Differential Item Functioning

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

Universidade Federal do Rio de Janeiro

Universidade Federal de Minas Gerais

Universidade Federal de Juiz de Fora

1 Introduction

Differential Item Functioning (DIF) occurs when individuals from different groups and same proficiency have different probability of correctly answering an item. These groups may be defined in terms of cultural, geographical, ethnical, social or economical differences. For example, in an international test, groups may be defined as developed and non-developed countries.

DIF analysis is a very important issue in educational research. Not considering DIF when this exists may lead to considerably distorted results in terms of individual and even population characteristics. For example, studies presented in O'Neil and McPeek [1993], Schmitt and Bleistein [1987], Berberoglu [1995], Gierl et al. [2003] showed differences in performance of items of educational assessment tests (like GRE, SAT, GMAT, etc.) for different groups associated with ethnic characteristics, gender, and/or socioeconomic status.

DIF may be uniform or not. DIF is said to be uniform when it is independent of the abilities, that is, the probability of correctly answering an item is uniformly greater for one group than the other, for all ability levels. In IRT models, this means that the characteristic curves of the item from the two groups are parallel. More specifically, in the 3PL model, it means that a difference is only found in the difficulty parameter. On the other hand, for a nonuniform DIF, there is interaction between DIF and ability.

There are basically two different approaches to deal with DIF in a test. In the first one, items with significant DIF are identified using some method and discarded from the main analysis which proceeds with the remaining ones in a DIF free context. However, item with DIF may contain important information about the groups, from cultural aspects to problems or differences in the teaching/learning system. The second approach then, does not discard item with DIF and incorporates DIF detection, quantification and maybe even explanation to the main analysis.

There are several well-established methods for DIF analysis in the literature that vary from each other in a variety of ways. They may differ in terms of the two different approaches described above, they may or not be based on IRT models, consider or not more than two groups or more than one factor (group division), deal or not with only uniform DIF, split or not the analysis into different steps. Ideally, a method should be as robust as possible to deal with all sorts of different DIF situations and, preferably, incorporate the DIF analysis to the main one using as much as possible the information present in the data.

The aim of this chapter is to provide an overview of some of the most well-known methods for DIF analysis, highlighting their main features, advantages and disadvantages. Special attention is given to model-based methodologies, that allows not only for DIF detection, but also its quantification and possible explanation.

The presentation follows the classification proposed by Wainer [1993], that separates empirically-based procedures from model-based procedures. Thus, the chapter is organized as follows: Section 2 presents empirically-based procedures for testing for the presence of DIF; Section 3 presents methodologies based on IRT models that are designed to test for and accommodate the presence of DIF, if it exists. Section 4 provides guidance on the use of an encompassing approach for DIF detection, quantification and explanation. Section 5 shows some simulation results to compare the efficiency of some of the methods presented. Section6 discusses some possible extensions and draws final conclusions.

2 Empirically based methods

Some of the most well-known empirically based methods for testing DIF are now presented. We have opted to present the Mantel-Haenszel procedure, perhaps the most well-known of them, and methods using random effects and logistic regression. Other methods can be found in Clauser and Mazor [1998] and Wang and Yeh [2003].

In order to present each of the methods and make the presentation as clear as possible we establish now the IRT and DIF notation used throughout the chapter. Typically, in a DIF analysis, one group is fixed as reference group and the other one(s) as focal group(s). We define Y_{ij} as the indicator variable that examinee j has correctly answered item i. We also define $P_{ij} = P(Y_{ij} = 1)$ as the probability that examinee j correctly answers item i. For IRT models, a_i, b_i and c_i represent the discrimination, difficulty and guessing parameters of item i, respectively. Finally, θ_j is the ability of examinee j.

2.1 The Mantel-Haenszel procedures

The Mantel-Haenszel (MH) procedure is based on the Mantel-Haenszel statistics [see Mantel and Haenszel, 1959] and was proposed in Holland and Thayer [1988]. It is particulary appealing because of its simplicity and low computational cost. The method is restricted to the comparison of two groups (reference and one focal) and, in principle, the abilities of the examinees ought to be known prior to the analysis, as the method requires the groups divided in matched comparable subgroups. As that is not the case in real situations, the groups may be matched using the total test score, which includes the item being studied. Some studies have shown that this is a good approximation when all the items follow the 3PL model for sufficiently long tests, but significant breakdowns may occur if the number of items is small (less than 25) and the difference between the mean score of reference and focal groups is large [see Stout, 1990, Allen and Donoghue, 1996, Donoghue et al., 1993, Roussos and Stout, 1996, Shealy and Stout, 1993b, Uttaro and Millsap, 1994]. Also, this matching procedure however is not dissociated from the DIF existence and, therefore,

the ability purification in successive stages is frequently recommended, where the items detected with DIF (in each stage) are eliminated from the ability calculation for the next analysis [see Holland and Thayer, 1988, Wang and Su, 2004].

Although very simple and practical, the MH procedure is not designed for and may not be powerful in detecting nonuniform DIF [see Hambleton and Rogers, 1989].

For the item under investigation, called *studied item* here, the MH procedure arranges the data of the studied item into 2×2 tables and the null hypothesis of no DIF is tested via Chi-square procedures. The test however does not measure the magnitude of the DIF (in a given scale) exhibited by the item in study.

Each one of the 2×2 tables refers to one matched set of examinees from reference and focal groups. For example, if examinees are matched using the total test score, there will be one table for each total test score found in the data. The tables are constructed as shown in table 1.

	$Y = 1$	$Y = 0$	Total
R	n_{R1k}	n_{R0k}	n_{Rk}
F	n_{F1k}	n_{F0k}	n_{Fk}
Total	n_{1k}	n_{0k}	n_k

Table 1: *Table constructed with data from examinees from the k^{th} matched set on the studied item. Entries are the number of occurrences in each subgroup. R and F refer to reference and focal groups, respectively. $Y = 1$ and $Y = 0$ correspond to correct and wrong answers, respectively.*

Under the null hypothesis H_0 of no DIF in the studied item, n_{R1k} and n_{F1k} are independent hypergeometric random variables with parameters (n_{Rk}, n_{1k}) and (n_{Fk}, n_{1k}), respectively.

The MH procedure tests this null hypothesis against the alternative hypothesis

$$H_a : \frac{p_{Rk}}{q_{Rk}} = \alpha \frac{p_{Fk}}{q_{Fk}}, \quad \forall k,$$

for $\alpha \neq 1$, where p_{Rk} and p_{Fk} are the probabilities of correct answer ($Y = 1$) for the reference and focal groups, respectively. Note that $\alpha = 1$ corresponds to the null hypothesis H_0. Parameter α is called the common odds-ratio and the MH chi-square statistics is given

$$\frac{(|\sum_k n_{R1k} - \sum_k E[n_{R1k}]| - 1/2)^2}{\sum_k Var[n_{R1k}]}, \tag{1}$$

where

$$E[n_{R1k}] = \frac{n_{Rk} n_{1k}}{n_k} \quad \text{and} \quad Var[n_{R1k}] = \frac{n_{Rk} n_{Fk} n_{1k} n_{0k}}{n_k^2 (n_k - 1)}.$$

Under H_0, the MH statistics has an approximate χ_1^2 distribution and the MH test of size γ rejects H_0 if the MH statistic is larger than the $(1 - \gamma)$ quantile of the χ_1^2 distribution.

If H_0 is rejected, the intensity and direction of DIF may be calculated. Assuming that DIF is uniform across strata k, one can compute an estimate of α given by

$$\hat{\alpha}_{MH} = \frac{\sum_k n_{R1k} n_{F0k} / n_k}{\sum_k n_{R0k} n_{F1k} / n_k}, \tag{2}$$

which re-scaled to

$$\Delta_{MH} = -2.35 \log(\hat{\alpha}_{MH}) \tag{3}$$

is a measure of the amount of DIF in the scale of differences in the item difficulty as measured in the ETS "delta scale" [see Holland and Thayer, 1985]. This estimate can be used to classify the magnitude and direction of DIF: favorable to the reference (focal) group if $\Delta_{MH} < 0$ ($\Delta_{MH} > 0$, respectively) and high if $|\Delta_{MH}| > 1.5$.

The standard error of Δ_{MH} is given by [see Phillips and Holland, 1987] as

$$se(\Delta_{MH}) = \left[(2.35)^2 \sum_k \frac{(n_{R1k} n_{F0k} + \Delta_{MH} n_{R0k} n_{F1k})}{2R^2 n_k^2} [n_{R1k} + n_{F0k} + \Delta_{MH}(n_{R0k} + n_{F1k})] \right]^{1/2}, \tag{4}$$

where $R = \sum_k (\frac{n_{R1k} n_{F0k}}{n_k})$. Based on the asymptotic normality of Δ_{MH}, Zwick et al. [1999] followed an approximate Bayesian procedure admitting that

$$\Delta_{MH_i} | \delta_i \sim N(\delta_i, S_i^2) \tag{5}$$

where $\delta_i = E(\Delta_{MH_i})$ is the DIF parameter of interest. Assuming additionally a prior distribution $\delta_i \sim N(\delta, \Gamma^2)$, gives the posterior distribution for δ_i given by

$$\delta_i|\Delta_{MH_i} \sim N(W_i\Delta_{MH_i} + (1 - W_i)\delta, W_i S_i^2) \text{ where } W_i = \frac{\Gamma^2}{S_i^2+\Gamma^2}.$$

The above authors assume the quantities S_i^2, Γ^2 and δ known and given respectively by the estimated values of $se(\Delta_{MH_i})$, mean and variance of the estimates of Δ_{MH_i} in an empirical Bayes procedure [see Zwick et al., 1999].

Finally, the method can be extended to account for polytomous items and/or multiple groups with generalized MH procedures. Two different extensions are more common to the case of polytomous items: generalized Mantel-Haenzel test - GMH [see Mantel and Haenszel, 1959, Zwick et al., 1993] for nominal responses and the Mantel test [see Mantel, 1963, Zwick et al., 1993] for ordinal responses. The latter is briefly described below. GMH test may also be found in the above references and generalizations for multiple groups may be found in Fidalgo and Madeira [2008].

Consider the generalization below for Table 1 for the case of items with more than 2 responses:

	$Y = 1$	$\cdots$	$Y = i$	$\cdots$	$Y = p$	Total
R	n_{R1k}	$\cdots$	n_{Rik}	$\cdots$	n_{Rpk}	n_{Rk}
F	n_{F1k}	$\cdots$	n_{Fik}	$\cdots$	n_{Fpk}	n_{Fk}
Total	n_{1k}	$\cdots$	n_{ik}	$\cdots$	n_{pk}	n_k

Table 2: p - number of categories

Under the hypothesis H_0 of no DIF, $\{n_{Fik},\ i = 1, \ldots, p\}$ and $\{n_{Rik},\ i = 1, \ldots, p\}$ are independent and follow multivariate hypergeometric distributions with parameters $(n_{Fk}, n_{ik}, i = 1, \ldots, p.)$ and $(n_{Rk}, n_{ik}, i = 1, \ldots, p.)$, respectively. Let $R_k = \sum_{i=1}^{p} n_{Rik} i$ and $F_k = \sum_{i=1}^{p} n_{Fik} i$ be the sums of scores of reference and focal groups respectively. Then,

$$E[R_k] = \frac{n_{Rk}}{n_k} \sum_{i=1}^{p} i\ n_{ik} \quad , \quad E[F_k] = \frac{n_{Fk}}{n_k} \sum_{i=1}^{p} i\ n_{ik}$$

$$Var(R_k) = Var(F_k) = \frac{n_{Rk} n_{Fk}}{n_k^2 (n_k - 1)} \left[(n_k \sum_{i=1}^{p} i^2 n_{ik}) - (\sum_{i=1}^{p} i \; n_{ik})^2 \right]$$

The Mantel statistics $\dfrac{(\sum_k R_k - \sum_k E[R_k])^2}{\sum_k Var(R_k)}$ has χ_1^2 distribution, under H_0 [see Mantel, 1963] and the Mantel test of size γ rejects H_0 if the MH statistic is larger than the $(1-\gamma)$ quantile of the χ_1^2 distribution.

2.2 Random effects models

This method decomposes the variation of the MH D-DIF statistics in (3) to address issues related to different administrations of the test. The authors argue that a separate DIF analysis of administrations of the same test are bound to be very inefficient due to sampling variation, which motivated their variance decomposition approach.

The random effects model for the MH D-DIF statistics from a single administration is given by

$$y_i = \mu + \xi_i + \epsilon_i \tag{6}$$

where, in this section, $y_i = -2.35 \log(\hat{\alpha}_{MH})$, μ is an unknown constant (usually close to 0), and ξ_i and ϵ_i are random terms.

The term ξ_i is the deviation of DIF in item i from the average DIF μ. The ξ_i's are considered to be random, all independent and identically distributed with distribution $\mathcal{N}(0, \tau^2)$. Furthermore, the ϵ_i's are independent with distribution $\mathcal{N}(0, s_i^2)$, where s_i^2 is the conditional variance of $\log(\hat{\alpha}_{MH})$ (conditional on ξ_i).

Finally, it is assumed that the administration of the test involves a large number of examinees in both groups so that the estimate of each s_i^2 has negligible sampling variance, which is therefore ignored. This way, the DIF in a test can be characterized by the variance of the ξ_i's, τ^2. The case where $\tau^2 = 0$ corresponds to complete absence of DIF, whereas large values of τ^2 implies that either a very large proportion of items have moderately large DIFs or a small number of items have very large DIFs.

The analysis with the model in (7) consists of four steps. Firstly, the y_i's and s_i's are obtained. Then, the global parameters μ and τ^2 are estimated. On the third step, the posterior distributions of the coefficients $\mu + \xi_i$ are computed. Finally, the final step consists of deciding for each item whether to retain or replace it, based on the chances of reduction in the value of τ^2.

If the same test is administered on more than one occasion, say K, the variation of the MH D-DIF coefficient across the administrations can be represented by an extra random term in the model. Denoting y_{ik} as the MH D-DIF statistic for item i in administration k, the new model is given by

$$y_{ik} = \mu + \xi_i + \alpha_{ik} + \epsilon_{ik} \tag{7}$$

where $\xi_i + \alpha_{ik} = \xi_{ik}$ is the population-specific random effect. It is assumed that the α_{ik}'s are i.i.d. with $\mathcal{N}(0, \sigma^2)$ distribution. The variance σ^2 measures the variation in DIF across the administrations.

2.3 Logistic regression method

The logistic regression method was proposed in Swaminathan and Rogers [1990] and is more robust than the MH procedure in the sense that it considers the continuous nature of the abilities and is suitable to detect non-uniform DIF. The method is based on the following logistic regression model for predicting the probability of a correct response to an item:

$$logit\left[P(Y_{jg} = 1)|\theta\right] = \beta_{0g} + \beta_{1g}\theta_{jg}, \quad j = 1, \ldots, n_g, \ g = 1, 2, \tag{8}$$

where Y_{jg} is the answer given by examinee j from group g and θ_{jg} is his/her ability. β_{0g} and β_{1g} are the intercept and slope parameters in group g, respectively.

Under this formulation, if the studied item exhibits no DIF, then $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. On the other hand, if $\beta_{01} \neq \beta_{02}$ and $\beta_{11} = \beta_{12}$, the curves are parallel and the item exhibits uniform DIF. For any other case, non-uniform DIF is present.

The model in (8) can be conveniently reparameterised as

$$logit\,[P(Y_{jg}=1)|\theta] = \tau_0 + \tau_1\theta_j + \tau_2 g_1 + \tau_3(\theta_j g_1), \tag{9}$$

where g_1 is the indicator variable of examinee j belonging to group 1. This implies that

$$\tau_2 = \beta_{01} - \beta_{02} \quad \text{and} \quad \tau_3 = \beta_{11} - \beta_{12}$$

The model requires prior knowledge of the abilities. Once again, these may be fixed as the total test scores or, preferably, some abilities purification procedure ought to be adopted. Given that the abilities are known, we have in hands a logistic regression model for with we observe some data (answers given to the studied item) and need to estimate the coefficients to come to a conclusion about the DIF. The original paper adopts a maximum likelihood approach and uses asymptotic theory for MLE's to carry out a hypothesis test for τ_2 and τ_3. Nevertheless, other inference methods may be also adopted, for instance, Bayesian methods.

Various extensions and adaptations of logistic regression were proposed for the case of polytomous data. Miller and Spray [1993] suggest its use with the grouping indicator g_R ($g_R = 1$ indicates the reference group and $g_R = 0$ indicates the focal group) as response variable as

$$logit[P(g_R = 1|\theta_j)] = \gamma_0 + \gamma_1\theta_j + \gamma_2 Y_j + \gamma_3\theta_j Y_j \tag{10}$$

where Y_j is the item response, θ_j is the proficiency, when it is known or the total test score, otherwise. Note that both cases disregard measurement errors. $\gamma_3 = \gamma_2 = 0$ indicates no DIF; $\gamma_3 = 0$ and $\gamma_2 \neq 0$ indicates the presence of uniform DIF; any other case indicates the presence of non-uniform DIF. The authors suggest choosing the model according to a series of likelihood ratio tests, but other methods may also be used.

French and Miller [1996] considered the case of ordinal responses and suggested the replacement of response Y_j, with p categories, by $p-1$ dichotomous variables Y_j^i, $i = 1, \ldots, p-1$, each representing a different contrast between categories. They suggested

three schemes for construction of Y_j^i: continuation ratio, cumulative and adjacent categories schemes. The first one admits that $Y_j^i = 0$ if $Y_j = i$ and $Y_j^i = 1$ if $Y_j > i, i = 1, \ldots, p-1$. In the second one, $Y_j^i = 0$ if $Y_j \leq i$ and Y_j^i if $Y_j > i$. In the latter scheme $Y_j^i = 0$ if $Y_j = i$ and $Y_j^i = 1$ if $Y_j = i + 1$. DIF detection tests can be obtained from inference about the regression coefficients γ_2 and γ_3 of model

$$logit[P(Y_j^i = 1|\theta_j)] = \gamma_0 + \gamma_1\theta_j + \gamma_2 g_F + \gamma_3\theta_j g_R, \quad i = 1, \ldots, p-1. \tag{11}$$

A simultaneous test for all $p-1$ variables may be only carried out for the first scheme because it is the only scheme that yields independent Y_j^i variables.

2.4 Hierarchical logistic regression method

This method was proposed in Swanson et al. [2002] and consists of a hierarchical logistic regression model. The main contribution of the paper lies on the second level of the model where the coefficients of the logistic regression are explained by covariates related to item characteristics. This approach helps to identify consistent sources of DIF across items and to quantify the proportion of variation in DIF explained by the covariates. This models is related to the one described in Rogers and Swaminathan [2000], where the authors use examinees characteristics at the second level to improve the matching of reference and focal groups.

The hierarchical logistic regression model of Swanson et al. [2002] is given by

$$logit\left[P(Y_{ij} = 1)\right] = b_{0i} + b_{1i}\theta_j + b_{2i} * group_j, \tag{12}$$

where θ_j is the ability of examinee j and $group_j$ is the indicator variable of examinee j belonging to the focal group. Note that the discrimination of the items are constrained to be the same for both groups and b_{2i} reflects the deviation of the item difficulty in the focal group from the reference group.

The second level of the model treats the coefficients of (12) as random variables where

item characteristics are used to explain DIF, for example:

$$\begin{aligned} b_{0i} &= \beta_{00} + \epsilon_{0i} \\ b_{1i} &= \beta_{10} + \epsilon_{1i} \\ b_{2i} &= \beta_{20} + \beta_{21} I_1 + \beta_{22} I_2 + \ldots + \beta_{2n} I_n + \epsilon_{2i}, \end{aligned}$$

where $I_1, \ldots, I_n$ are interval or dummy-coded item characteristics and the ϵ's are the unexplained variances.

Estimation is performed using empirical Bayes methods with estimates of the abilities assumed to be known and fixed.

3 Methods based on IRT Models

Let us assume that the response to an item is governed by the IRT model family. Then, differential item functioning could be verified by allowing the model parameters to vary across groups of respondents. Substantial difference in these estimates would provide evidence in favor of the presence of DIF. for example, in the 2PL model one may assume for item i that

$$\log\left(\frac{P_{ijg}}{1 - P_{ijg}}\right) = 1.7[a_{ig}(\theta_j - b_{ig})] \tag{13}$$

where P_{ijg} is the probability of correctly answering the question for item i by respondent j in group $g, g = 1, \ldots, G$.

One problem that emerges in these contexts is the identifiability of these models. If a test has I items and it is admitted that the parameters of all items vary across groups, then it is easy to show that for any constants c_{1g}, c_{2g}:

$$a_{ig}(\theta_j - b_{ig}) = a_{ig} \cdot c_{1g}(\tfrac{\theta_j - b_{ig} + c_{2g} - c_{2g}}{c_{1g}}) = a_{ig} c_{1g}[\tfrac{(\theta_j + c_{2g})}{c_{1g}} - \tfrac{(b_{ig} - c_{2g})}{c_{1g}}]$$

The above non-identification of the model reflects the problem of maintaining the comparability of the proficiency scales across groups when DIF is present. For example,

it is difficult to tell whether the better result of a group when compared to another group was due to an improved proficiency of the first group or due to the presence of positive DIF for this group, making the items easier. So, identification of these models is related to further assumptions regarding the presence and intensity of DIF.

One example if the method proposed by Zimowski et al. (1996) for the 3PL model. The assume that DIF occurs only in the difficulty parameter (as also available in the `BILOG-MG` software) and impose the restriction that the average item difficulty $\frac{1}{I}\sum_{i=1}^{I} b_{ig}$ is the same for all groups. This restriction ensure identifiability but pays the price of assuming that DIF is always compensated, that is, is there are easier items for one group there must be more difficult items when compared to another group. Another common identification restriction is to assume from the start that a few items do not present DIF. These items are usually refereed to as anchor items.

3.1 IRT-LR methods

The IRT-LR methods were proposed in Thissen et al. [1993] and one of them was implemented in the software `IRTLRDIF v.2.0b` [see Thissen, 2001], for two models (3PL and Samejimas's (1969, 1997) graded model). These methods are IRT model based and, generally speaking, detect items with DIF by first estimating then testing the item parameters for the reference and focal groups. Thissen et al. [1993] proposes three different IRT-LR procedures that differ in the methods used to estimate the parameters and in the IRT model being used, but that all use LR tests. The authors also propose an alternative to LR tests, the IRT-D^2, that uses maximum likelihood estimation and ratios of parameter estimates to their standard erros to test the DIF hypothesis.

The IRT-LR methods fix the distribution of abilities for the reference group, generally $\mathcal{N}(0,1)$, and estimates these parameters for the focal group. Naturally, some strategy must be adopted to match the examinees from different groups. Typically, a set of anchor items (those which do not exhibit DIF) is fixed and the procedures are applied to each of

the remaining items, separately. Alternatively, `IRTLRDIF` also offers the option of fixing the parameters of all the other items apart from the one being studied. This last option may not be very efficient, specially if many items exhibit DIF and mostly to the same direction.

The IRT-LR methods proceed as follows: given that an item i is at study, the parameters of the IRT model adopted are estimated using some estimation method - `IRTLRDIF` uses the Bock and Aitkin [1981] algorithm. The parameters to be estimated are those of item i and the items of the anchor set (if this is fixed) or of all the other items, the abilities and the parameters from the focal group ability distribution. Two models are fit here - the compact model [C], where the item parameters are assumed to be the same for both groups, and the augmented model [A], for which the parameter(s) of item i differ between the groups. The next step consists of obtaining the LR test statistic given by

$$G^2(d.f.) = 2 \log \frac{Likelihood[A]}{Likelihood[C]}, \tag{14}$$

where $Likelihood[\cdot]$ is the likelihood of model $[\cdot]$ given the estimates of the model parameters. It is well documented that, under very general assumptions, $G^2(d.f.)$ follows a $\mathcal{X}^2(d.f.)$ distribution under the null hypothesis that model [C] is "correct", where $d.f.$ is the difference between the number of parameters in [A] and the number of parameters in [C].

Depending on the IRT model in hands, successive tests may be performed to test parameters individually, see, for example, the procedure adopted by the `IRTLRDIF` software for the 3PL model. In the simplest case of the 1PL model, the null hypotheses states that the studied item has no DIF against the alternative hypotheses that the difficulty parameter of this item is different for the reference and focal groups.

3.2 The SIB test

The SIB test, proposed in Shealy and Stout [1993a], is an statistical test designed to simultaneously detect DIF present in one or more items of a test. It was the first IRT

based method to simultaneously detect DIF in more than one item, but has limited power to detect nonuniform DIF.

The SIB test looks at the DIF problem from a multidimensionality perspective using a multidimensional non-parametric IRT model. The ability vector $\boldsymbol{\theta}$ is decomposed into $\{\theta, \boldsymbol{\eta}\}$, where θ is the ability intended to be measured by the test and $\boldsymbol{\eta}$ are nuisance abilities not meant to be measured in the test but that do influence in the answer to one or more items. This way, DIF items are those which, apart from the target ability, measure one or more nuisance abilities.

In order to statistically detect DIF for a subset of items, it is necessary to identify a subset of anchor items, i.e. items that only measure the target ability, in order to match the examinees of equal target ability. The SIB test proceeds by constructing a DIF index $\beta_{\boldsymbol{U}}$ against the focal group considering the studied item subset. More specifically:

$$\beta_{\boldsymbol{U}} = \int_{\Theta} B(\theta) f_F(\theta) d\theta, \tag{15}$$

where $f_F(\theta)$ is the probability density function of θ for the focal group and

$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta) = E[h(\boldsymbol{U})|\theta, g = R] - E[h(\boldsymbol{U})|\theta, g = F], \tag{16}$$

where $\boldsymbol{U}$ is the vector of answers given to the subset of studied items by an examinee chosen at random and $h(\boldsymbol{U})$ is a test score.

The SIB test is carried on by testing the hypothesis:

$$H_0 : \beta_{\boldsymbol{U}} = 0 \quad \text{vs} \quad H_1 : \beta_{\boldsymbol{U}} > 0.$$

The test statistics, which is essentially an estimate of β_U normalized to have unit variance, is given by

$$B = \frac{\hat{\beta}_{\boldsymbol{U}}}{\hat{\sigma}(\hat{\beta}_{\boldsymbol{U}})}. \tag{17}$$

See Shealy and Stout [1993b] for the full expression. Finally, B is approximately standard normal when $\beta_{\boldsymbol{U}} = 0$ and the target ability distributions are the same.

3.3 Multilevel Bayesian IRT method

May [2006] proposes a multilevel Bayesian IRT model to compared SES scores across different nations. A set of anchors items are fixed and the remaining ones are allowed to operate differently across nations. This way, the resultant scores are internationally comparable.

The proposed model consists of the standard graded response model of Samejima [1997] with the discrimination and threshold parameters being specific to each nation in the non-anchor items:

$$\log\left(\frac{\Omega_{hjki}}{1-\Omega_{hjki}}\right) = 1.7\left[a_{hi}(\theta_j - b_{hi} + \delta_{ki})\right], \tag{18}$$

where Ω_{hjki} is the cumulative probability that student j from nation h responds in category k or higher on item i, θ_j is the SES score of student j, a_{hi} is the discrimination power for item i in nation h, b_{hi} is the overall threshold for item i in nation h and δ_{ki} is the category parameter for response category k on item i.

Inference is conducted via MCMC in a Bayesian framework with vague priors. This approach also allows the treatment of the missing data problem encountered in SES data by simply sampling these data values from their sampling distribution in a Gibbs sampler step.

3.4 Integrated Bayesian DIF model (IBDM)

The main idea behind the integrated DIF model is to reconcile the two contrasting activities of testing and estimating DIF into a single, unified framework. In passing, it also reconciles with simultaneous estimation of all model parameters (item characteristics, individual proficiencies and DIF magnitude). This task is achieved more naturally under the Bayesian paradigm but its use under the frequentist paradigm may be possible, albeit substantially more cumbersome.

The model that serves as a basis for this proposal can be viewed as an extension and

generalization of (18). The presentation here will concentrate on the 2PL IRT model with dichotomous responses but it can be generalized to other IRT settings. The model has observational equation with DIF quantification

$$logit\,[P(Y_{jg}=1)|\theta] = a_g(\theta_{jg}-\beta_g), \quad j=1,\ldots,n_g,\ g=1,...,G, \tag{19}$$

where $a_g = ad_g^a$ and $b_g = b + d_g^b$, with $d_g^a = 1$ and $d_g^b = 0$, for the reference group $g = 1$. The choice of multiplicative DIF effect for the discrimination a and additive DIF effect for the difficulty seem to make sense given their respective natures but these effect could have been defined to act differently.

These models allow for DIF explanation and DIF detection. DIF explanation is achieved in a (mixed) regression setting

$$d_g^h = \ z_g'\gamma(\ + w_g^h\ ), \quad g=2,...,G,\ h=a,b \tag{20}$$

where z_g are characteristics associated with group g and the item under consideration, γ is their coefficient and the added random terms $w_g^h \sim N(0,\tau_g^h)$ may account for extra-variability. A typical example is a question involving monetary issues in Mathematics exam at an elementary level. Our experience shows that children from rural areas appear to be less exposed to money and end up having more difficulty in correctly answering than urban children, irrespective of their mathematical ability.

DIF detection is achieved by assuming that the explanation in 20 is only valid with probability π_g^h. The complementary probability $1-\pi_g^h$ is associated with a DIF model concentrated around 0, representing no DIF. Soares et al. [2009] use the representation $d_g^h \sim N(0, c\tau_g^h)$, where c is set at a suitably large value to ensure d_g^h remains close to 0. This provides an indication that d_g^h is a reasonable hypothesis and thus DIF is assumed not to exist in this case.

The model is quite general and solves many previously existing difficulties. It only requires a single anchor item even in the presence of vague prior information. When prior information exists, no anchor item may be required. The prior definition of an anchor

item does not impose that all other items are not anchor and have DIF. The mixture component above allows only the detection of items with DIF. The other items will then be set as anchor as well. The integrated model also allows for different distributions for the proficiencies of each group, with the standard normal distribution retained for the reference group. A possible generalization is suggested by Fox and Verhagen [2010] where a regression model for the mean proficiencies explains their differences across groups. Some of the above characteristics of the integrated model are detailed in the simulations of the next sections.

4 Practical Guidance for IBDM

The integrated model of the previous section can be used in many different ways to detect DIF. Many of the existing models previously presented are recovered with suitable restrictions on this formulation, depending on the choice of π_{ig}^h. The model is general enough to circumvent many limitations of traditional DIF mehodologies. All methods below will only differ in the specification of π_{ig}^h, $h = a, b$. Note that $\pi_{ig}^h = 0$, for $h = a, b$, implies item i is an anchor item and $\pi_{ig}^h = 1$, for $h = a, b$, implies that item i is non-anchor and its DIF must be estimated. Finally, $0 < \pi_{ig}^h < 1$, for $h = a, b$ is the uncertain scenario where the item may be identified as anchor or non-anchor.

A few practical points are better illustrated empirically. A simulated data set consist of a typical setup with 20 items from which 5 (and 2) of them have DIF in difficulty (discrimination, respectively). There will be assumed 1,000 respondents at each of 2 groups (reference and focal). Results are obtained via MCMC with `OpenBUGS` ([Lunn et al., 2009]) and are based on the code provided in the Appendix A. 25,000 iterations were required to ensure convergence. Additional 10,000 iterations were obtained and a final sample of 1,000 values was used for inference after thinning at every 10 iterations. Readers are referred to Gamerman and Lopes [2006] for more details on MCMC algorithms.

Among the most important situation in practice, one may list:

1. General Case (Non informative prior for DIF existence);

   In this case, $\pi_{ig}^h = 0.5$ may be set for all items or non-informative prior distributions may be set form them. DIF identification will then depend on the estimation of the parameters associated with d_{ig}^h, $h = a, b$. DIF can be detected in a number of ways. The simplest ones are by comparison of the posterior mean or median of π_{ig}^h against a chosen cutoff point (eg 0.5) or by assessing whether the credibility intervals of d_{ig}^h include 0. Generally, more sensitivity is obtained with smaller prior variances for the components of d_{ig}^h (eg τ_g^h). Table 2 presents the results obtained with the estimation procedures. It clearly shows good recovery of the simulated values of the parameters and good DIF detection. The model correctly identified 4 out of 5 items with difficulty DIF and 1 out of 2 items with discrimination DIF. The not identified DIF discrimination for item 20 had very small DIF values implying very little difference between analyses with and without DIF for this item. For item 11, the difficult parameter for focal group is very small and the method identified the DIF in discrimination parameter, what may be reasonable. Generated (estimated) values for means and standard deviations were 0.00 (0.00) and 1.00 (1.01) for reference group and -0.73 (-0.75) and 1.00 (0.98) for focal group.

   Difficulties associated with correct DIF identification may occur in some situations. This typically occurs when the focal group mean was substantially larger than the the reference group mean and vice versa or when DIF is substantially asymmetric. This problem may be mitigated by using more informative priors for π_{ig}^h or for the other DIF parameters. A simulated example illustrating this point will be presented in Appendix B.

2. Fixing Anchor Items

   One may assume that some items do not have DIF. These items form the Fixed

Table 3: **Results for simulated dataset**

item	Reference group			Focal group		DIF detection for a			DIF detection for b		
						Credibility limits		π^a_{i2}	Credibility limits		π^b_{i2}
	a	b	c	a	b	5%	95%		5%	95%	
1	0,98(0,92)	-0,76(-0,74)	0,10(0,11)	0,98(0,92)	-0,76(-0,74)	0,00	0,00	0,06	-0,02	0,00	0,10
2	0,70(0,76)	-0,63(-0,52)	0,16(0,21)	0,70(0,78)	-0,63(-0,53)	0,00	0,24	0,16	0,00	0,10	0,13
3	1,47(1,50)	0,50(0,51)	0,24(0,25)	1,47(1,49)	0,50(0,51)	-0,22	0,03	0,19	0,00	0,00	0,10
4	0,89(0,81)	1,96(2,27)	0,39(0,37)	0,89(0,83)	1,96(2,35)	-0,74	0,20	0,33	-1,18	0,56	0,40
5	1,49(1,36)	1,02(1,14)	0,17(0,19)	1,49(1,41)	1,02(1,15)	-0,08	0,24	0,20	-0,08	0,00	0,12
6	0,77(0,77)	-1,81(-1,92)	0,26(0,23)	0,77(0,76)	-1,81(-1,85)	-0,12	0,00	0,12	-0,35	0,00	0,33
7	1,45(1,34)	0,17(0,14)	0,19(0,19)	1,45(1,37)	0,17(0,15)	0,00	0,20	0,15	-0,09	0,01	0,14
8	1,56(1,27)	-2,17(-2,18)	0,24(0,23)	1,56(1,36)	-2,17(-2,19)	0,00	0,38	0,18	0,00	0,13	0,17
9	1,42(1,31)	-0,35(-0,33)	0,25(0,23)	1,42(1,56)	-0,35(-0,32)	0,00	0,65	0,40	-0,03	0,00	0,08
10	0,92(1,06)	0,35(0,37)	0,28(0,26)	0,92(0,95)	0,35(0,28)	-0,55	0,00	0,36	0,00	0,35	0,38
11	1,52(1,50)	-2,43(-2,48)	0,11(0,20)	1,52(4,62)	**-3,03(-2,60)**	**0,00**	**1,93**	**0,79**	0,00	0,64	0,30
12	2,02(2,01)	-0,56(-0,51)	0,21(0,22)	2,02(1,98)	-0,56(-0,50)	-0,19	0,02	0,15	-0,05	0,00	0,10
13	0,67(0,86)	-0,91(-0,68)	0,06(0,12)	0,67(0,85)	-0,91(-0,74)	-0,11	0,00	0,10	0,00	0,27	0,35
14	1,73(1,68)	-0,65(-0,57)	0,22(0,22)	1,73(1,77)	**-0,41(-0,33)**	0,00	0,36	0,23	**-0,35**	**-0,13**	**1,00**
15	1,05(1,22)	0,17(0,24)	0,23(0,23)	1,05(1,20)	0,17(0,25)	-0,25	0,00	0,18	-0,01	0,00	0,09
16	0,83(0,90)	-0,68(-0,51)	0,12(0,17)	0,83(0,89)	**-1,03(-0,90)**	-0,18	0,00	0,14	**0,20**	**0,56**	**0,97**
17	0,66(0,95)	1,26(1,09)	0,21(0,23)	0,66(0,85)	1,26(1,42)	-0,85	0,00	0,36	-1,43	0,00	0,51
18	0,68(0,81)	-0,71(-0,45)	0,13(0,22)	**0,47(0,44)**	**-1,70(-1,46)**	**-0,87**	**-0,39**	**1,00**	**0,70**	**1,31**	**1,00**
19	1,16(1,24)	0,53(0,70)	0,34(0,37)	1,16(2,25)	**1,11(1,42)**	-0,18	1,55	0,34	**-1,20**	**-0,26**	**1,00**
20	0,95(0,85)	1,01(1,12)	0,12(0,11)	**0,75(0,76)**	1,01(1,12)	-0,48	0,00	0,40	-0,19	0,18	0,21

DIF values: simulated (estimated). Bold represents DIF items.

Anchor Item Set and $\pi^h_{ig} = 0$, if item i belong to this set. The remaining items may be imposed to have DIF ($\pi^h_g = 1$) or may allow to have DIF detected and possibly explained via regression. Soares et al. [2009] showed a superior performance in identification when the Fixed Anchor Item Set increases.

3. Imposing Prior Information about DIF Existence

Fixing one or more items to be anchor items ensures model identifiability and leads to good estimation is this assumption is correct. However, absolute certainty is rarely achieved in practice. It seems more reasonable to set as many non-DIF items

as possible. If there is some prior information, it may be included in the model when specifying the prior for π_g^h. Relevant prior information may also lead to model identification. This point will be illustrated in the results for the simulated data of Appendix B.

5 Comparison of methods

This section presents presents a simulation study to compare the performance of different procedures to detect DIF. This exercise was performed with a set of 30 databases generated from 3PL models with 2 groups with 1,000 respondents each. The average proficiency in the reference group was set to 0 for identification and average proficiency in the focal group were randomly chosen in the interval $[-1.5, 1.5]$.

Table 4: Results for 30 tests with 20 items each

		discrimination		difficulty		Error percentage
		number of items without DIF	number of items with DIF	number of items without DIF	number of items with DIF	
Real		600	0	469	131	
IRTLRDIF	Without DIF	506 (84.3%)	–	359 (76.5%)	22 (16.8%)	20%
	With DIF	94 (15.7%)	–	110 (23.5%)	109 (83.2%)	
MH	Without DIF	–	–	422 (90,0%)	39 (29,8%)	19.9%
	With DIF	–	–	47 (10.0%)	92 (70.2%)	
BILOG-MG	Without DIF	–	–	450 (95.9%)	40 (30.5%)	17.3%
	With DIF	–	–	19 (4.1%)	91 (69.5%)	
IBDM $\pi_{i2}^h = 0.5$	Without DIF	557 (92.8%)	–	426 (90.8%)	33 (25.1%)	17.1%
	With DIF	43 (7.2%)	–	43 (9.2%)	98 (74.9%)	
IBDM $\pi_{12}^b = 0.1$ $\pi_{i2}^h = 0.5, i > 1$	Without DIF	557 (92.8%)	–	438 (93.4%)	31 (23.7%)	15.2%
	With DIF	43 (7.2%)	–	31 (6.6%)	100 (76.3%)	

Soares et al. [2009] evaluate the methods by comparing the estimated values of DIF against their generated values, as suggested by Zwick et al. [1993]. Here, comparison is

based on the percentage of correct DIF detection in a effort to simplify the comparison and thus help the practitioner. This comparison is obviously affected by the significance levels chosen to declare that DIF was detected. The significance levels adopted for DIF detection were chosen as the default value in the software `IRTLRDIF v2.0b` and 0.05 for Mantel-Haenszel procedures and the cases that were run in `BILOG-MG`. An item was identified with DIF in IBDM when the posterior mean of π_{ig}^h was larger than 0.5.

The choice of the significance levels reflects a trade-off between sensitivity and specificity, making it difficult to define an universally accepted level. Assuming that sensitivity and specificity are equally important, the average percentages of misclassified items were 20%, 19.9%, 17.3%, 17.1% and 15.2%, respectively for methods IRTLRDIF, MH, BILOG-MG, IBDM with $\pi_{i2}^h = 0,5$ and IBDM with $\pi_{12}^b = 0,1$. This study shows some evidence of superior performance of IBDM when informative prior is assumed for a given item. Since this amount of prior information is usually encountered in real studies, it seems safe to recommend its use in practical studies. Appendix A details the code required to perform the analysis with these models with open source software.

6 Concluding remarks

There are a number of outstanding issues related to DIF that were not addressed in this chapter for conciseness. Some of them will be briefly touched upon in this Section.

There is clearly a connection between DIF and multidimensionality. Consider the responses of students paired according to their proficiencies but belonging to different groups. A DIF analysis aims at ascertaining whether their probability of correct answer remains the same and quantifying their diference otherwise. In the multidimensional case more than one cognitive ability is measured. It is likely that students paired according to one ability may not match at their other abilities, even when the selected ability is the most important. The responses will be different and the multidimensionality may be an obvious cause of DIF. One must verify whether the probability of correct answer remains

the same for students in different groups but paired according to all their proficiencies in these multidimensional settings. Thus, the presence of DIF may be an indication that an important dimension is not being considered. The DIF analysis may help unveiling the existence of other dimensions in these cases. The most common scenario is that DIF is caused by a special skill or knowledge, that is required for a subset of items. These special abilities commonly end up influencing the pattern of responses of items even when the test was not intended to be multidimensional. The SIB test is an example of procedures that analyze DIF under multidimensionality, considering these additional dimensions as nuisance.

The ideas of this chapter were restricted to a single factor for classification of groups. There are practical situations that may lead to the presence of many classification factors, such as type of school (public/private), race, socio-economic background, ... There are many possible models to be entertained here. The most complete is usually referred to as saturated model which considers a different DIF for all possible combination of the groups at each factor. This setup can be seen as equivalent to a single classification factor where each group consists on a given combination of the groups associated with the classification factor (eg white, middle-class, public schools pupils). Other model configurations are possible however if some or all of the interactions are removed. They lead to a more parsimonious model formulation. Different configurations can be tested against each other as performed in ANOVA tables. Gonçalves et al. [2013] provide the details of this extension.

It was clear from the presentation that the model is overparameterised and some form of additional information is required. The options within the realm of frequentist inference are inherently more limited and involve deterministic assumptions about some model components. The Bayesian approach allows for *lighter* restrictions in the form of probability statements. These can be seen as favoring (rather than imposing) a given set of restrictions. Thus, they may offer substantial improvements because data can overturn

prior information but will never be able to change a deterministic restriction. In any case, the problem is aggravated when different distributions are assumed for the proficiencies at each group. There is no simple solution to this problem and the most promising alternatives seem to be in the form of prior information but further research is clearly required on this subject.

Appendix A

OpenBUGS code

```
model{
    for (i in 1:I[1]) {
        theta[i, 1] ~ dnorm(0,1)
    }
    for (k in 2:K) {
        mu[k] ~ dnorm(0,1)
        tau[k] ~ dgamma(0.1, 0.1)
        sigma[k] <- 1/sqrt(tau[k])
        for (i in 1:I[k]) {
            theta[i, k] ~ dnorm(mu[k],tau[k]).
        }
    }
    for (j in 1:J) {
        a.geral[j] ~ dlnorm(0, 2.778)
          b.geral[j] ~ dnorm(0, 0.25)
          c[j] ~ dbeta(5, 17)
    }
    for (j in 1:J) {
        d.a[j,i.group[j, 1]] <- 0
        d.b[j,i.group[j, 1]] <- 0
        for (k in 2:n.group[j]) {
            Zb[j,i.group[j, k]] ~ dbern(piZb[j, i.group[j, k]])
            Za[j,i.group[j, k]] ~ dbern(piZa[j, i.group[j, k]])
            auxd.a[j,i.group[j, k]]~ dnorm(0, 1)
            auxd.b[j,i.group[j, k]]~ dnorm(0, 1)
            d.a[j,i.group[j, k]] <- Za[j,i.group[j, k]]*auxd.a[j,i.group[j, k]]
            d.b[j,i.group[j, k]] <- Zb[j,i.group[j, k]]*auxd.b[j,i.group[j, k]]
        }
    }
    for (j in 1:J) {
        for (k in 1:n.group[j]) {
            a[j,i.group[j,k]] <- a.geral[j]*exp(d.a[j,i.group[j,k]])
            b[j,i.group[j,k]] <- b.geral[j]-d.b[j,i.group[j,k]]
        }
    }
    for (k in 1:K) {
```

```
        for (i in 1:I[k]) {
             for (l in 1:n.item[G[i,k]]) {
                  Y[i,item.i[G[i,k],l],k] ~ dbern(p[i,item.i[G[i,k],l],k])
p[i,item.i[G[i,k],l],k] <- c[item.i[G[i,k],l]] +((1-c[item.i[G[i,k],l]])*
    *phi(a[item.i[G[i,k],l],k]*(theta[i,k]-b[item.i[G[i,k],l],k])))
             }
        }
    }
}


list(
I=c(2000, 2000), J=30, K=2,
Y= structure(
.Data= c(
   0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1,
   0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0,
   1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0,
   ...
   1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0,
   0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1),
   .Dim=c(2000, 30, 2)),
piZa= structure(
.Data= c(
   NA, 0.5,
   NA, 0.5,
   NA, 0.5,
   ...
   NA, 0.5,
   NA, 0.5,
   NA, 0.5),
   .Dim=c(30, 2)),
piZb= structure(
.Data= c(
   NA, 0.5,
   NA, 0.5,
   NA, 0.5,
   ...
   NA, 0.5,
   NA, 0.5,
   NA, 0.5),
   .Dim=c(30, 2)),
```

```
item.i= structure(
.Data= c(
   1, 2, 3,... , 28, 29, 30,
   1, 2, 3,... , 28, 29, 30),
   .Dim=c(2, 30)),
n.item=c(30, 30),
i.group= structure(
.Data= c(
   1, 2,
   1, 2,
   1, 2,
   ...
   1, 2,
   1, 2,
   1, 2),
   .Dim=c(30, 2)),
n.group=c(2, 2,..., 2, 2),
G= structure(
.Data= c(
   1,1,
   1,1,
   1,1,
   ...
   1,1,
   1,1,
   1,1)
.Dim=c(2000, 2))
)
```

Database details:

J: Number of items

K: Number of groups

I: Vector with number of respondents per group

piZa: Prior probability of DIF in parameter a for item i (π_{ig}^{a})

piZb: Prior probability of DIF in parameter b for item i (π_{ig}^{b})

i.group: Matrix with each row corresponding to an item and each column to the group it belongs to. First row is reserved for the reference group in case of DIF. Remaining blank

slots must be filled with *NA*.

n.group: Vector of length J with number of groups that responded each item. Its elements are equivalent to the valid rows of matrix i.group.

item.i: Matrix in which each row corresponds to a form of a test. Rows are filled with the number of items that the different formats were presented.

n.item: Vector indicating how many items were presented in each form.

G: Matrix where each column corresponds to a group and each row corresponds to a respondent from this group. It indicates the form of the test taken by each respondent.

Appendix B

This is an example with simulated data on a test containing 20 items. Table 4 below presents the results with the non-informative prior $\pi_{ig}^{h} = 0.5$ for all items. The model identified almost all items as presenting DIF and the average proficiency for the focal group was substantially underestimated.

Table 5: $\pi_{ig}^{h} = 0.5$

item	Reference group			Focal group		DIF detection for a			DIF detection for b		
						Credibility limits		π_{i2}^{a}	Credibility limits		π_{i2}^{b}
	a	b	c	a	b	5%	95%		5%	95%	
1	1,26(1,30)	-0,40(-0,35)	0,11(0,20)	1,26(1,58)	-0,40(-0,99)	0,00	0,45	0,48	**0,58**	**0,72**	**1,00**
2	1,41(1,32)	-0,13(-0,23)	0,25(0,19)	1,41(1,29)	-0,13(-1,00)	-0,33	0,00	0,15	**0,56**	**0,94**	**1,00**
3	1,67(1,41)	-1,36(-1,40)	0,14(0,15)	1,67(1,41)	-1,36(-2,03)	0,00	0,00	0,00	**0,43**	**0,78**	**1,00**
4	0,59(0,62)	0,05(0,03)	0,24(0,23)	0,59(0,55)	0,05(-0,67)	-0,75	0,00	0,32	**0,32**	**1,00**	**1,00**
5	1,90(1,68)	0,44(0,39)	0,25(0,23)	1,90(1,86)	0,44(-0,38)	-0,06	0,89	0,20	**0,54**	**0,95**	**1,00**
6	0,98(0,85)	0,29(-0,07)	0,39(0,33)	0,98(0,83)	0,29(-0,24)	-0,41	0,04	0,18	0,00	0,54	0,46
7	0,92(0,93)	-0,10(-0,25)	0,22(0,18)	0,92(0,91)	-0,10(-0,86)	-0,25	0,09	0,16	**0,35**	**0,79**	**1,00**
8	1,54(0,91)	1,30(1,30)	0,26(0,23)	1,54(0,80)	1,30(0,76)	-0,71	0,10	0,44	**0,00**	**1,08**	**0,83**
9	1,19(1,15)	1,43(1,76)	0,33(0,35)	1,19(1,16)	1,43(0,93)	-0,56	0,26	0,13	**0,00**	**1,30**	**0,95**
10	1,07(0,92)	1,01(1,08)	0,19(0,20)	1,07(0,92)	1,01(0,44)	-0,23	0,11	0,11	**0,00**	**0,96**	**0,92**
11	1,59(1,23)	1,47(1,64)	0,31(0,31)	1,59(1,87)	**2,98(1,82)**	-0,85	1,63	0,30	-1,58	0,40	0,39
12	1,39(1,47)	-0,11(-0,09)	0,28(0,28)	1,39(1,42)	**0,85(-0,02)**	-0,46	0,09	0,22	-0,51	0,01	0,30
13	1,84(1,31)	-0,78(-1,13)	0,39(0,23)	1,84(1,32)	-0,78(-1,79)	-0,06	0,12	0,10	**0,48**	**0,83**	**1,00**
14	1,18(0,90)	1,29(1,21)	0,28(0,23)	1,18(0,64)	1,29(1,14)	-1,25	0,00	0,62	-1,46	0,81	0,59
15	1,08(0,57)	1,86(2,29)	0,23(0,18)	1,08(0,55)	**1,65(1,75)**	-0,55	0,37	0,36	-0,15	1,50	0,72
16	1,05(0,78)	-0,29(-0,73)	0,35(0,21)	1,05(0,78)	-0,29(-1,34)	-0,11	0,03	0,07	**0,35**	**0,85**	**1,00**
17	1,09(0,90)	-1,85(-2,01)	0,21(0,27)	1,09(0,90)	-1,85(-2,58)	0,00	0,00	0,00	**0,36**	**0,76**	**1,00**
18	2,11(2,17)	0,28(0,28)	0,13(0,13)	2,11(2,84)	**1,18(0,50)**	0,00	0,60	0,65	-0,59	0,00	0,71
19	1,32(1,31)	-1,58(-1,70)	0,23(0,19)	1,32(1,04)	-1,58(-2,48)	-0,79	0,00	0,48	**0,48**	**1,25**	**1,00**
20	0,99(0,89)	-1,09(-1,34)	0,23(0,18)	0,99(0,92)	**-1,54(-2,40)**	0,00	0,30	0,23	**0,83**	**1,29**	**1,00**

DIF values: simulated (estimated). Bold represents DIF items.

Average proficiency of the focal groups: -0.522 (-1.225)

Analysis was repeated with the same data sets and same prior but for item 1 where it

was assumed now that $\pi^b_{12} = 0.1$, giving a small (but positive) probability for the presence of DIF in the difficulty parameter for this item. Table 5 presents the estimation results. These are clearly better results than those obtained in Table 4 with non-informative priors. These results show that stronger and typically unverified restrictions of fixing items as anchor ones are not needed. Use of appropriate prior distributions avoids their use while providing good recovery of all model parameters, including DIF.

Table 6: $\pi^b_{12} = 0.1$

item	Reference group			Focal group		DIF detection for a			DIF detection for b		
						Credibility limits		π^a_{i2}	Credibility limits		π^b_{i2}
	a	b	c	a	b	5%	95%		5%	95%	
1	1,26(1,31)	-0,40(-0,39)	0,11(0,16)	1,26(1,32)	-0,40(-0,38)	0,00	0,13	0,05	-0,14	0,00	0,07
2	1,41(1,34)	-0,13(-0,23)	0,25(0,21)	1,41(1,32)	-0,13(-0,25)	-0,28	0,02	0,14	0,00	0,18	0,18
3	1,67(1,45)	-1,36(-1,36)	0,14(0,14)	1,67(1,45)	-1,36(-1,36)	0,00	0,00	0,03	-0,04	0,01	0,07
4	0,59(0,65)	0,05(0,07)	0,24(0,24)	0,59(0,55)	0,05(0,06)	-0,71	0,00	0,45	-0,08	0,21	0,14
5	1,90(1,69)	0,44(0,36)	0,25(0,22)	1,90(1,70)	0,44(0,35)	-0,22	0,24	0,15	0,00	0,19	0,15
6	0,98(0,92)	0,29(0,07)	0,39(0,35)	0,98(0,90)	0,29(0,38)	-0,39	0,14	0,17	**-0,65**	**0,00**	**0,87**
7	0,92(0,95)	-0,10(-0,21)	0,22(0,18)	0,92(0,90)	-0,10(-0,20)	-0,44	0,00	0,25	-0,12	0,00	0,09
8	1,54(0,92)	1,30(1,33)	0,26(0,24)	1,54(0,88)	1,30(1,34)	-0,51	0,03	0,22	-0,33	0,21	0,20
9	1,19(1,24)	1,43(1,64)	0,33(0,35)	1,19(1,31)	1,43(1,59)	-0,41	0,65	0,19	-0,10	0,52	0,25
10	1,07(0,98)	1,01(1,10)	0,19(0,21)	1,07(0,99)	1,01(1,10)	-0,12	0,20	0,13	-0,13	0,16	0,14
11	1,59(1,22)	1,47(1,68)	0,31(0,31)	1,59(2,11)	**2,98(2,39)**	-0,63	2,12	0,33	**-2,23**	**0,00**	**0,83**
12	1,39(1,51)	-0,11(-0,06)	0,28(0,28)	1,39(1,44)	**0,85(0,71)**	-0,50	0,05	0,23	**-1,03**	**-0,56**	**1,00**
13	1,84(1,37)	-0,78(-1,06)	0,39(0,26)	1,84(1,37)	-0,78(-1,06)	-0,08	0,09	0,09	-0,06	0,06	0,09
14	1,18(1,05)	1,29(1,35)	0,28(0,28)	1,18(1,21)	1,29(1,37)	-0,34	0,96	0,25	-0,34	0,10	0,19
15	1,08(0,56)	1,86(2,34)	0,23(0,18)	1,08(0,56)	**1,65(2,31)**	-0,16	0,09	0,10	-0,21	0,45	0,23
16	1,05(0,80)	-0,29(-0,70)	0,35(0,19)	1,05(0,78)	-0,29(-0,70)	-0,30	0,01	0,16	-0,11	0,05	0,10
17	1,09(0,89)	-1,85(-2,00)	0,21(0,24)	1,09(0,88)	-1,85(-1,99)	-0,11	0,00	0,07	-0,19	0,05	0,14
18	2,11(1,96)	0,28(0,27)	0,13(0,13)	2,11(5,59)	**1,18(1,26)**	-0,02	3,18	0,40	**-1,26**	**-0,74**	**1,00**
19	1,32(1,19)	-1,58(-1,72)	0,23(0,20)	1,32(1,06)	-1,58(-1,78)	-0,58	0,00	0,36	-0,11	0,49	0,28
20	0,99(0,87)	-1,09(-1,34)	0,23(0,19)	0,99(0,91)	**-1,54(-1,71)**	0,00	0,41	0,17	**0,00**	**0,62**	**0,92**

DIF values: simulated (estimated). Bold represents DIF items.

Average proficiency of the focal groups: -0.522 (-0.550)

Bibliografia

N. L. Allen and J. R. Donoghue. Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement*, 33:231–251, 1996.

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

R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters. *Psychometrika*, 46:443–459, 1981.

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

J. Donoghue, P. W. Holland, and D. T. Thayer. *Differential Item Functioning*, chapter A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning, pages 137–166. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

A. M. Fidalgo and J. M. Madeira. Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, 68:940–958, 2008.

J. P. Fox and A. J. Verhagen. *Cross-cultural Analysis: Methods and Applications*, chapter Random item effects modeling for cross-national survey data, pages 467–488. Routeledge Academic, London, 2010.

A. W. French and T. R. Miller. Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33:315–332, 1996.

D. Gamerman and H. L. Lopes. *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. Taylor and Francis, 2nd edition, 2006.

M. J. Gierl, J. Bisanz, G. Bisanz, and K. Boughton. 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, 2003.

F. B. Gonçalves, D. Gamerman, and T. M. Soares. Simultaneous multifactor DIF analysis and detection in Item Response Theory. *To appear in Computational Statistics and Data Analysis*, 2013.

R. K. Hambleton and H. J. Rogers. Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education*, 2:313–334, 1989.

P. W. Holland and D. T. Thayer. An alternative definition of the ets delta scale of item difficulty. Research Report RR-85-43, Educational Testing Service, Princeton, NJ, 1985.

P. W. Holland and D. T. Thayer. *Test validity*, chapter Differential item performance and the Mantel-Haenszel procedure, pages 129–145. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28:3049–3067, 2009.

N. Mantel. Chi-square tests with one degrees of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58:690–700, 1963.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

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

T. R. Miller and J. A. Spray. Logistic discrimant function analysis for dif identification of polytomously scored items. *Journal of Educational Measurement*, 30:107–122, 1993.

K. A. O'Neil and W. M. McPeek. *Differential item functioning*, chapter Item and test characteristics that are associated with differential item functioning, pages 255–276. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

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

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

L. A. Roussos and W. F. Stout. Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33:215–230, 1996.

F. Samejima. *Handbook of modern item response theory*, chapter Graded response model, pages 85–100. Springer Verlag, New York, 1997.

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

R. T. Shealy and W. F. Stout. *Differential item functioning*, chapter An item response theory model for test bias and differential test functioning, pages 197–239. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993a.

R. T. Shealy and W. F. Stout. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58:159–194, 1993b.

T. M. Soares, F. B. Gonçalves, and D. Gamerman. An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, 34(3):348–377, 2009.

W. F. Stout. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrica*, 55:293–325, 1990.

H. Swaminathan and H. J. Rogers. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27:361–370, 1990.

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

D. Thissen. *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill, 2001. Computer software.

D. Thissen, L. Steinberg, and H. Wainer. *Differential Item Functioning*, chapter Detection of differential Item Functioning using the parameters of item response models, pages 67–114. Lawrence Erlbaum, Hillsdale, NJ, 1993.

T. Uttaro and R. E. Millsap. Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, 18:15–25, 1994.

H. Wainer. *Differential Item Functioning*, chapter Model-Based Standardized Measure-

ment of an Item's Differential Impact, pages 123–135. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

W.-C. Wang and Y.-H Su. 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, 2004.

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

R. Zwick, J. R. Donoghue, and A. Grima. Assessment of differential item for performance tasks. *Journal of Educational Measurement*, 30:233–251, 1993.

R. Zwick, D. Thayer, and C. Lewis. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36:1–28, 1999.