b_{03}	-0.7	-0.555	-0.584
γ^b_{13}	1.4	1.084	1.140
$\left(\tau^b_3\right)^2$	0.04	0.127	0.116
	MSE	1.116	0.915

Table 3: *Real value and posterior mean of the parameters of the abilities' distribution and of the parameters of the DIF regression in both Models. MSE is their mean square error.*

These results of Table 3 indicate that little difference is found between the results obtained in each model for most of the parameters. Larger differences are found in the estimation of the parameters π^a_{i3}. Interestingly, this is the group for which there are less items with DIF. This indicates that Model 1 seems to be less efficient than Model 2 in situations with few DIF items.

4.2 Two factors example

In this section a data set with two factors drawn from the multifactor model with no interaction is analyzed using the no interaction and the saturated versions of the multifactor Model 2. The first factor has three groups and the second factor has two groups. A binary covariate is used to explain DIF in difficulty in both factors. The models are compared using DIC.

	Discrimination	Difficulty
Factor 1 - group 2	95%	87.5%
Factor 1 - group 3	82.5%	95%
Factor 2 - group 2	95%	100%

Table 4: *Percentage of correct classification for discrimination and difficulty in both factors.*

The model is again capable of good recovery of parameters as shown in previous simulated example. Table 4 shows that most of the items are correctly classified as having or not DIF. The model also estimates well the DIF parameters and, as in the previous example, only the parameters τ's are not well estimated.

Table 5 seems to favour the multifactor model having only main effects (with no interaction), as expected. Note that the value of $\bar{D}$ is a little smaller for the saturated model, which is reasonable, since it has more parameters but the difference is very small. But the value of p_D is considerably smaller in this model as the multifactor model with no interaction has much less parameters.

Another way to choose between models is to compare the estimates of the parameters of the

	$\bar{D}$	p_D	DIC
Multifactor Model	123150	-90669	32479
Saturated Model	123140	-88036	35108

Table 5: *$\bar{D}$, p_D and DIC in both Models.*

DIF regression structure and the parameters of the proficiencies' distributions in both models. Similarities between the estimates in both model is another indication of absence of interaction. Figure 5 presents this comparison. The estimates of the multifactor model with no interaction are obtained by summing (for the parameters of DIF regression structure and the mean of the proficiencies' distribution) or multiplying (for the variance of the proficiencies' distribution) the parameters of the respective group the student belongs to in each factor. For example, γ_0 of group 6 is obtained by adding γ_{03} from factor 1 with γ_{02} from factor 2, since group 6 correspond to the intersection of these two groups. All the parameters presented in Figure are very similar in both models providing indication of no interaction between the two factors in the DIF in discrimination and difficulty. It can also been seen that there is no interaction between the factors when considering the proficiencies.

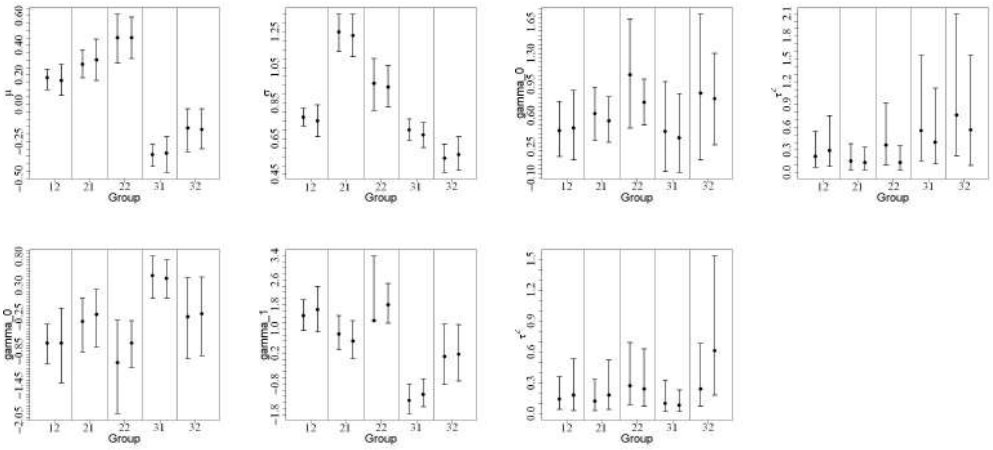

Figure 5: *Posterior mean and* 95% *credibility interval of the parameters μ, σ, γ_0^a, τ_a^2 (top row), γ_0^b, γ_1^b, τ_b^2 (bottom row) in each group for both models. Left interval: multifactor model with no interaction, right interval: saturated model.*

5 Application

The real data set analysed in this paper refers to SAEB, which stands for "Sistema de Avaliação da Educação Básica" (System for Assessment of Basic Education). It is coordinated by the Brazilian government and aims the assessment of the basic education system of the country since 1990. SAEB's assessments produce information on the Brazilian educational reality and, specifically, by regions, in private and public schools, through a bi-annual (since 1993) proficiency exam in Mathematics and Portuguese. It is applied to a sample of students in the 4^{th} and 8^{th} year of elementary school and in the 3^{rd} year of high school (for more information

on SAEB, visit *http://www.inep.gov.br/basica/saeb/ingles.htm*).

Brazil is separated in five great regions: South-East (SE), South (S), North-East (NE), Center-West (CW) and North (N), as shown in figure 6. South-East is the most economically developed one and has around 42% of the Brazilian population, South is the second most economical developed one. There are four types of elementary and high school in the country: private, federal, state and municipal schools. Private schools are thought to be better than state and municipal ones. Federal schools are known to provide good education but there are very few of them in Brazil. The data set analysed in this paper does not contain students from federal schools.

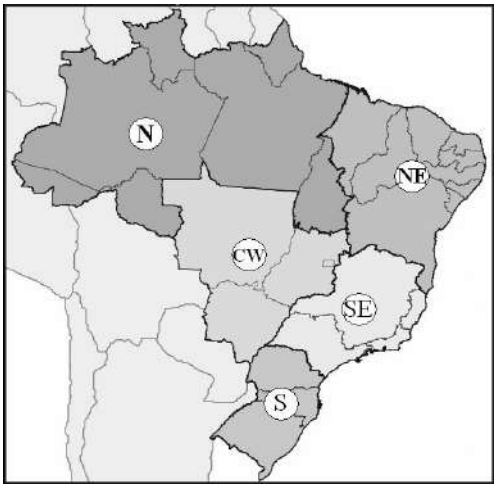

Figure 6: *Brazilian regions.*

The data set analysed in this paper refers to the Maths test applied to the 4^{th} grade in the year 2003. The 169 items are separated into thirteen blocks of thirteen items and the tests are formed by three blocks via incomplete balanced blocks. Each student answers to one of these tests. The tests are composed of multiple choice items with four options each, where only one is correct. 46131 students were evaluated. Due to computational price, a sample of 9930 students is randomly sampled to be analysed: 1986 from each region; 2686 from private schools, 3669 from state schools and 3575 from municipal schools. The Regions are chosen as the group division for the one factor DIF analysis. Region SE is considered the reference group (group 1), Region S is group 2, followed by Regions NE, CW and N, respectively. A two factor analysis incorporating the type of school is also performed with the multifactor model with no interactions and with the saturated model. Private school is the reference group, state school is group 2 and municipal school is group 3.

5.1 One factor analysis

Based on the content of the test's items and on the opinion of specialists on education, it is believed that some item contents may have some influence on DIF explanation for difficulty. Four covariates are then introduced in the model and the results are reported below. The covariates are binary and indicate if the item content is related to: numeric problems involving sum and subtraction, use of decimal system in number decomposition, area and/or perimeter calculation drawn in a square-lined mesh, and rational number representation using fractions. From the

169 items of the test, 85 are related to some of these contents. Table 6 shows the items related to each content.

Content	Items
Numeric problems involving sum and subtraction	2 6 7 13 14 23 24 29 37 38 39 44 50 56 58 61 62 65 67 71 77 80 84 90 103 104 106 113 123 124 135 136 139 141 142 146 148 149 152 153 156 160 162 165
Use of decimal system in number decomposition	4 8 30 33 43 54 64 75 91 107 117 143 144 151 163 168
Area and/or perimeter calculation drawn in a square-lined mesh	11 32 35 40 57 68 81 83 92 94 109 115 125 128 131 138
Rational number representation using fractions	20 21 72 86 95 98 114 137 158

Table 6: *Items related to each covariate.*

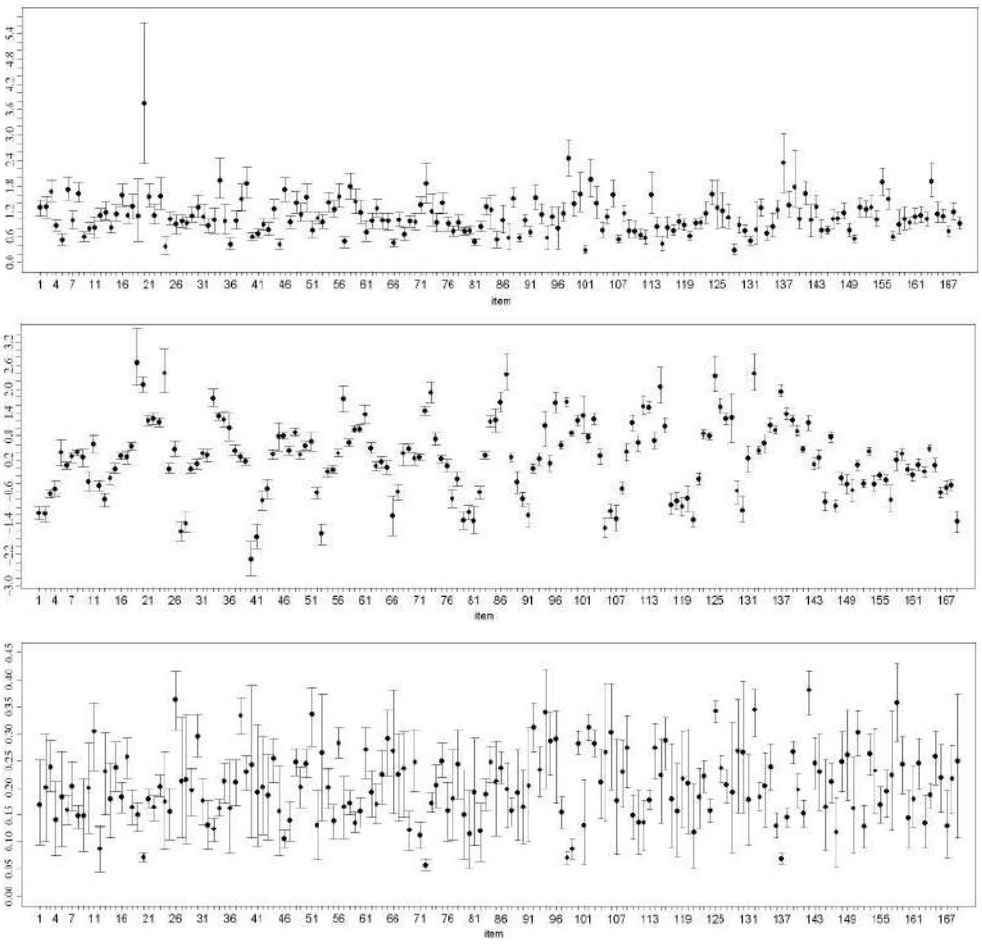

Figure 7: *Posterior mean and* 95% *credibility interval of the items parameters. Top row: Discrimination, Middle row: Difficulty, Bottom row: Guessing.*

Figure 7 presents the estimates of the item parameters in the reference group (South-East). The estimates of the discrimination parameters vary from 0.4 to 2.5 (except item 20 with discrimination 3.8) with mean 1.07 and variance 0.20. The items with highest discrimination in SE are 20, 98 and 137. They are also difficult items and have very small values (< 0.10) for the

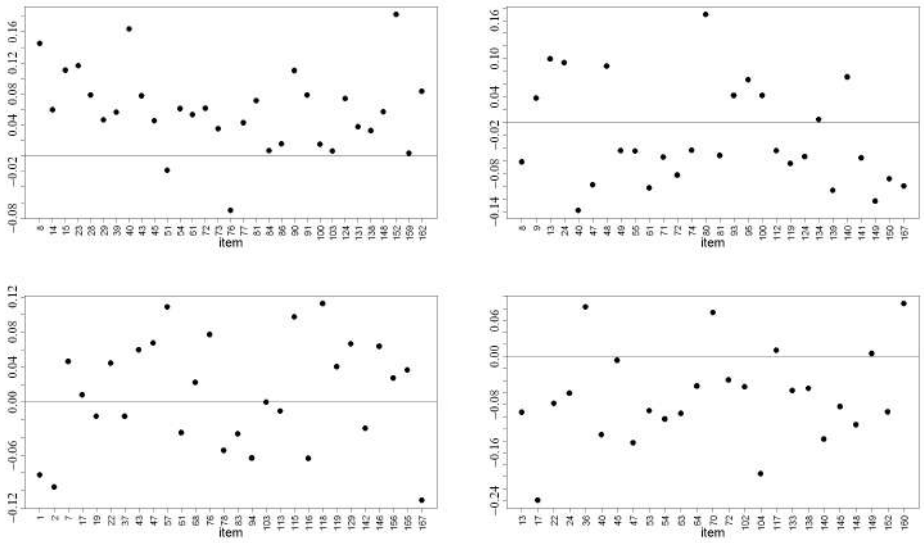

Figure 8: *Posterior mean and* 95% *credibility interval of the DIF parameters in the discrimination of the items detected as DIF ones. Top-left: S, Top-right: NE, Bottom-left: CW, Bottom-right: N.*

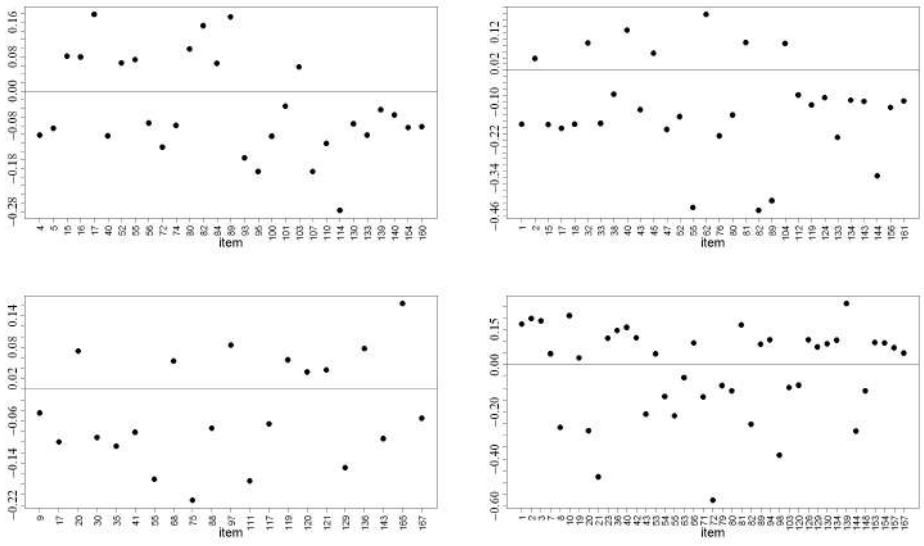

Figure 9: *Posterior mean and* 95% *credibility interval of the DIF parameters in the difficulty of the items detected as DIF ones. Top-left: S, Top-right: NE, Bottom-left: CW, Bottom-right: N.*

guessing parameter. The estimates of the difficulty parameters vary between -2.5 and 3.0 with mean 0.26 and variance 0.92. The most difficult items are 19, 20, 24, 87, 115, 125 and 132. The easiest ones are 27, 40 and 41. The estimates of the guessing parameters vary from 0.07 to 0.38 with mean 0.20 and variance 0.004.

Considering the threshold probability $p^* = 0.5$, 31 items have been detected as having DIF in discrimination in Region S, 28 in NE, 28 in CW, and Region N has 25 items detected as having

Parameter	Post. mean	Cred. interval
μ_2	-0.045	(-0.109 , 0.028)
μ_3	-0.676	(-0.746 , -0.607)
μ_4	-0.399	(-0.463 , -0.330)
μ_5	-0.784	(-0.867 , -0.691)
σ_2	0.960	(0.913 , 1.005)
σ_3	1.128	(1.063 , 1.198)
σ_4	1.003	(0.954 , 1.055)
σ_5	0.989	(0.934 , 1.041)
γ^a_{02}	0.037	(-0.032 , 0.104)
$\left(\tau^a_2\right)^2$	0.024	(0.012 , 0.043)
γ^b_{02}	-0.01	(-0.114 , 0.096)
γ^b_{12}	0.031	(-0.109 , 0.166)
γ^b_{22}	-0.086	(-0.366 , 0.141)
γ^b_{32}	-0.050	(-0.549 , 0.314)
γ^b_{42}	**-0.167**	**(-0.503 , 0.106)**
$\left(\tau^b_2\right)^2$	0.023	(0.012 , 0.042)
γ^a_{03}	-0.020	(-0.109 , 0.064)
$\left(\tau^a_3\right)^2$	0.030	(0.013 , 0.058)
γ^b_{03}	**-0.105**	**(-0.215 , 0.003)**
γ^b_{13}	**0.115**	**(-0.036 , 0.269)**
γ^b_{23}	-0.025	(-0.334 , 0.210)
γ^b_{33}	**0.158**	**(-0.181 , 0.483)**
γ^b_{43}	0.068	(-3.828 , 3.857)
$\left(\tau^b_3\right)^2$	0.033	(0.018 , 0.058)
γ^a_{04}	-0.013	(-0.064 , 0.101)
$\left(\tau^a_4\right)^2$	0.024	(0.012 , 0.044)
γ^b_{04}	-0.030	(-0.134 , 0.057)
γ^b_{14}	0.060	(-0.069 , 0.194)
γ^b_{24}	-0.078	(-0.278 , 0.155)
γ^b_{34}	0.000	(-0.442 , 0.298)
γ^b_{44}	0.074	(-2.364 , 2.77)
$\left(\tau^b_4\right)^2$	0.019	(0.009 , 0.034)
γ^a_{05}	-0.044	(-0.152 , 0.058)
$\left(\tau^a_5\right)^2$	0.032	(0.015 , 0.059)
γ^b_{05}	0.039	(-0.073 , 0.163)
γ^b_{15}	-0.019	(-0.185 , 0.137)
γ^b_{25}	**-0.201**	**(-0.511 , 0.046)**
γ^b_{35}	-0.050	(-0.304 , 0.439)
γ^b_{45}	**-0.514**	**(-1.082 , -0.070)**
$\left(\tau^b_5\right)^2$	0.034	(0.017 , 0.059)

Table 7: *Posterior mean and* 95% *credibility interval of the parameters of the abilities' distribution and of the parameters of the DIF regression. The first index of parameters* γ *refers to the covariate: 1- numeric problems involving sum and subtraction, 2- use of decimal system in number decomposition, 3- area and/or perimeter calculation drawn in a square-lined mesh, 4- rational number representation using fractions.*

DIF in this parameter. Concerning difficulty, 29 DIF items were detected in Region S, 30 in NE, 21 in CW and 41 in N. Note, from Figures 8 and 9, that some of the items detected as DIF ones have a very small value for the DIF parameter, meaning that this difference is practically irrelevant. Forty three (25% of total) items are detected as not having DIF in any groups and any parameters, these items are important to assure good equalization of the proficiencies.

It can be noticed from Figure 8 that most of the items with DIF in discrimination are more discriminant in region NE than in all other regions. The same happens in Region N. In Region S most of the DIF items are less discriminant than in the other regions. No item presents DIF in discrimination for all regions.

It is observed from Figure 9 that most of the items with DIF in difficulty are more difficult in Region S than in all other regions. The same happens in Regions NE and CW. In Region N, 58% of the items present positive DIF (less difficulty). From these, most of them are less difficult than in all other regions. Nevertheless, most of the other DIF items (with negative DIF) have greater absolute values that the ones with positive DIF and they are more difficulty in Region N than in the other regions.

Table 7 shows that items related to rational number representation using fractions seems to be harder for students from S ($E(\gamma^b_{42}|Y) = -0.16$). Items with DIF in NE are, on average, harder in this region than in SE ($E(\gamma^b_{03}|Y) = -0.10$). Numeric problems involving sum and subtraction seems to be easier in NE ($E(\gamma^b_{13}|Y) = 0.11$). The estimate of the regression coefficient related to area and/or perimeter calculation drawn in a square-lined mesh has a considerable estimated value ($E(\gamma^b_{33}|Y) = 0.15$) in Region NE, which indicates that this kind of item is also easier in Region NE.

Items related to use of decimal system in number decomposition ($E(\gamma^b_{25}|Y) = -0.20$) seems to

be harder for students from the Region N. Items related to rational number representation using fractions have a high value for the estimate of the regression coefficient ($E(\gamma_{45}^b|Y) = -0.51$), which means that this kind of item is harder for students from Region N than for students from the other regions of the country. Four out of the nine items related to this content are detected as having DIF in difficulty in Region N and these four items are among the five items with highest DIF parameter estimate. According to specialists, this difference is probably due to curricular differences related to this content between the teaching system in Region North and in the other regions.

The credibility intervals of some of the coefficients cited above contain 0, but these are not the maximum density intervals and zero is close to one of the interval limits. So, these coefficients may be considered significant.

Item 55 presents DIF in difficulty for all regions, this item is a subtraction operation that demands decompositions of a decimal order unit into units of a inferior decimal order. It is easier in South than in all other regions, more difficult in CW and N (DIF parameter is smaller in N, but there seems to be no significant difference between these two regions) than in SE and S, and it is harder in NE than in all other regions.

Figure 10 presents the estimate of the characteristic curve of item 55 in all Regions. Besides presenting DIF in difficulty in all regions, this item has DIF in discrimination in NE (more discriminant than in all other regions). Note that NE is the region where it is more difficult. This relation between difficulty and discrimination is very commonly found: the more difficult is an item, the more discriminant it is.

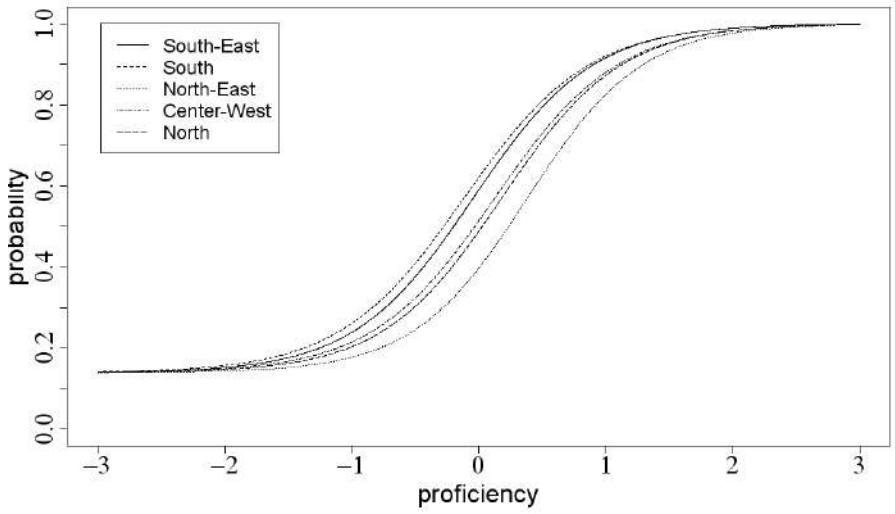

Figure 10: *Estimate of the characteristic curve of item 55 in all Regions.*

5.2 Two factors analysis

In this section, the real data set will be analysed considering two factors: the country's regions and the type of school. This last factor is composed by three groups: private school, state school and municipal school. The same four covariates used in the one factor analysis will be considered now. The analysis will be performed with the main effects model and with the

saturated model, that considers the interaction between the two factors. The models will be compared using the DIC and also analysing the parameters of the DIF regression structure and the parameters of the proficiencies' distribution in both models, as done in section 4.3.

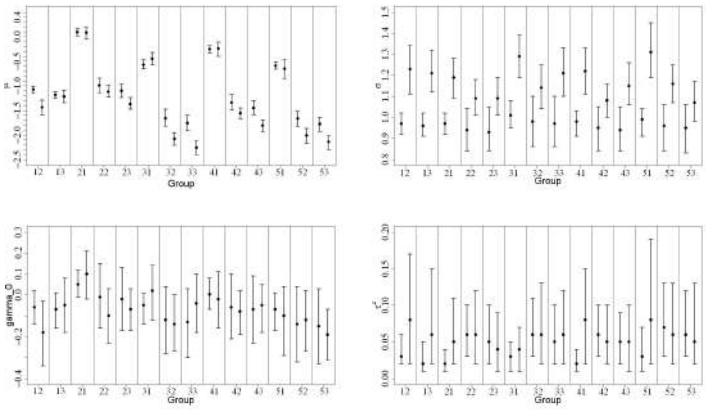

Figure 11: *Posterior mean and* 95% *credibility interval of the parameters* μ, σ, γ_0^a, τ_a^2 *in each group for both models. Left interval: multifactor model with no interaction, right interval: saturated model.*

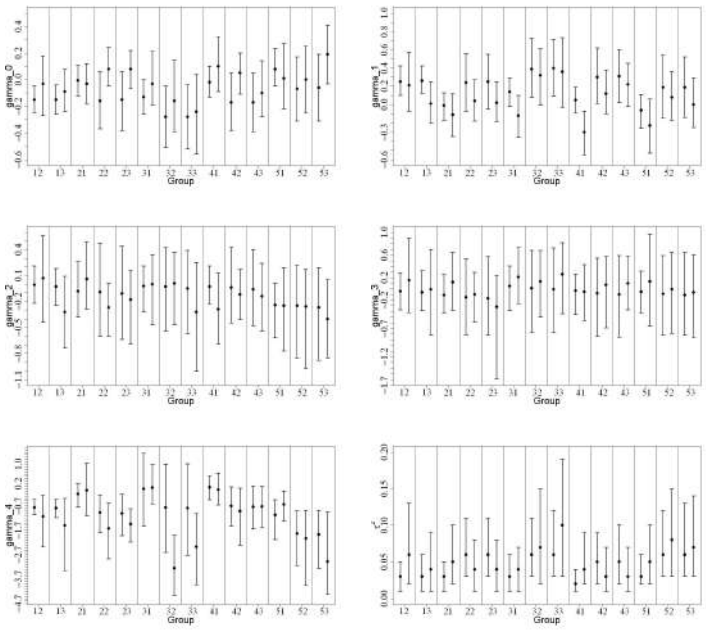

Figure 12: *Posterior mean and* 95% *credibility interval of the parameters* γ_0^b, γ_1^b, γ_2^b, γ_3^b, γ_4^b, τ_b^2 *in each group for both models. Left interval: multifactor model with no interaction, right interval: saturated model.*

In Figure 11, Group ij means group i in factor 1 and group j in factor 2. It can be noticed that the interaction between the factors in the mean of the proficiencies happens in two situations. The first one is that municipal school is a little better than state school in SE

	μ			σ		
	1FM	MM	SM	1FM	MM	SM
SE	0.00	-0.71	-0.83	1.00	1.06	1.27
S	-0.04	-0.71	-0.85	0.96	0.99	1.19
NE	-0.67	-1.37	-1.66	1.12	1.07	1.33
CW	-0.40	-1.11	-1.32	1.00	1.00	1.20
N	-0.78	-1.53	-1.83	0.99	0.95	1.15

Table 8: *Posterior mean and variance of the proficiencies in each region obtained with tree different models. 1FM is the one factor model, SM is Saturated Model and MM is Multifactor Model with no interactions.*

	μ		σ	
School	MM	SM	MM	SM
Private	-0.19	-0.18	0.96	1.15
State	-1.37	-1.66	0.90	1.07
Municipal	-1.46	-1.77	0.90	1.09

Table 9: *Posterior mean and standard-deviation of the proficiencies in each school obtained with two different models. SM is Saturated Model and MM is Multifactor Model with no interactions.*

Parameter	Post. mean	Cred. interval
Factor 1 - Region		
γ^b_{42}	-0.21	(-0.74 , 0.20)
γ^b_{03}	-0.13	(-0.26 , 0.00)
γ^b_{13}	0.14	(-0.02 , 0.30)
γ^b_{25}	-0.24	(-0.62 , 0.01)
γ^b_{45}	-1.07	(-2.09 , -0.46)
Factor 2 - School		
γ^b_{02}	-0.15	(-0.25 , -0.05)
γ^b_{12}	0.25	(0.10 , 0.43)
γ^b_{42}	-0.77	(-1.07 , -0.45)
γ^b_{03}	-0.15	(-0.25 , -0.04)
γ^b_{13}	0.26	(0.12 , 0.42)
γ^b_{43}	-0.79	(-1.18 , -0.43)

Table 10: *Posterior mean and* 95% *credibility interval of the parameters of the DIF regression that seem to be significantly different from zero. Estimates obtained by the model with no interaction.*

and is the opposite in the other regions. The second situation is that the difference between private and public (state and municipal) schools is a little different in each region. Concerning the variance of the proficiencies, the interaction is more complex, that is, the difference between the three types of schools is considerably different in each region.

Table 8 shows that the ordering of the means of the proficiencies in each region is the same in the three models. Notice that the values between the one factor model and the other two models can not be compared since the reference group is not the same. But it can be done between the last two models. It shows that the difference among the regions is larger in the saturated model. Concerning the variance os the proficiencies, the one factor model has a different ordering from the models that consider two factors and, comparing these last two, the estimates obtained with the saturated model are greater.

Table 9 confirms the information extracted from Figure 11; the difference between the mean of the proficiencies in private and public school is greater in the saturated model. Besides that, the variances of the proficiencies are greater in the saturated model and the schools are differently ordered.

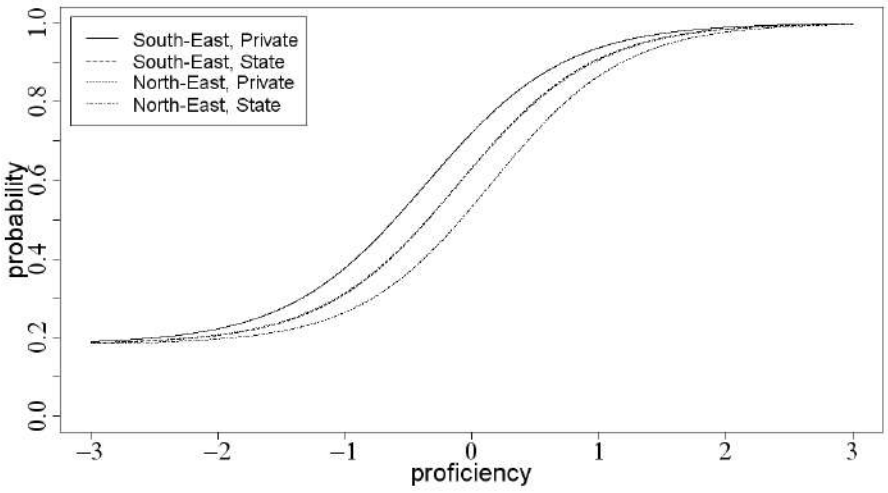

Figure 13: *Estimate of the characteristic curve of item 133 in four subgroups obtained by the model with no interaction.*

	$\bar{D}$	p_D	DIC
One Factor Model	413712	-222771	190941
Main Effects model	410648	-221366	189282
Saturated Model	410442	-212881	197561

Table 11: $\bar{D}$, p_D *and DIC in three Models.*

Note from Table 11 that the DIC for the main effects model is the smallest one, indicating that this model is preferred. The values of $\bar{D}$ and p_D show that the multifactor models fit the data much better. They also show that, naturally, the one factor model is the less complex one, followed by the main effects model and finally the saturated model.

The moderate difference between the values of $\bar{D}$ of the multifactor models indicate that a small interaction between the two factors may exist. Indeed, no intervals are disjoint for the DIF parameters, indicating that there is no or little interaction between the factors considering DIF. In the case of the distribution of the proficiencies, there seems to be a small but considerable interaction between the factors.

All the arguments above support the importance of considering the second factor in the DIF analysis, although the one factor analysis is not worthless. Moreover, the results obtained by comparing the analysis with the models with and without interaction indicate that another possible option would be to consider interaction only in the proficiencies.

The mean of the differences of the DIF probability parameters between the one factor model and the multifactor model with no interaction is 0.007. This is an evidence that the results of the one factor analysis are important and valid.

The high complexity of the saturated model is due to the interaction on the DIF parameters, which brings more parameters into consideration. The modelling of the proficiencies' distribution have a great impact in the analysis only in small sample situations. Then, unless the researcher is sure about the non interaction on the proficiencies' distribution, such interaction should be

considered in the analysis.

From the coefficients that were significant in the one factor analysis, only γ_{33}^b, which indicated that items related to area and/or perimeter calculation drawn in a square-lined mesh are easier in Region NE, is not significant in the two factor analysis with no interaction. This may have happened due to two reasons: the parameter value is not high in the one factor analysis, so the small difference on the estimates between the two models is caused the difference on the significance of the covariate; the other reason is that the inclusion of the new factor accommodated the influence of the covariate.

Concerning the type of school, the same coefficients are significant for state and municipal school, and their values are basically the same. They show that items with DIF are, on average, more difficult in public school. Besides that, they indicate that items concerning numeric problems involving sum and subtraction that present DIF are easier for students from public school while items related to rational number representation using fractions that present DIF are harder for these students.

Note that the coefficient related to rational number representation using fractions (γ_{45}^b) is even higher in the two factors analysis for Region N and this content is also much harder for students from public schools (γ_{42}^b and γ_{43}^b).

More generally, it can be concluded that students from private school have, on average, higher proficiencies than students from public schools. Moreover, state and municipal schools are very similar on both DIF and proficiency.

Figure 13 shows the estimated characteristic curve of item 133 in four subgroups. This item has DIF (in difficulty) in Region NE (harder than in SE) and in state school (harder than in private school), and these DIF's are practically the same. For that reason the curve is almost equal for state school in SE and private school in NE.

The results of the analysis show that it is possible to perform DIF detection, DIF explanation (in a mixed regression model) and multifactor DIF in a single step. This approach seems to be promising for practical applications since it is rarely known in advance which items have DIF. Additionally, multifactor DIF can be combined with the regression setup in an unified framework, as introduced in this paper, giving meaningful answers to educational issues that require them.

6 Conclusions

In this paper, two integrated Bayesian models for detection of items with differential functioning and explanation of the differential functioning by regression structures with covariates associated to the items were presented and studied. These models also consider the hypotheses of more than one grouping factors. The Bayesian inference procedures in both model were also presented along with important issues concerning prior distributions, the models' identifiability and the MCMC algorithm convergence.

The two models were compared in simulated studies, and it was concluded that Model 1 seems to be less efficient than Model 2 in situations with few DIF items. The DIC indicates that both models are equally good to fit the data and Model 2 is a little more complex than Model 1. Simulated studies showed the efficiency of the Bayesian methodology for parameters estimation and for solving identifiability problems of the likelihood function. A real analysis showed the viability of using the model in practical situations with satisfactory and intuitively consistent results.

This paper shows that many ideas that were explored separately for DIF analysis in IRT can be successfully accomplished in a simultaneous setup. This has the advantage of avoiding performance of data analysis in 2 or more steps. These procedures are bound to make ad-hoc assumptions and misleadingly assume with certainty when uncertainty is clearly present. Nevertheless, improvements in the model can still be achieved. Examples are: incorporation of correlation structures between the DIF magnitude and the item's difficulty, and among the DIF in the different items. These are possible directions for future works.

Appendix

Proof of Lemma 1

Suppose that a prior distribution $beta(\alpha, \beta)$ is assumed for π^h_{ig}. Then, the full conditional distribution $p(\pi^h_{ig_k}|\Psi_{-\pi^h_{ig_k}}, Y)$ of $\pi^h_{ig_k}$ is a $beta(\alpha, \beta+1)$ if $d^h_{ig_k} = 0$ and is a $beta(\alpha+1, \beta)$ if $d^h_{ig_k} \neq 0$, where $\Psi_{-\pi^h_{ig_k}}$ denotes Ψ without its component $\pi^h_{ig_k}$. These distributions have means $\dfrac{\alpha}{\alpha+\beta+1}$ and $\dfrac{\alpha+1}{\alpha+\beta+1}$, respectively.

Let $z^h_{ig_k}$ be the indicator of $d^h_{ig_k} = 0$. For a discrete $z^h_{ig_k}$, the posterior distribution of $\pi^h_{ig_k}$ can be written as

$$\begin{aligned} p(\pi^h_{ig_k}|Y) &= \sum_{z^h_{ig_k}} p(\pi^h_{ig_k}|Y, z^h_{ig_k}) p(z^h_{ig_k}|Y) \qquad (7) \\ &= p(\pi^h_{ig_k}|Y, z^h_{ig_k}=0) P(z^h_{ig_k}=0|Y) + p(\pi^h_{ig_k}|Y, z^h_{ig_k}=1) P(z^h_{ig_k}=1|Y). \end{aligned}$$

Taking expectations on both sides and setting $w^h_{ig_k} = P(z^h_{ig_k}=0|Y)$ gives

$$E(\pi^h_{ig_k}|Y) = E(\pi^h_{ig_k}|Y, z^h_{ig_k}=0) w^h_{ig_k} + E(\pi^h_{ig_k}|Y, z^h_{ig_k}=1)(1 - w^h_{ig_k}).$$

Since $0 < w^h_{ig_k} < 1$, $E(\pi^h_{ig_k}|Y)$ will be a weighted mean of the conditional expectations and thus will always be between them.

The proof is completed by noting that since $\pi^h_{ig_k}|\Psi_{-\pi^h_{ig_k}}, Y \sim \pi^h_{ig_k}|z^h_{ig}$, then, $\pi^h_{ig_k}|Y, z^h_{ig_k} \sim \pi^h_{ig_k}|z^h_{ig_k}$. Therefore, the posterior mean of $\pi^h_{ig_k}$ is a weighted mean of the two values the expectation of its full conditional distribution can assume, $\forall\, i = 1, \ldots, I$, $\forall\, k = 1, \ldots, K$, $\forall\, g = 2, \ldots, G$ and $\forall\, h = a, b$.

References

Agresti, A. (2002). *Categorical Data Analysis*. (2nd. edition). New York: Wiley.

Albert, J. H. (1992). Bayesian estimation of normal ogive item responses curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251-269.

Béguin, A. A. & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541-562.

Berberoglu, G. (1995). Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation, 21*, 439-456.

Birnbaum, S. (1968). Some latent traits models and their use in inferring an examinee's ability. In Lord, F. & Novick, M. (Eds.), *Statistical Theories of Mental Test Scores*, 397-472. Reading, MA: Addison Wesley.

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

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

Fox, J. P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Gamerman, D. & Lopes, H. L. (2006). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference* (2rd ed.). New York: Taylor & Francis.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of American Statistical Association, 85*, 398-409.

Gierl, M. J., Bisanz, J., Bisanz, G. & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40*, 281-306.

Gonçalves, F. B. (2006). Bayesian Analysis of the Item Response Theory: a Generalized Approach. Unpublished *M.Sc. dissertation*, IM-UFRJ (in Portuguese).

Hanson, B. A. (1998). Uniform DIF and DIF defined by Differences in Item Response Functions. *Journal of Educational and Behavioral Education, 23*, 244-253.

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

Janssen, R., Schepers, J. & Peres, D. (2004). Models with item and item group predictors. In Boeck, P. D. & Wilson, M. (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.*, 189-212. New York: Springer.

Jeffreys, H. (1939) *Theory of Probability.* 1a. ed., Clarendon Press, Oxford.

Liseo, B. & Loperfido N. (2006). Default Bayesian analysis of the skew-normal distribution. *Journal of Statistical Planning and Inference, 136, 2*, 373-389.

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*, 171-196. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational Behavioral Statistics, 31*, 63-79.

Migon, H. S. & Gamerman, D. (1999). *Statistical Inference: An Integrated Approach.* Arnold, London.

O'Neil, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*, 255-276. Hillsdale, NJ: Lawrence Erlbaum Associates.

Patz, R. J. & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J. & Junker, B. W. (1999b). Applications and Extensions of MCMC in IRT. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Rogers, H. J. & Swaminathan, H. (2000). *Identification of factors that contribute to DIF: A hierarchical modeling approach.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*, 85-100. New Yorg: Springer Verlag.

Schmitt, A. P. & Bleistein, C. A. (1987). *Factors affecting differential item functioning for black examinees on scholastic aptitude test analogy items* (ETS RR-87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Holland, P. W. & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*, 281-316. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sinharay, S., Dorans, N. J., Grant, M. C., Blew, E. O. & Knorr, C. M. (2006). Using past data to enhance small-sample DIF estimation: A Bayesian approach. Technical report ETS RR-06-09. Princeton, NJ: Educational Testing Service.

Soares, T. M., Gonçalves, F. B. & Gamerman, D. (2006). An integrated Bayesian model for DIF analysis. Technical report 197, LES-UFRJ (submitted).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64,* 583–640.

Steel, M. F. J. & Fernández, C. (1999). Multivariate Student-t regression models: pitfalls and inference. *Biometrika, 86*, 153-168.

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

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

Thissen, D., 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*, 67-114. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W.-C. (2000). Simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online, 5*, 57-76.

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

Wang, W.-C. & Yeh, Y-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 1-20.

West, M., Lucas, J., Carvalho, C., Wang, Q., Bild, A. & Nevins, J.R. (2006) Sparse statistical modelling in gene expression genomics. In K. A. Do, P. Müller & M. Vannucci (Eds.), *Bayesian Inference for Gene Expression and Proteomics*, 155-176. New York: Cambridge University Press.

Zwick, R. & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*, 57-76.

Zwick, R., Thayer, D. T & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

Zwick, R., Thayer, D. T & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*, 225-247.