Using Item Response Mixed (IRM) Models to Improve the Comparability of Educational Assessment Scores

Tufi M. Soares*
Federal University of Juiz de Fora, Juiz de Fora, Brazil
E-mail: tufi@caed.ufjf.br

Abstract

In educational assessment, Differential Item Functioning (DIF) occurs when students with the same proficiency levels have different probabilities of giving the same answer to some of the test items. DIF can be a hindrance to the quality of the comparability of measures produced by large-scale evaluations. The problem can be particularly serious in evaluations that aim at making international comparisons of education quality, such as for instance the PISA program, which deals with a considerable amount of cultural, linguistic and curricular differences among its participating countries. Traditional methods that analyze the problem have a generally good performance when DIF is an exception, restricted to a few items of the test. However, this is not the case when, for instance, one compares the functioning of the items in a test administered in Japan and Brazil. Recently, new methods for DIF detection have been proposed, based on IRT procedures. Among them, it is worth mentioning those that employ Item Response Mixed (IRM) models for DIF analysis, which allows them to treat the problem even under unfavorable circumstances, as in the cases when there is a considerable amount of DIF in the test items. This article makes a review of these models and proposes new DIF analysis methods that are alternatives which are safe enough to guarantee the quality of the comparability of the results. Simulations and applications using results of PISA and some educational tests in Brazil are presented in order to show the efficacy of the methods.

Keywords: Comparability of international assessment scores, Differential Item Functioning, Item Response Theory.

1. Introduction

Nowadays, different international programs for educational assessment have been implemented. The objective of these programs is to compare the different educational systems of the different countries. Probably the most important of them is the PISA program. The OECD – Organization for Economic Co-Operation and Development - Program 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 and prepared for life in a large context and also to compare themselves to other students in other countries. The data collected in PISA surveys contains valuable information for researchers, policy makers, educators, parents and students. These types of programs impose a great challenge for the comparability of the educational assessment results, since it involves different kind of cultures, *curricula*, languages, etc. One of the most important problems that may be observed in this case is the known Differential Item Functioning (DIF) problem. For example, if an item is strongly related to some cultural aspect of a 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. Generally, DIF is defined as the property of some items of a test being easier or harder for the students of a specific group than students of a reference group, still that the students of both groups have the same proficiency or cognitive measure by the test. Studies conducted by the Educational

Testing Service (see Stricker and Emmerich (1999)), in the U.S.A., indicate that DIF may exist due to three factors in a large-scale assessment: 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.

In this work, is presented a new method to identify the DIF based on the Integrated Bayesian Model for DIF analysis proposed by Soares et al. (2009). The method was used to analyse the math assessment data of PISA-2003.

2. The Integrated Bayesian ("Mixed") DIF Model

The Item Response Theory (IRT) is a psychometric theory extensively used in educational assessment and cognitive psychology to analyze data arising from answers given to items belonging to a test. Its basic idea is to apply models, generally parametric ones, where the parameters represent important features of the items and subjects. Some common item parameters are discrimination, difficulty and guessing. The subject parameters may be the proficiency in math for example. The increasing use of IRT in educational assessment, particularly in international comparison assessment, is leading researchers to propose new methods to take DIF in account. In particular, Soares et al (2009) propose a new IRT Bayesian model that is a generalization of the known three parameters logistic model to accommodate the possible Differential Functioning of the Items. That model is presented here and it is the base of the method for DIF detection that will be proposed in the next section.

Suppose that a test has I items. Let $Y_{ij},\ j = 1, \dots, J,$ be the score attributed to the answer given by the student j to the item i. In the case where I is a dichotomous item, $Y_{ij} = 1$ if the answer is right and $Y_{ij} = 0$ if it is wrong. Define $p_{ij} = \frac{1}{1+e^{-\Delta_{ij}}}$, where $\Delta_{ij} = D\ a_i\ (\theta_j - b_i)$. In this case, the three parameters logistic model proposed by Birnbaum (1968) are given by:

$$P(Y_{ij} = 1 | \theta_j, a_i, b_i) = c_i + (1 - c_i)\, p_{ij}$$

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

In general, there can be different types of DIF (see Hanson (1998) for a wider characterization). For the three parameters model, the types of DIF can be characterized 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 used in this chapter associates the student's answer to his/her ability via (1) with

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

where a_{ig} is the discrimination parameter for item i and group g, b_{ig} is the difficulty parameter for item i and g, $c_{ig} \in [0,1]$ is the guessing parameter for item i, for *i = 1, ..., I, J = 1, ..., J* and *g = 1, ..., G.*

Since the item difficulty is a location parameter, it is natural to think about its DIF in the additive form an thus it is set as $b_{ig} = b_i - d^b_{ig}$. Analogously, since the discrimination is a scale parameter, it is natural to think about its DIF in the multiplicative form and thus it is set as $a_{ig} = e^{d_{ig}^a} a_i$. Therefore, $d_{ig}^b \left(d_{i1}^b = 0\right)$ represents the DIF related to the difficulty of the item in each group and $e^{d_{ig}^a}$ $(d_{i1}^a = 0)$ represents the DIF related to the discrimination of the item in each group. The 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 $\theta_j \,\middle|\, \lambda_{g(j)} \sim N(\mu_{g(j)}, \sigma_{g(j)}^2)$, where *g(j)* means the group which the student j belongs to. It is admitted that $\lambda_1 = (\mu_1, \sigma_1^2)$= (0, 1) to guarantee the identification of the model. On the other hand, $\lambda_g = \left(\mu_g, \sigma_g^2\right)$ is unknown for $g \geq 2$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 $a_i \sim LN(\mu_a, \sigma_a^2)$, $b_i \sim N\left(\mu_b, \sigma_b^2\right)$, and $c_i \sim Be(\alpha_c, \beta_c)$, for $i = 1, \dots, I$. The prior distributions for the parameters of the abilities distributions are $\mu_g \sim N\left(\mu_{0g}, \sigma_{0g}^2\right)$ and $\sigma_g^2 \sim IG(\alpha_g, \beta_g)$.

The set of anchor items (items for which $d_{i1}^b = d_{i1}^a = 0$, $\forall g$) is represented by $I_A \subset \{1, \dots, I\}$. The set of items for which the parameters vary between the groups is represented by I_{dif}. Moreover, $I_{dif}^a \subset I_{dif}$ is the set of items with DIF in the discrimination and $I_{dif}^b \subset I_{dif}$ is the set of items with DIF in the difficulty. 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.

Let $Z_{ig}^h, h = a, b$, be the DIF Indicator variable of item *I* in group *g*, for parameter h. Therefore, $Z_{ig}^h = 1$ if the parameter *h* of item *I* has DIF in group *g*, and $Z_{ig}^h = 0$, otherwise. Two possibilities may be considered: Z_{ig}^h is known or unknown and must be identified. The method proposed here is a combination of these two situations as can be seen in the next section. When $Z_{ig}^h = 0$, $d_{ig}^h \sim N(0, s^2)$, where is chosen to be small, such that d_{ig}^h is concentrated around 0, otherwise, if $Z_{ig}^h = 1$, $d_{ig}^h \sim N(m_g^h, \left(\tau_g^h\right)^2)$. Suitable prior distributions are $m_g^h \sim N\left(m_0^h, Sh_0^h\right)$, $\left(\tau_g^h\right)^2 \sim IG(\alpha_g^h, \beta_g^h)$ and $Z_{ig}^h \sim Ber(\pi_{ig}^h)$.

The estimation of the parameter is performed by using MCMC methods. The method used is Gibbs Sampling with Metropolis-Hastings steps (see Gamerman & Lopes (2006) for details).

3. A Method for DIF Detection

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. If one wishes to set an item as anchor one, it is sufficient to set $\pi_{ig}^h = 0$. 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 DIF´s magnitude, it is sufficient to make $\pi_{ig}^h = 1$. However, the central idea proposed in Soares et al. (2009) is to consider previous information and beliefs about the items' functioning to identify the DIF more precisely and effectively. Otherwise, the authors have shown that in many situations the identification of the DIF may be done also with non informative priors (setting $\pi_{ig}^h = 0.5$).

The method proposed here is quite different from the methods proposed in Soares et al. (2009), and it consists in to impose, iteratively, one item as anchor one in each step and compares, in a posterior analysis, the model results. This posterior analysis may be based on Bayesian decision rules. The algorithm for this method is the following:

i) Choose one item (k) to be anchor (set for this $\pi_{kg}^{h} = 0$); for all items set $\pi_{kg}^{h} = 0.5$;

ii) Identify all others items without DIF using The Integrated Bayesian DIF Model, consistently with the choose made in step i) (call the set of these items of $I_A(k)$);

iii) Choose another item, k', not presented in $I_A(k)$ and again find $I_A(k')$; repeat as if there is not more items not presented in all $I_A(k) \cup I_A(k') \cup$ ...;

iv) Apply one Bayesian Decision Rule to decide for one of the models correspondent to one of the anchor sets $I_A(k), I_A(k'), ...$

4. Some Results

The method was used for PISA math test of 2003. In the table 1, are compared the results for the both cases: considering and not considering the DIF effect over the proficiencies of the students. The results for the method described in the previous section are presented and they are compared with the results reported by PISA (that uses Rasch and Credit Partial Models) and 3PL model not considering DIF.

Table 1 – Comparison of the Means and Standard Deviations of the proficiencies for the countries of English Spoken

Country	Parameter	PISA report – Rasch Model	3PL model – not considering DIF	3PL model – considering DIF
Australia	Mean	524.11	522.95	521.87
	Standard Dev.	95.60	94.40	94.10
Canada	Mean	532.70	529.43	529.34
	Standard Dev.	87.33	90.00	87.65
Great Britain	Mean	508.14	508.14	508.14
	Standard Dev.	92.47	92.47	92.47
Ireland	Mean	503.52	502.87	508.50
	Standard Dev.	85.32	85.62	84.45
New Zealand	Mean	524.17	521.56	522.78
	Standard Dev.	98.17	96.72	93.76
U.S.A.	Mean	483.64	484.85	475.23
	Standard Dev.	95.37	91.40	96.78

The results suggest that small differences are observed for the means of Ireland and USA when the DIF is considered.

6. Conclusion

The analysis presented here shows the importance of consider appropriately the DIF in educational assessment. Although small differences was observed for the results – with or without DIF consideration, still that these differences are relevant of the point view of the countries governments. The algorithm proposed demonstrated be feasible and may be an alternative in some situations where identification problems constraint the traditional use of IRM models. Results not reported here shown these situations in practices.

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.

Gamerman, D. and H. F. Lopes (2006). *Markov chain Monte Carlo*: Stochastic simulation for Bayesian inference (2nd ed.). New York: Chapman & Hall/CRC.

Soares, T. M., F. B. Gonçalves, and D. Gamerman (2009). An integrated bayesian model for dif analysis. *Journal of Educational and Behavioral Statistics*. September 2009 vol. 34 no. 3 348-377.

Stricker, L. J. and W. Emmerich (1999). Possible determinants of differential item functioning: Familiarity, interest and emotional reaction. *Journal of Educational Measurement* Vol. 36, No. 4, Winter, 1999, 347-366.