Bayesian analysis in Item Response Theory applied in a large-scale educational assessment

Dani Gamerman, Tufi M. Soares and Flávio B. Gonçalves

Instituto de Matemática, Universidade Federal do Rio de Janeiro, Brazil
Centro de Políticas Públicas e Avaliação da Educação, Departamento de Estatística,
Universidade Federal de Juiz de Fora, Brazil
Department of Statistics, University of Warwick, UK

Abstract

This chapter provides an analysis of the outcomes of the international PISA test, concentrating on the Mathematics test for the English speaking countries. Data inspection suggest the need for departure from standard Item Response Theory (IRT) models to account for differential item functioning (DIF). An integrated model to incorporate detection, quantification and explanation of DIF is applied. Prior information plays a crucial role in this setting with quantification of the knowledge accumulated by research in the area. It also helps model identification in this highly parameterised setting. Results of the analysis highlight differences between the countries and suggest possible explanation for it.

1 Introduction

The OECD's[1] Programme for International Student Assessment (PISA) collects information, every three years, about 15-year-old students in several countries around the world. It examines how well students are prepared to meet the challenges of the future, rather than how well they master particular curricula. The data collected in PISA surveys contains valuable information for researchers, policy makers, educators, parents and students. It allows countries to assess how

[1]OECD - Organisation for Economic Co-Operation and Development

well their 15-year-old students are prepared for life in a large context and to compare themselves to other countries. As it is mentioned in PISA 2003 Technical Report[2], it is now recognised that the future economic and social well-being of countries is closely linked to the knowledge and skills of their population.

The first PISA survey was in 2000 and 43 countries participated in the assessment. The other ones were in 2003 with 41 countries and in 2006 with 57 countries. In 2009, 67 countries will participate. At least one country from each continent is present in the surveys, and the majority is from Europe. A list with the countries participating in 2003 is presented in Appendix A.

PISA is applied in a large number of countries. For this reason, there are many differences among the students participating, from cultural to curricular ones. These differences may have influence in characteristics of some items in each country. For example, if an item is strongly related to some cultural aspect of an specific country, it is expected that this item would be easier for students from this country than for students from another one where such cultural aspect is not presented. If this difference is not taken into account, such item is not good to assess the abilities of students from these two countries. Such phenomenon is called Differential Item Functioning (DIF) and has been heavily studied in psychometric literature. In general, one group is fixed as the *reference group* and the other ones as the *focal groups*, and the item functioning in the latter ones is compared to the item functioning in the reference group.

DIF Analysis can be separated in two different stages. The first one is the detection of items with DIF and the second one is the quantification and explanation of the DIF detected in the first stage. Normally, the latter one is based on regression structures to identify features of the items that present differential functioning. Usually, the first stage is divided in two sub-stages: the estimation of the students abilities or a matching of the students using groups, where students in the same group are assumed to have the same ability; and the DIF detection using these estimated abilities. Both sub-stages are linked because, in general, it is impossible to generate abilities which are totally free of DIF. Therefore, the most natural approach would be to treat them together.

However, along with the integrated modelling of the DIF, comes the problem of model identifiability, which reflects the conceptual difficulty of the problem. For this reason, additional hypotheses on the DIF parameters of the model are assumed in order to guarantee identifiability.

[2]PISA 2003 Technical Report. http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf

Nevertheless, such hypotheses may be too restrictive.

Situations like this, where, besides the information from the data, initial hypotheses on the parameters are needed, are very suitable for Bayesian approaches. The use of appropriate prior distributions, elicited in a Bayesian framework and based on previous knowledge on the items' DIF, can be less restrictive than usual hypotheses, but still enough to guarantee identifiability.

Preparation of a large test like PISA somehow screens the DIF to avoid such items to compose the test. It is then not expected that items with large DIF, that have a great influence on the proficiencies, are found. On the other hand, it is impossible to eliminate all DIF in the tests, specially in such large scale tests.

This chapter presents a DIF analysis of the Mathematics test in PISA 2003, for the English speaking countries, using an integrated Bayesian model. Such model allows items to present DIF without affecting the quality of the estimated proficiencies. Moreover, the DIF analysis may be useful to detect educational differences among the countries.

2 Programme for International Student Assessment (PISA)

2.1 What does PISA assess?

PISA is based on a dynamic learning model that takes into account the fact that knowledge and ability must be continuously acquired to have a successful adaptation to a constant changing world. Therefore, PISA aims to assess how much of these knowledge and ability, essential for an effective participation in society, is acquired by students near the end of compulsory education. Differently from previous international assessments (IEA, TIMMS, OREALC, etc)[4], PISA does not focus only in curricular contents. It also emphasizes the knowledge required in modern life. Besides that, data about the student's study habits and their motivation and preferences for different types of leaning situation is also collected. The word Literacy is used to show the amplitude of the knowledge, abilities and competencies being assessed, and it covers:

- Contents or structures of the knowledge the students have to acquire in each domain;

[3]The information presented in this section is based on the PISA 2003 Technical Report. http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf

[4]IEA- International Association for Evaluation of Educational Achievement, TIMSS - Trends in International Mathematics and Science Study, OREALC/UNESCO - Oficina Regional de Educación para América Latina e Caribe.

- Processes to be executed;
- The context in which these knowledge and abilities are applied.

In Mathematics, for example, the competence is assessed in items that cover from basic operations to high order abilities involving reasoning and mathematical discoveries. The Literacy in Maths is assessed in three dimensions:

- The content of Mathematics - firstly defined in terms of wide mathematical concepts and then related to branches of the discipline;
- The process of Mathematics - general mathematical competence in use of mathematical language, selection of models and procedures, and abilities to solve problems;
- Situations where Mathematics is used - varying from specific contexts to the ones related to wider scientific and public issues.

In PISA 2003, four subject domains were tested, with mathematics as the major domain, and reading, science and problem solving as minor domains. It was a paper-and-pencil test. The tests were composed by three item formats: multiple-choice response, closed-constructed response and open-constructed response. Multiple choice items were either standard multiple choice, where the students had to choose the best answer out of a limited number of options (usually four), or complex multiple choice where the students had to choose one of several possible responses (true/false, correct/incorrect, etc.) for each of the statements presented. In closed-constructed response items, a wider range of responses was possible. Open-constructed response items required more extensive writing, or showing a calculation, and frequently included some explanation or justification. Pencils, erasers, rulers, and in some cases calculators, were provided. The decision on providing or not calculators was made by the National centres and was based on standard national practice. No items in the pool required a calculator but, in some items, its use could facilitate computation. Since the model used here to analyse the data is for standard multiple-choice items, the items which do not have this format were converted into dichotomic items. Most of them were already dichotomised by PISA. The five items remaining were dichotomised by putting 1 for full credit and 0 for partial or no credit.

The complex probabilistic sample, involving stratification and clustering, was obtained by randomly selecting the schools. Inside the selected schools, the students who were from 15 years

and 3 months to 16 years and 2 months old were also randomly chosen. The students had to be in the 7^{th} or 8^{th} year of Basic School or in High School. The sample was stratified by the schools' location (urban or rural). Besides, information on physical infra-structure of the school, geographic region, type of school (private or public) and number of enroled students were also used as variables in the implicit stratification.

2.2 The structure of PISA 2003 Maths test

More than a quarter of a million students, representing almost 30 million 15-year-old students enrolled in the schools of the 41 participating countries, were assessed in 2003.

The Maths test was composed by 85 items, from which 20 were retained from PISA 2000 for linking purposes. These items were selected from a previous set of 217 items by expert groups based on several criteria. Some important characteristics of the 85 select items are presented in Tables 1, 2 and 3. The characteristics presented in these tables will be considered in the DIF analysis presented here.

	Competence cluster			
Item format	Reproduction	Connections	Reflection	Total
Multiple-choice response	7	14	7	28
Closed-constructed response	7	4	2	13
Open-constructed response	12	22	10	44
Total	26	40	19	85

Table 1: *Mathematics main study items (item format by competency cluster).*

	Competence cluster			
Content category	Reproduction	Connections	Reflection	Total
Space and shape	5	12	3	20
Quantity	9	11	3	23
Change and relationships	7	8	7	22
Uncertainty	5	9	6	20
Total	26(31%)	40(47%)	19(22%)	85

Table 2: *Mathematics main study items (content category by competency cluster).*

3 Differential Item Functioning (DIF)

The Differential Item Functioning (DIF) is the phenomenon where the characteristics of an item differ among groups of individuals. Such characteristics may include discrimination,

Content category	Item format			
	Multiple-choice response	Closed-constructed response	Open-constructed response	Total
Space and shape	5	12	3	20
Quantity	9	11	3	23
Change and relationships	7	8	7	22
Uncertainty	5	9	6	20
Total	26(31%)	40(47%)	19(22%)	85

Table 3: *Mathematics main study items (content category by item format).*

difficulty, etc. For example, students with same level of knowledge and ability but different cultural background may have different probabilities of correctly answer an item with DIF.

The concern about DIF arose with the desire of creating tests' items that were not affected by cultural and ethnic characteristics of the individuals taking the tests (cf. Cole, 1993). If one does not consider DIF when it exists, one may be led to wrong conclusions regarding, specially, the students knowledge and abilities.

Studies conducted by the Educational Testing Service (ETS) (see Stricker and Emmerich, 1999), in the U.S.A., indicate that DIF may exist due to three factors in a large-scale assessment context: the familiarity to the item's content, which can be associated to the exposure to the theme or to a cultural factor; the personal interest in the content; and a negative emotional reaction caused by the item's content.

It seems reasonable to think that items with DIF should be avoided since these would favor some group(s) of students, besides having some technical and statistical implications and other ethical issues. Nevertheless, DIF items may be very useful to study social and cultural differences that are not easily noticed. In particular, in educational assessment, DIF items can help detecting contents that are treated differently among the groups and may point out in which groups the instruction of such contents should change.

Explaining the DIF is a very hard task. Besides that, the pedagogical and technical structure of a large-scale assessment like PISA aims to create high quality items that do not in general present differential functioning. However, it is known that the characteristics of a country and its level of economic development have great influence in the social and cultural life of its population, which reflects in education. For this reason, different countries are expected to organise the curriculums in different ways, some countries may give more importance to some themes and explore more some contents. The challenge in explaining the DIF in PISA here is

to notice patterns in the items that present DIF. In order to do so, it is important to have a large number of items that are quite different.

4 Bayesian Model for DIF Analysis

The most common approach to undertake a statistical analysis of educational assessment data is by using Item Response Theory (IRT). It is a psychometric theory extensively used in education assessment and cognitive psychology to analise data arising from answers given to items belonging to a test, questionnaire, etc., and it is very useful to produce scales.

The IRT arose, formally, in the work of Lord (1952) and Rasch (1960). The basic idea of the theory is to apply models, generally parametric, where the parameters represent important features of the items and of the subjects answering the items. Some common item parameters are discrimination, difficulty and guessing. The subject parameters are individual characteristics, for example, a type of ability or some other latent trait possibly associated to his/her psychological condition.

An item can be dichotomous, if the answer is either correct or not; polytomous, if the answer can be classified in more than two categories; and also continuous, if the answer is classified in a continuous scale. There are IRT models for all those situations, but only models for dichotomic items will be considered here.

The increasing use of IRT in educational assessment and the concerns about Differential Item Functioning are leading researchers to propose new models that take DIF into account. In this context, Soares et al. (2008) propose a new IRT Bayesian model that is a generalisation of the three parameters logistic model. The proposed model incorporates the detection, quantification and explanation of DIF in an integrated approach.

Typically, in educational assessment, a test is formed by I items, but a 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 $i \in I(j) \subset \{1, \ldots, I\}$. In the case where i is a dichotomic item, $Y_{ij} = 1$ if the answer is correct and $Y_{ij} = 0$ if the answer is wrong.

Define $P(Y_{ij} = 1|\theta_j, a_i, b_i) = \pi_{ij}$, and consider $logit(\pi_{ij}) = \ln\left(\dfrac{\pi_{ij}}{1-\pi_{ij}}\right)$. If $logit(\pi_{ij}) = \Delta_{ij}$, then $\pi_{ij} = logit^{-1}(\Delta_{ij}) = \dfrac{1}{1+e^{-\Delta_{ij}}}$. One of the most used models in IRT is the three parameters logistic model proposed by Birnbaum (1968), which models the probability of $Y_{ij} = 1$

the following way

$$\pi_{ij} = c_i + (1 - c_i) logit^{-1}(\Delta_{ij}) \tag{1}$$
$$\Delta_{ij} = Da_i(\theta_j - b_i)$$

where a_i, b_i and c_i are the discrimination, difficulty and guessing of item i, respectively. θ_j is the ability, proficiency or knowledge of student j, and D is a scale factor designed to approximate the logistic link to the normal one (and set here to 1.7).

A good way to interpret the three parameters logistic model is by analysing the item's characteristic curve generated by the model, presented in Figure 1. Note from the model that the greater is a_i, the greater will be the slope of the curve. This means that, the more discriminant an item is, the larger is the difference in the probability of correct answer for students with different proficiencies, that is, the larger is the capability of the item to discriminate the students. Moreover, the maximum slope is attained at $\theta = b$, as it can be seen in the figure.

The difficulty parameter b_i interferes in the height of the curve. If the value of b_i is increased, the curve moves down and the probability of correct answer is reduced (the item becomes harder). Alternatively, if b_i is decreased, the curve moves up and the probability of correct answer is increased (the item becomes easier).

Furthermore, note that $\lim_{\theta_j \to -\infty} Pr(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i$. Hence, c_i is the minimum probability that a student has to correctly answer item i. c_i is then called guessing parameter because, since it is a multiple choice item, there is always a probability that the student answers it correctly by guessing. Lord (1952) noticed that, in general, the percentage of correct answers for an item, in very low levels of proficiency, was smaller than the inverse of the number of options in the item. Experts in the area report that they have observed varied behaviors for those percentages.

In general, there can be different types of DIF (see Hanson (1998) for a wider characterisation). For the three parameters model, the types of DIF can be characterised according to the difficulty, discrimination and guessing. The model proposed here does not consider the possibility of DIF in the guessing parameter. Although it is possible, the applicability of this case is substantially limited by the known difficulties in the estimation of this parameter and by practical restrictions.

Suppose that the students are grouped in G groups; the Integrated Bayesian DIF Model

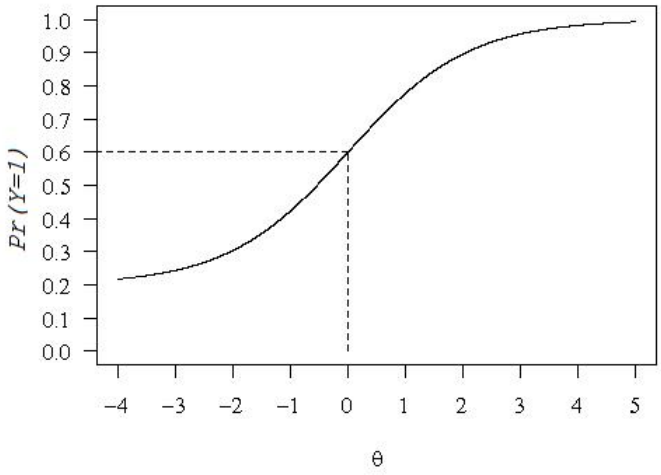

Figure 1: *Item's characteristic curve for the three parameters logistic model with* $a = 0.5$, $b = 0$ *e* $c = 0.2$.

used in this chapter associates the student's answer to his/her ability via (1) with

$$\Delta_{ij} = Da_{ig}(\theta_j - b_{ig}),$$

where a_{ig} is the discrimination parameter for item i and group g, b_{ig} is the difficulty parameter for item i and group g, c_i ($\in [0,1]$) is the guessing parameter for item i, for $i = 1, \ldots, I$, $j = 1, \ldots, J$ e $g = 1, \ldots, G$.

Since the item difficulty is a location parameter, it is natural to think about its DIF in the additive form and thus it is set as $b_{ig} = b_i - d^b_{ig}$. Analogously, since the the discrimination is a scale parameter, it is natural to think about its DIF in the multiplicative form and thus it is set as $a_{ig} = e^{d^a_{ig}} a_i$ (> 0). Thus, d^b_{ig} (with $d^b_{i1} = 0$) represents the DIF related to the difficulty of the item in each group and $e^{d^a_{ig}}$ (with $d^a_{i1} = 0$) represents the DIF related to the discrimination of the item in each group. Use of the exponential term in the discrimination places the DIF parameter over the line and combines naturally with a normal regression equation setting. Alternatively, the DIF parameter can be specified directly without the exponential form. This leads to a log-normal regression model. The two forms are equivalent but the former was preferred here.

It is assumed, *a priori*, that $\theta_j | \lambda_{g(j)} \sim N(\mu_{g(j)}, \sigma^2_{g(j)})$, where $g(j)$ means the group which the student j belongs to and $\lambda_g = (\mu_{g(j)}, \sigma^2_{g(j)})$, $j = 1, \ldots, J$. It is admitted that $\lambda_1 = (\mu_1 , \sigma^2_1) = (0, 1)$ to guarantee the identification of the model. On the other hand, $\lambda_g = (\mu_g, \sigma^2_g)$, $g =$

$2, \ldots, G$, is unknown and must be estimated along with the other parameters.

The model is completed with specifications of the prior distributions for the parameters. Let N be the Normal distribution, LN the Log-Normal distribution, Be the Beta distribution and IG the Inverse-Gamma distribution, so, the prior distributions assumed for the structural parameters are:

$$a_i \sim LN\left(\mu_{a_i}, \sigma^2_{a_i}\right) \ , \ b_i \sim N\left(\mu_{b_i}, \sigma^2_{b_i}\right) \quad \text{and} \quad c_i \sim Be\left(\alpha_{c_i}, \beta_{c_i}\right) \ , \ \text{for } i = 1, \ldots, I.$$

The prior distributions for the parameters of the abilities' distributions are:

$$\mu_g \sim N\left(\mu_{0g}, \sigma^2_{0g}\right) \quad \text{and} \quad \sigma^2_g \sim IG\left(\alpha_g, \beta_g\right) \qquad \forall\, g = 2, \ldots, G$$

The set of anchor items (items for which $d^a_{ig} = d^b_{ig} = 0, \ \forall \ g = 1, \ldots, G$) is represented by $I_A \subset \{1, \ldots, I\}$. The set of items for which the parameters vary between the groups is represented by $I_{dif} = \{1, \ldots, I\} - I_A$. Moreover, $I^a_{dif} \subset I_{dif}$ is the set of items with DIF in the discrimination and $I^b_{dif} \subset I_{dif}$ is the set corresponding to the DIF in the difficulty. Naturally $I_{dif} = I^a_{dif} \cup I^b_{dif}$. Notice that if an item belongs to I_{dif}, it does not necessarily mean that this item has DIF in the usual meaning of the term. It means that it is not an anchor item and it can potentially have DIF. Besides that, it can be used as an admissible information for the explanatory structure imposed to the DIF, which cannot be performed with the anchor items.

Let Z^h_{ig}, $h = a, b$, be the DIF indicator variable of item i in group g, for parameter h. Therefore, $Z^h_{ig} = 1$ if parameter h of item i has DIF in group g, and $Z^h_{ig} = 0$, otherwise. Two possibilities may be considered: one where Z^h_{ig} is known, which means that the anchor items are known *a priori* and the DIF analysis considers all the other items; and another one where Z^h_{ig} is unknown and must be identified. In other words, it is not known *a priori* if the item has or not DIF. The latter one will be used in the analysis of PISA.

Finally, a regression structure is considered for d^h_{ig} in the DIF explanation as follows

$$d^h_{ig} = \gamma^h_{0g} + \sum_{k=1}^{K^h} \gamma^h_{kg} W^h_{ik} + \eta^h_{ig} \ , \ \text{ if } Z^h_{ig} = 1. \tag{2}$$

γ^h_{kg} are the fixed parameters of the DIF model, W^h_{ik} are the explanatory variables associated to the items and η^h_{ig} is the item specific random factor for each group. It is also assumed for modelling simplification that $\eta^h_g \sim N(0, T_g)$, where $T_g = (\tau^h_g)^2 I$, $\forall\, g = 2, \ldots, G$.

The regression structure is imposed for all items but the anchor ones. Consider $W^h_i = (1, W^h_{i1}, \ldots, W^h_{iK^h})$ and $\gamma^h_g = (\gamma^h_{0g}, \ldots, \gamma^h_{K^h g})'$. When $Z^h_{ig} = 1$, the conditional distribution of d^h_{ig}

is given by $(d_{ig}^h|\gamma_g^h, W_i^h, (\tau_g^h)^2) \sim N\left(W_i^h\gamma_g^h, (\tau_g^h)^2\right)$. When $Z_{ig}^h = 0$, d_{ig}^h will be assumed to have a reduced variance normal distribution $(d_{ig}^h|(\tau_g^h)^2, Z_{ig}^h = 0) \sim N\left(0, s^2(\tau_g^h)^2\right)$, where s^2 is chosen to be small enough to ensure that d_{ig}^h is tightly concentrated around (but not degenerated at) 0. This strategy was proposed for variable selection in a regression model by George and McCulloch (1993).

The distribution of $d_{ig}^h|\gamma_g^h, W_i^h, Z_{ig}^h, (\tau_g^h)^2$ can then be written as follows

$$\left(d_{ig}^h|\gamma_g^h, W_i^h, Z_{ig}^h, (\tau_g^h)^2\right) \sim N\left(\left(W_i^h\gamma_g^h\right) Z_{ig}^h\ ,\ [s^2]^{1-Z_{ig}^h}(\tau_g^h)^2\right).$$

Suitable prior distributions are $\gamma_g^h \sim N\left(\gamma_0^h, S_0^h\right)$, $(\tau_g^h)^2 \sim IG\left(\alpha_g^h, \beta_g^h\right)$ and $Z_{ig}^h \sim Ber(\pi_{ig}^h)$, where Ber is the Bernoulli distribution.

The model proposed here is very general. Apart from the usual parameters, it also has a DIF indicator variable Z_{ig}^h which can either be estimated along with all the other parameters of the model or be fixed *a priori*. The items for which Z_{ig}^h is not fixed at 0 include additional variables to the model. These variables are used for the DIF explanation and constitute a regression model which may or not have covariates.

Prior distributions play a very important role in a Bayesian model. In this particular model, they are very important in the selection of the anchor items, as it will be seen. If one wishes to set an item as an anchor one, it is sufficient to make $\pi_{ig}^h = 0,\ \forall\ g = 2, \ldots, G,\ \forall\ h = a, b$. Naturally, π_{ig}^h may be set as zero for some but not all groups. In the same way, if one wishes to include an item in the DIF analysis, independent from the DIF's magnitude, it is sufficient to make $\pi_{ig}^h = 1,\ \forall\ g = 2, \ldots, G$, for some h. However, the most interesting use of this prior distributions is to consider previous information and beliefs about the items' functioning to identify parameters more precisely and effectively. The estimation of the parameter is done by using MCMC methods to obtain a sample from the joint posterior distribution of all the parameters. Details on these methods are presented in Appendix B.

The level of generality introduced by this model aims to represent the complexity of the problem. However, along with the integrated modelling of the DIF, comes the problem of model identifiability. Such problems are described and studied in Soares et al. (2008). The authors show that identifiability is achieved by either imposing informative prior distributions for some Z_{ig}^h parameters or fixing some items not to have DIF. Nevertheless, in many cases, the model is identifiable without any of these two actions because of the informative priors attributed to the other parameters.

5 DIF Analysis of PISA 2003

The Bayesian model presented in Section 4 is now used to perform a DIF analysis of the PISA 2003 Maths test in English speaking countries. The database obtained had 84 of the 85 items of the test. The estimates of such missing item's parameters are not presented in the PISA Technical Report 2003 either. Great Britain is defined as the reference group (group 1). Table 4 shows the other groups.

Country	**Group**
Great Britain	1
Canada	2
Australia	3
Ireland	4
U.S.A.	5
New Zealand	6

Table 4: *E*nglish speaking countries used in the DIF analysis.

5.1 Prior distribution

The following prior distributions are used:

$$a_i \sim LN(0,2)\ ,\ b_i \sim N(0,1)\ ,\ c_i \sim Be(5,17)\ ,\ \mu_g \sim N(0,1)\ ,\ \sigma_g^2 \sim IG(0.1,0.1)\ ,\ \gamma_g^h \sim N(0,I)\ ,\ (\tau_g^h)^2 \sim IG(0.5,0.5) \text{ and } Z_{ig}^h \sim Ber(0.5)\ ,\ \text{for } i=1,\ldots,84,\ g=2,\ldots,6,\ h=a,b.$$

The value chosen for parameter s^2 is $1/200000$.

The prior distributions of the item parameters are chosen according to what is expected, in general, when the proficiencies follow a standard Normal distribution, which is the case of the reference group. The discrimination parameters are expected to vary mostly between 0 and 3, and the $LN(0,2)$ has probability 0.70 of being smaller than 3. Most of the difficulty parameters are expected to be between -2 and 2, and a few of them to have absolute value greater than 2. The standard Normal distribution is then a suitable prior to describe this behavior. The guessing parameter is expected to be low (less than 0.3), since many items were not multiple choice ones, and then should have a low probability of correct guessing. The $Be(5,17)$ distribution gives 0.8 probability to values smaller than 0.3.

The mean of the proficiencies in the focal groups are not expected to be very far (more than 1 unit) from the mean of the reference group. For this reason, a standard Normal distribution is

used for these parameters. For the variance of the proficiencies, a large variance prior distribution was preferred.

The DIF parameters are expected to have absolute value smaller than 0.5 and, for a few items, between 0.5 and 1. For this reason, the effect of a binary covariate is also expected to be around these values. Therefore, a standard Normal distribution is used for the coefficients in the regression analysis. For the variance of the regression error, a prior distribution with large variance is adopted.

Since there is no prior information about how likely each item is to present DIF, a symmetric prior distribution Ber(0.5) is used for the DIF indicator variables Z_{ig}^h.

5.2 DIF detection

Figures 2, 3 and 4 show some results for the item parameters. For the discrimination and difficulty parameters, the estimates in the reference group (GBR) and in the groups where the item is likely to have DIF *a posteriori* are presented.

A nice feature of the model is that it incorporates the uncertainty about DIF in the estimates. It means that the model does not have to "choose" if an item has or not DIF. It outputs a posterior distribution on the parameters that describes the uncertainty about that, particularly, for the DIF parameters' posterior distribution. It is up to the researcher to analyse this uncertainty and draw conclusions from it.

If an item i is likely to have DIF *a posteriori* in group g and parameter h, say $P(Z_{ig}^h = 1|Y) > 0.5)$, the posterior distribution of d_{ig}^h will be bimodal with one mode in 0. In order words, this posterior distribution is a mixture of a distribution with mean 0 and a very small variance ($\approx s_i^2$) and another distribution. The former one is the distribution of $(d_{ig}^h|Z_{ig}^h = 0, Y)$, which is approximately $N(0, s_i^2)$ and the latter is the distribution of $(d_{ig}^h|Z_{ig}^h = 1, Y)$. The more likely the item is to have DIF *a posteriori*, the further away from 0 this second distribution is.

The decision on an item presenting or not DIF is based on the posterior distribution of the respective Z_{ig}^h. If one item is assumed to have DIF, the estimation of this DIF is made by the distribution of $(d_{ig}^h|Z_{ig}^h = 1, Y)$. For this reason, the estimates of the item parameters in the focal groups, when the item has a posterior probability of DIF greater than 0.5, presented in Figures 2 and 3, is the posterior mean of $(d_{ig}^h|Z_{ig}^h = 1, Y)$.

Concerning difficulty, twenty items have posterior probability of DIF smaller than 0.5 in all groups, that is, they are more likely not to present any DIF. No item has this DIF probability

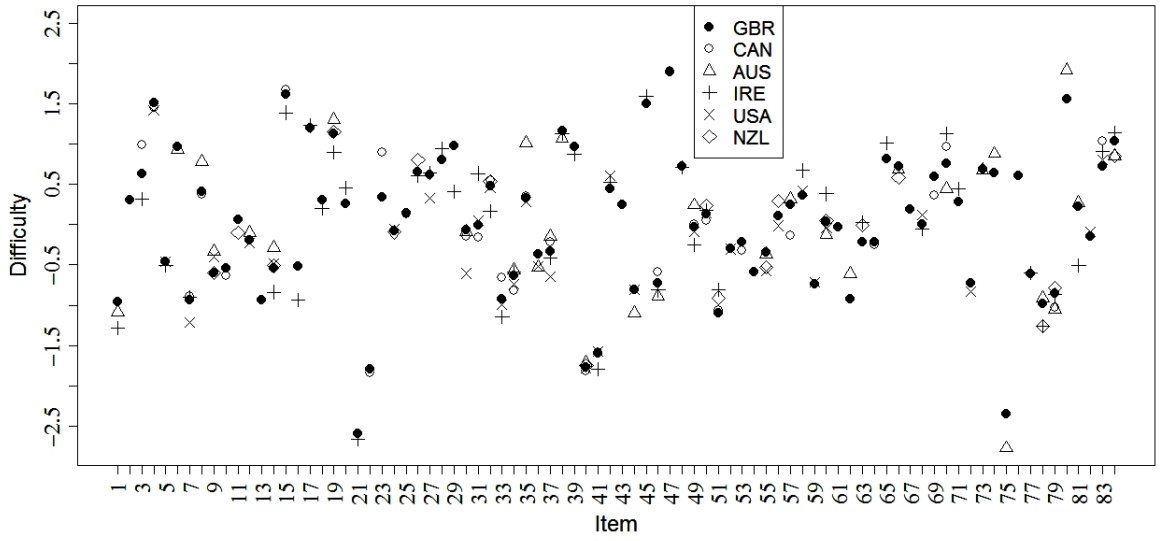

Figure 2: *Estimates of difficulty parameters in the reference group and in the groups where they had a posterior probability of DIF greater than 0.5.*

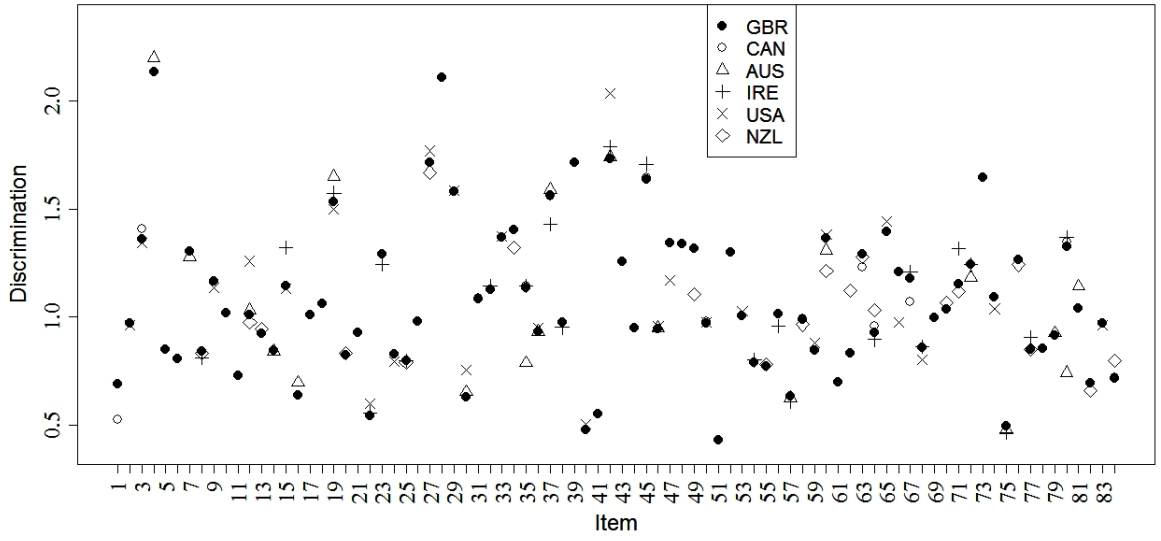

Figure 3: *Estimates of discrimination parameters in the reference group and in the groups where they had a posterior probability of DIF greater than 0.5.*

greater than 0.5 for more than three countries. Forty five DIF parameters have absolute value greater than 0.3, which is a considerable magnitude for DIF, and nine have this value greater than 0.5, which is a large value for DIF.

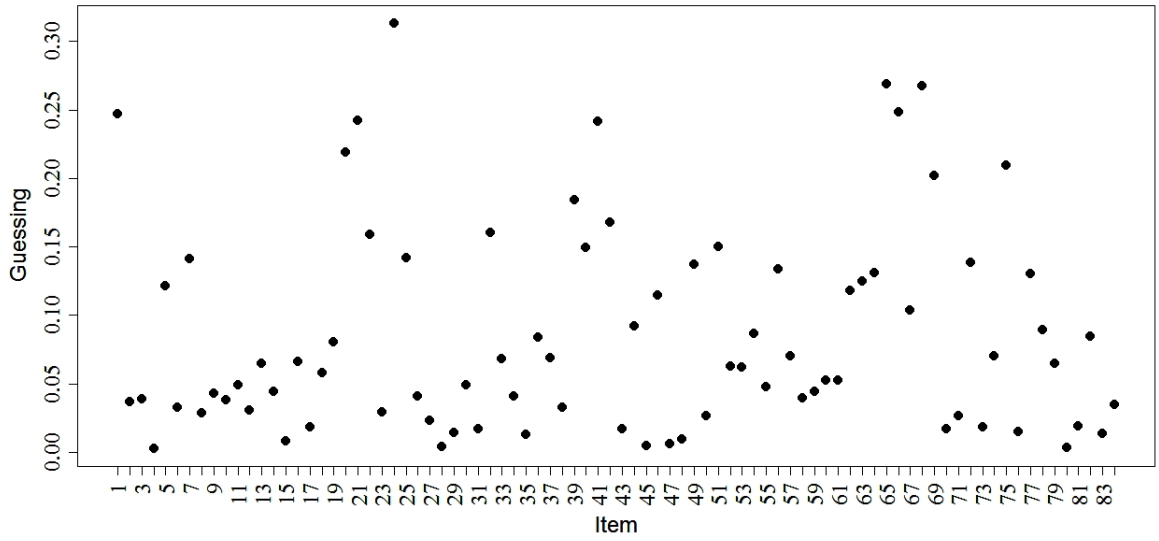

Figure 4: *Estimates of guessing parameter.*

Considering discrimination, seven items have posterior probability of DIF smaller than 0.5 in all groups. Nine items have this DIF probability greater than 0.5 for four countries and no item has this probability greater than 0.5 for all the countries. On the other hand, only eight DIF parameters are greater than 0.3 (which increases the discrimination in 35% if it is positive and decreases in 26% if it is negative). Only one DIF parameters is greater than 0.5, this value increases the discrimination in 65% if it is positive and decreases in 40% if it is negative.

Finally, note that most of the guessing parameters estimates are less than 0.15. This was expected since most of the items are not multiple choice ones. It would be reasonable to fix $c = 0$ for these items. However, it was not done, in order to make possible the comparison of the results obtained with the integrated Bayesian model to the ones from Bilog-mg software (see Thissen, 2001) without DIF. Bilog-mg uses the same scale for the proficiencies and fits a three parameters logistic model.

5.3 DIF explanation

Four covariates were chosen to explain the DIF in the English speaking countries. The first one is the Content category shown in Table 2. It is a categorical variable with four categories: Space and shape; Quantity; Change and relationships; and Uncertainty. Three dummy variables are used to introduce this covariate to the regression model. Quantity is the base category, the

first dummy variable refers to Space and shape, the second one to Change and relationships, and the last one to Uncertainty.

The second covariate used in the DIF explanation is Competence cluster shown in Table 1. It is also a categorical variable and has three categories: Reproduction; Connections; and Reflections. Two dummy variables represent the covariate where Connections is the base category, the first dummy variable refers to Reproduction and the second one represents Reflection.

The third covariate is a binary variable and indicates if the item has support of graphical resource. Finally, the fourth covariate represents the size of the question, measured by the number x of words and standardised using the rule $(x - 75)/100$.

Although the complete analysis can be performed in an integrated way, incorporating all the uncertainties, the DIF explanation is performed here in a separate step. The DIF magnitude among the six countries is not large, since few items with large DIF were detected in the previous section. This way, it would be difficult to highlight possible effects of covariates in the DIF.

The analysis is performed by fixing the items that had a small probability of presenting DIF in all groups in the first analysis as anchor items (13, 21, 43, 61, 67 and 74). All the other items are fixed to have DIF, that is $\pi_{ig}^{h} = 1$. This way, the analysis presented in this section is designed only to identify possible effects of covariates in the DIF.

Several models were fitted for the DIF explanation. The first model for DIF in difficulty has only the covariates to indicate the Content category. In each of the following steps, a new covariate was included and the covariates that were not significant at 95% in all group in the previous model were removed. For the discrimination, the same models used for difficulty were fitted.

The results, presented in Tables 5 and 6, show that Space and shape items are more difficult for students from Ireland compared to the other countries. These items are also somewhat easier for students from New Zealand. Moreover, items related to Uncertainty are harder for students from the U.S.A. compared to the other five countries. They are also slightly harder for students from Australia and Ireland than for the ones from the other three countries.

Regarding the discrimination of the items, the results show that, in general, items with DIF in discrimination are more discriminant for students from Great Britain, followed by New Zealand and Ireland, Canada and Australia, and they are less discriminant for students from the U.S.A. On the other hand, if only items related to Uncertainty are considered, they are,

Covariates	Coefficients				
	Canada	**Australia**	**Ireland**	**U.S.A**	**New Zealand**
Model 1					
Intercept	0.03	-0.09	0.001	0.01	-0.02
Space and shape	-0.11	0.02	-0.27**	-0.13	0.11*
Change	-0.13*	-0.01	-0.06	-0.15	0.01
Uncertainty	-0.14**	-0.15**	-0.15	-0.33*	-0.07
Model 2					
Intercept	-0.03	-0.05	0.08	-0.03	0.03
Space and shape	-0.04	0.06	-0.24**	-0.07	0.10*
Uncertainty	-0.08	-0.15**	-0.11*	-0.29**	-0.10*
Reproduction	0.04	0.003	-0.09	-0.003	0.009
Reflection	-0.02	-0.03	-0.10	0.08	0.002
Model 3					
Intercept	-0.03	-0.19*	0.03	0.02	-0.04
Space and shape	-0.05	0.02	-0.25**	-0.03	0.10*
Uncertainty	-0.08	-0.12*	-0.13*	-0.27**	-0.09*
Graphical support	0.003	0.11	-0.02	-0.09	0.01
Model 4					
Intercept	0.02	-0.03	0.03	-0.01	0.02
Space and shape	-0.04	0.05	-0.24**	-0.02	0.11*
Uncertainty	-0.08	-0.15*	-0.13*	-0.23**	-0.08
Question size	0.03	-0.03	-0.01	0.07	0.002

Table 5: *R*esults of the DIF explanation regression model for DIF in difficulty. * means significant at 90% and ** at 95%.

on average, as discriminant for students from the U.S.A. as they are for students from Great Britain and they are more discriminant in these two countries than in the other four countries. Furthermore, the question size has an influence on the DIF in discrimination in Ireland and New Zealand. Larger questions make the items less discriminant in these countries, specially in Ireland.

5.4 Analysis of the proficiencies

The results obtained for the distribution of the proficiencies in the DIF analysis without covariates is presented. They are compared with the results from the original analysis of PISA and from the analysis with the Bilog-mg software.

Bilog-mg also allows the existence of DIF, but only in the difficulty. The scale in Bilog-mg is defined by assuming a standard normal distribution for the proficiencies from the reference group. The same is used for the Bayesian model proposed in this chapter.

Covariates	Coefficients				
	Canada	**Australia**	**Ireland**	**U.S.A**	**New Zealand**
Model 1					
Intercept	-0.14**	-0.14**	-0.12*	-0.27**	-0.05
Space and shape	0.06	-0.006	0.12	0.09	0.008
Change	-0.01	-0.02	0.02	-0.01	-0.02
Uncertainty	-0.006	0.05	0.07	0.15	0.03
Model 2					
Intercept	-0.16**	-0.17**	-0.06	-0.26**	-0.11*
Space and shape	0.08	0.008	0.10	0.09	0.02
Uncertainty	-0.003	0.04	0.05	0.15*	0.03
Reproduction	-0.01	0.01	-0.08	-0.02	-0.02
Reflection	0.02	0.10	-0.03	0.07	0.05
Model 3					
Intercept	-0.19**	-0.16**	-0.16**	-0.27**	-0.09*
Space and shape	0.06	0.02	0.09	0.09	0.05
Uncertainty	0.01	0.07	0.07	0.17*	0.03
Graphical support	0.04	0.02	0.05	0.04	-0.05
Model 4					
Intercept	-0.17**	-0.16**	-0.12**	-0.22**	-0.09*
Space and shape	0.08	0.01	0.08	0.11	0.02
Uncertainty	-0.02	0.05	0.01	0.16*	0.01
Question size	-0.06	-0.09	-0.20**	-0.08	-0.13*

Table 6: *R*esults of the DIF explanation regression model for DIF in discrimination. * - significant at 90% and ** - significant at 95%. Significant at α % means that the α % posterior credibility interval does not include 0.

Table 7 shows the mean and variance of the distributions of the proficiencies in each country considering:

- the original PISA proficiency (pv1math), which does not consider DIF;
- the results obtained with Bilog-mg without DIF;
- the results obtained with Bilog-mg without DIF in a modified scale;
- The results obtained with the integrated Bayesian model (IBM) proposed in this chapter;
- The results obtained with the integrated Bayesian model (IBM) in a modified scale;

The modified scale referred above consists in transforming the estimates in order to make the mean and variance of the reference group (GBR) the same as in the PISA scale.

Country	Parameter	PISA	Bilog-mg*	IBM*	Bilog-mg	IBM
Australia	Mean	524.11	522.95	522.15	0.1602	0.1515
N=235,486	Std. Dev.	95.60	94.40	93.78	1.0209	1.1042
Canada	Mean	532.70	529.43	528.77	0.2303	0.2231
N=330,098	Std. Dev.	87.33	90.00	85.26	0.9734	0.9221
Great Britain	Mean	508.14	508.14	508.14	0	0
N=696,215	Std. Dev.	92.47	92.47	92.47	1	1
Ireland	Mean	503.52	502.87	508.10	-0.0570	-0.0005
N=54,838	Std. Dev.	85.32	85.62	82.37	0.9259	0.8908
New Zealand	Mean	524.17	521.56	523.68	0.1452	0.1681
N=48,606	Std. Dev.	98.17	96.72	92.71	1.0459	1.0026
U.S.A.	Mean	483.64	484.85	474.99	-0.2518	-0.3585
N=3,140,301	Std. Dev.	95.37	91.40	97.52	0.9884	1.0546
Total	Mean	493.81	494.33	487.45	-0.1494	-0.2238
N=4,505,544	Std. Dev.	95.72	92.88	97.46	1.0045	1.0540

Table 7: *E*stimates of the mean and standard deviation of the proficiencies in each country. * refers to the modified scales.

PISA uses the Rasch model and a partial credit model and fixes the mean and standard deviation of all the proficiencies to be 500 and 100, respectively. For this reason, the proficiencies obtained with the integrated Bayesian model can not be directly compared to the ones from the original PISA results. They should be compared to the results from Bilog-mg, with GBR as the reference group. The transformed scales of the IBM and Bilog-mg are presented to just to give an idea about the differences among the countries, compared with the original results from PISA.

Table 7 shows that the results are very similar with and without DIF in four of the six countries. Differences are only found in Ireland, where the mean increases when DIF is considered, and in the U.S.A., where the mean decreases if DIF is accounted. It is important to mention that the data was weighted by the sampling weights.

6 Conclusions

The analysis presented here show the importance of appropriately account for all sources of heterogeneity present in educational testing. Incorporation of differentiation in the education pattern of countries allows the possibility for explanation of possible causes for it. This can lead the way for improvement in the schooling systems. The use of the Bayesian paradigm provides a number of advantages ranging from inclusion of relevant background information to allowance

for model identification.

In the context of the specific application considered, a host of indicators differentiating the educational systems of the English-speaking countries were identified. These may help to understand the nature and possible origins of the difference between them and lead the way for incorporation of beneficial practices in the currently available systems.

Appendix A

Countries that participated in PISA 2003:

OECD countries:

Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, United Kingdom, Scotland, United States.

Partner countries:

Brazil, Hong Kong-China, Indonesia, Latvia, Liechtenstein, Macao-China, Russian Federation, Serbia and Montenegro, Thailand, Tunisia, Uruguay.

Appendix B

The estimation of the parameter is performed by using MCMC methods to obtain a sample from the joint posterior distribution of all the parameters. The method used is Gibbs Sampling with Metropolis-Hastings steps.

Basically, all the parameters that appear explicitly in the likelihood are drawn using a Metropolis-Hastings step with a suitable random walk as the proposal distribution, because it is not possible to directly draw from their full conditional distributions. These parameters includes: the proficiencies, the item parameters and the DIF parameters. For all the other parameters, it is possible to directly draw from their full conditional distributions.

The parameters are simulated as follows:

- Abilities

To draw samples from $p(\theta|\beta, d, \lambda, \gamma, T, Y, W, Z)$, samples are drawn from:

$$p(\theta_j|\theta_{\neq j}, \beta, d, \lambda, \gamma, T, Y, W, Z) = p\left(\theta_j|\beta_{I(j)}, d_{I(j)g(j)}, \lambda_{g(j)}, Y_j\right) \propto$$
$$p\left(Y_j|\theta_j, \beta_{I(j)}, d_{I(j)g(j)}, \lambda_{g(j)}\right) p\left(\theta_j|\beta_{I(j)}, d_{I(j)g(j)}, \lambda_{g(j)}\right) =$$
$$p\left(Y_j|\theta_j, \beta_{I(j)}, d_{I(j)g(j)}\right) p\left(\theta_j|\lambda_{g(j)}\right) =$$
$$\prod_{i \in I(j)} p\left(Y_{ij}|\theta_j, \beta_i, d_{ig(j)}\right) p\left(\theta_j|\lambda_{g(j)}\right), \quad \forall j = 1, \ldots, J.$$

The calculations above are obtained by assuming independence between the students' abilities, and between the answers Y_{ij} when conditioned to the abilities and to the items' parameters. It is not possible to directly draw from the distribution above because of its complexity. So, the Metropolis-Hastings algorithm is used. A normal transition kernel is used and the proposal for the new state is

$$\theta_j^l \sim q\left(\theta_j|\theta_j^{l-1}\right) = N(\theta_j^{l-1}, c_\theta^2)$$

The tuning parameter is set as $c_\theta = 0.1$, chosen after a pilot study to assure an appropriate acceptance rate of the chain.

- Parameters of the distributions of the groups' abilities

  It is possible to directly draw from the full conditional distributions of the means and variances of the abilities' distribution since conjugate prior distributions were chosen.

  - Mean of the distributions of the groups' abilities

    If a prior distribution $\mu_g \sim N\left(\mu_{0g}, \sigma_{0g}^2\right)$ is chosen, the following full conditional distribution is obtained:

    $$p(\mu_g|\cdot) = p\left(\mu_g|\theta_{J(g)}, \sigma_g^2\right) \propto p\left(\theta_{J(g)}|\mu_g, \sigma_g^2\right) p\left(\mu_g|\sigma_g^2\right) =$$
    $$\prod_{j \in J(g)} p\left(\theta_j|\mu_g, \sigma_g^2\right) p\left(\mu_g|\sigma_g^2\right) = \prod_{j \in J(g)} p\left(\theta_j|\mu_g, \sigma_g^2\right) p(\mu_g)$$

    so:

    $$(\mu_g|\cdot) \sim N\left(m_g, s_g^2\right), \text{ where:}$$
    $$m_g = \frac{\sum_{j \in J(g)} \theta_j \sigma_{0g}^2 + \mu_{0g}\sigma_g^2}{n_g\sigma_{0g}^2 + \sigma_g^2} \quad and \quad s_g = \frac{\sigma_g \sigma_{0g}}{\sqrt{n_g\sigma_{0g}^2 + \sigma_g^2}}.$$

    J(g) represents the set of the students in the g group, $g = 2, \ldots, G$, and n_g is the number of students in group g.

– Variance of the distributions of the groups' abilities

$$p\left(\sigma_g^2|\cdot\right) = p\left(\sigma_g^2|\theta_{J(g)}, \mu_g\right) \propto \prod_{j\in J(g)} p\left(\theta_j|\mu_g, \sigma_g^2\right) p\left(\sigma_g^2\right)$$

If a prior distribution $\sigma_g^2 \sim GI(\alpha_g, \beta_g)$ is chosen, the following full conditional distribution is obtained:

$$\left(\sigma_g^2|\cdot\right) \sim GI\left(\alpha_g + \frac{n_g}{2}\ ,\ \frac{\sum_{j\in J(g)}(\theta_j - \mu_g)^2) + 2\beta_g}{2}\right)\ ,\quad g = 2, \ldots, G.$$

- Structural parameters β

Under the hypotheses of local independence of the items, samples of $p(\beta|\cdot)$ are drawn from:

$$p\left(\beta_i|\theta_{J(i)}, d_i, Y_{J(i)}\right) \propto p\left(Y_{J(i)}|\theta_{J(i)}, \beta_i, d_i\right) p\left(\beta_i|\theta_{J(i)}, d_i\right) =$$

$$\prod_{j\in J(i)} p\left(Y_{ij}|\theta_j, \beta_i, d_{ig(j)}\right) p(\beta_i) =$$

$$\prod_{j\in J(i)} p\left(Y_{ij}|\theta_j, \beta_i, d_{ig(j)}\right) p(a_i)p(b_i)p(c_i), \quad \forall i = 1, \ldots, I$$

The last equality comes from the prior independence of the parameters. The chosen prior distributions of the parameters are:

$a_i \sim LN\left(\mu_{a_i}, \sigma_{a_i}^2\right)$; $b_i \sim N\left(\mu_{b_i}, \sigma_{b_i}^2\right)$ e $c_i \sim beta\left(\alpha_{c_i}, \beta_{c_i}\right)$

In general, $\mu_{a_i} = 0,\ \sigma_{a_i}^2 = 2,\ \mu_{b_i} = 0,\ \sigma_{b_i}^2 = 1,\ \alpha_{c_i} = 5,\ \beta_{c_i} = 17$. These values are used, for example, as default values in the software Bilog-mg. In some cases, these values have to be modified to assure a better fit of current data or to add past information about the parameters.

Once again, it is not possible to directly draw from the full conditional distribution and the Metropolis-Hastings algorithm is used assuming the following transition kernels:

$a_i^l \sim LN\left(\ln(a_i^{l-1}), c_a\right)$, $b_i^l \sim N(b_i^{l-1}, c_b^2)$ e $c_i^l \sim U[c_i^{l-1} - \delta, c_i^{l-1} + \delta]$

In general, the values $c_a = 0.02,\ c_b = 0.1,\ \delta = 0.05$ are used.

- Structural DIF parameters

To draw samples from $p(d^h|\cdot)$, $h = a, b$, samples are independently drawn from:

$$p(d_{ig}^h|\cdot) = p\left(d_{ig}^h|d_{\neq i,g}^h, d_g^{\neq h}, Z^h, \theta_{J(i,g)}, \beta_i, \gamma_g, T, Y_{J(i,g)}, W\right) \propto$$

$$p\left(Y_{J(i)}|\theta_{J(i,g)}, \beta_i, d_{ig}^h\right) p\left(d_{ig}^h|d_{\neq i,g}^h, d_g^{\neq h}, W, \gamma_g, T, Z_{ig}^h\right) =$$

$$\prod_{j\in J(i,g)} p\left(Y_{ij}|\theta_j, \beta_i, d_{ig}^h\right) p\left(d_{ig}^h|W_i^h, \gamma_g^h, \tau_g^h, Z_{ig}^h\right),\ \ \forall i \in I_{dif}^h,\ g = 2, \ldots, G.$$

In the last equality, it is assumed that $T^h = (\tau_g^h)^2 I$, where I is the identity matrix $nid_h \times nid_h$, and nid_h is the number of items for which DIF in parameter h is assumed, $h = a, b$. The conditional prior distribution of the DIF parameters is $\left(d_{ig}^h | W_i^h, \gamma_g^h, \tau_g^h, Z_{ig}^h\right) \sim N\left(W_i^h \gamma_g^h, (\tau_g^h)^2\right)$ if $Z_{ig}^h = 1$, and the transition kernel used in the Metropolis-Hastings algorithm is the following:
$(d_{ig}^h)^{l+1} \sim N\left((d_{ig}^h)^l, 0.1\right)$, $\forall i$. On the other hand, $\left(d_{ig}^h | W_i^h, \gamma_g^h, \tau_g^h, Z_{ig}^h\right) \sim N\left(0, s_i^2 (\tau_g^h)^2\right)$ if $Z_{ig}^h = 0$. In practical situations, since s_i is very small, $d_{ig}^h \cong 0$ if $Z_{ig}^h = 0$.

- Parameters of the DIF regression structure

For the parameters γ_g, $g = 2, \ldots, G$, samples are drawn from:
$p\left(\gamma_g^h | \cdot\right) = p\left(\gamma_g^h | d_g^h, T_g^h, W^h, Z^h\right) \propto p\left(d_g^h | \gamma_g^h, W^h, T_g^h, Z^h\right) p\left(\gamma_g^h\right)$.
If a prior distribution $\gamma_g^h \sim N(\gamma_0^h, S_0^h)$ is assumed, the following full conditional distribution is obtained:
$\left(\gamma_g^h | d_g^h, T_g^h, W^h, Z^h\right) \sim N(H, L)$, where:

$$L = \left[\left(W_{I(Z_{ig}^h=1)}^h\right)^T \left(T_g^h\right)^{-1} W_{I(Z_{ig}^h=1)}^h + \left(S_0^h\right)^{-1}\right]^{-1} \quad \text{and}$$

$$H = L\left[\left(W_{I(Z_{ig}^h=1)}^h\right)^T \left(T_g^h\right)^{-1} d_{I(Z_{ig}^h=1),g}^h + \left(S_0^h\right)^{-1} \gamma_0^h\right]$$

Samples of $(\tau_g^h)^2$ are drawn from:
$p\left((\tau_g^h)^2 | \cdot\right) = p\left((\tau_g^h)^2 | d_g^h, \gamma_g^h, W_{I(Z_{ig}^h=1)}^h, Z^h\right) \propto$
$p\left(d_g^h | (\tau_g^h)^2, \gamma_g^h, W_{I(Z_{ig}^h=1)}^h, Z^h\right) p\left((\tau_g^h)^2\right)$
If a prior distribution $(\tau_g^h)^2 \sim GI(\alpha_{\tau_g^h}, \beta_{\tau_g^h})$ is assumed, the following full conditional distribution is obtained:

$$\left((\tau_g^h)^2 | \cdot\right) \sim GI\left(\alpha_{\tau_g^h} + \frac{\sum Z_{ig}^h}{2}\ ,\ \left[\frac{1}{2}\left(d_g^h - W_{I(Z_{ig}^h=1)}^h \gamma_g\right)^T \left(d_g^h - W_{I(Z_{ig}^h=1)}^h \gamma_g\right) + \beta_{\tau_g^h}\right]\right)$$

$g = 2, \ldots, G$.

- DIF indicator variable

$p\left(Z_{ig}^h | \cdot\right) = p\left(Z_{ig}^h | d_{ig}^h, T^h, W_i^h, \gamma^h\right) = p\left(d_{ig}^h | Z_{ig}^h, T^h, W_i^h, \gamma^h\right) p(Z_{ig}^h)$
If a prior distribution $Z_{ig}^h \sim Ber(\pi_{ig}^h)$ is assumed, the full conditional distribution $(Z_{ig}^h | \cdot) \sim Ber(\omega_{ig}^h)$ is obtained, where:

$$\omega_{ig}^h = \frac{1}{c_\omega} \pi_{ig}^h exp\left(-\frac{1}{2(\tau_g^2)}\left(d_{ig} - W_i^h \gamma_g^h\right)^2\right), \text{ and:}$$

$$c_\omega = \pi_{ig}^h exp\left(\frac{-1}{2(\tau_g^2)}\left(d_{ig} - W_i^h\gamma_g^h\right)^2\right) + \left(1 - \pi_{ig}^h\right) exp\left(\frac{-1}{2(s_i^2\tau_g^2)}(d_{ig}^h)^2\right)$$
for $g = 2, \ldots, G$.

References

Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*, Chapter Some Latent Traits Models and Their Use in Inferring Examinee's Ability. Reading: Ma. Addison-Wesley.

Cole, N. S. (1993). *Differential Item Functioning*, Chapter History and Development of DIF. Hillsdale, NJ: Lawrence Erlbaum.

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of American Statistical Association 85*, 398–409.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, 7*.

Rasch (1960). *Probabilistic Models for Some Intelligence and Attainment Test.* Copenhagen: Institute for Educational Research.

Soares, T. M., F. B. Gonçalves, and D. Gamerman (2008). An integrated bayesian model for dif analysis. *Journal of Educational and Behavioral Statistics*. To appear.

Stricker, L. J. and W. Emmerich (1999). Possible determinants of differential item functioning: Familiarity, interest and emotional reaction. *Journal of Educational Measurement 36*, 347–366.

Thissen, D. (2001). *IRTLRDIF v.2.0.b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning.* http://www.unc.edu/ dthissen/dl.html: Scientific Software International, Inc.