Brazilian Journal of Probability and Statistics
2019, Vol. 33, No. 4, 782–800
https://doi.org/10.1214/19-BJPS443

Bayesian modelling of the abilities in dichotomous IRT models via regression with missing values in the covariates

Flávio B. Gonçalves and Bárbara C. C. Dias

Universidade Federal de Minas Gerais

Abstract. Educational assessment usually considers a contextual questionnaire to extract relevant information from the applicants. This may include items related to socio-economical profile as well as items to extract other characteristics potentially related to applicant's performance in the test. A careful analysis of the questionnaires jointly with the test's results may evidence important relations between profiles and test performance. The most coherent way to perform this task in a statistical context is to use the information from the questionnaire to help explain the variability of the abilities in a joint model-based approach. Nevertheless, the responses to the questionnaire typically present missing values which, in some cases, may be missing not at random. This paper proposes a statistical methodology to model the abilities in dichotomous IRT models using the information of the contextual questionnaires via linear regression. The proposed methodology models the missing data jointly with the all the observed data, which allows for the estimation of the former. The missing data modelling is flexible enough to allow the specification of missing not at random structures. Furthermore, even if those structures are not assumed a priori, they can be estimated from the posterior results when assuming missing (completely) at random structures a priori. Statistical inference is performed under the Bayesian paradigm via an efficient MCMC algorithm. Simulated and real examples are presented to investigate the efficiency and applicability of the proposed methodology.

1 Introduction

The usual approach in educational assessments consists in applying a test in which the responses to items in the test are dominantly determined by the latent trait(s) (ability) one is interested in measuring. This is usually done via Item Response Theory (IRT) models which relate the probability of a given response to the applicants' ability and to the item's characteristics. In many cases, however, the applicants also complete a contextual questionnaire to obtain their pedagogic-socio-economical profile. These questionnaires may even be applied in other levels such as teachers and schools.

Data extracted from the questionnaires may be used to produce descriptive statistics to help understand the profile of the applicants, teachers and schools involved in the assessment or even to report the results in the test by profile groups.

Key words and phrases. 3PNO model, Bayesian inference, MCMC.
Received October 2018; accepted April 2019.

A deeper investigation, however, may reveal relevant association between profiles and the performance in the test (represented by the respective abilities). This is typically done by fitting linear regression models with the estimated ability as the response variable being explained by the covariates from the questionnaire (see, for example, da Silva Fernandes et al., 2010; Laros and Marciano, 2008; Soares and Alves, 2003; Alves et al., 2015). This kind of approach may be compromised by the fact that significant uncertainty is ignore when using only the point estimate of the abilities (see Zwinderman, 1991).

It is common to have a considerable percentage of missing data in the contextual questionnaires and there is usually no explicit missing pattern. This implies that, in order to remove the missing data from the analysis, a significant percentage of the applicants would have to be removed, which means loosing a lot of useful information. Moreover, if a missing not at random process exists, the results obtained from the reduced dataset could be seriously misleading. It is then crucial to model the missing data jointly with the observed data (responses to the test and questionnaire).

This paper proposes a model-based approach to simultaneously fit a dichotomous IRT model and explain the abilities through covariates from the contextual questionnaires, allowing for the presence of missing data in the covariates. The fully Bayesian approach allows for the quantification of all sources of uncertainty. A flexible modelling structure is able to capture a variety of missing patterns. Furthermore, even if a missing (completely) at random structure is assumed, the results from the posterior distribution allow for the estimation of missing not at random structures, as long as the respective covariate is related to the ability being measured. The relation between the ability of the applicants and the covariates is modelled via linear regression on the second level of the model which has any standard dichotomous IRT model on the first level. We particularly explore the 3PNO model.

The regression approach to model the abilities has been previously considered in the literature. For example, Zwinderman (1991) considers this approach for the abilities in the 1PL model. In a more general approach, Fox (2005) proposes a multilevel linear regression modelling of the abilities to account for covariates from the questionnaires applied to different levels (student, school, etc.). The author considers the 2PNO model for dichotomous and polytomous responses and performs Bayesian inference via MCMC. However, due to the difficulty to fix the abilities' scale in a multilevel regression context, the author adopts a particular strategy inside the MCMC by standardasing the values of the abilities sampled on each iteration of the algorithm, which implies in a non-fully Bayesian approach.

A general approach to account for modelling both the abilities via covariates in a simultaneous setup as well as the missing data in the covariates has, to the best of our knowledge, never been considered before in the literature.

This paper is organised as follows. Section 2 presents the proposed Bayesian model, which includes the 3PNO model on the first level, the linear regression

to explain the abilities on the second level and the missing structure on the third one. Section 3 describes the MCMC algorithm used to explore the posterior distribution. Simulated and real examples are presented in Section 4. Some extensions of the regression model for the abilities is presented in Section 5. This includes multilevel modelling, latent covariates, graded response models and more general structures for the regression errors. Finally, Section 6 brings some final remarks.

2 Proposed model

We propose a Bayesian model that basically combines three components. First, the 3PNO model for dichotomous items, second, a regression structure to the abilities using covariates from the contextual questionnaires, and third, a joint model for the covariates which will account for modelling and inference of the missing data. Note that the data consist of the responses to the dichotomous items and a portion of the covariate values.

For a dataset with I items and J individuals, let Y_{ij} be the indicator variable of individual j correctly responding item i. The proposed model is given by

$$(Y_{ij} = 1|\theta_j, a_i, b_i, c_i) \sim \text{Ber}\big(c_i + (1 - c_i)\Phi(a_i\theta_j - b_i)\big), \tag{2.1}$$

$$\theta_j \overset{\text{ind.}}{\sim} N\big(X_{j\cdot}\beta, \sigma_e^2\big), \tag{2.2}$$

$$X \sim \pi(X), \tag{2.3}$$

where a_i, b_i and c_i are the discrimination, difficulty and guessing parameters of item i, respectively, and θ_j is the ability of individual j; $\Phi(\cdot)$ is the standard normal c.d.f., $\beta = (\beta_0, \beta_1, \ldots, \beta_Q)'$ are the regression coefficients, $\mathbf{X}_{j\cdot}$ is the jth row of the design matrix X, containing the covariates from the questionnaries. Furthermore, $\pi(X)$ is the prior distribution of the covariates, to be discussed further ahead in the text.

The model above is not identifiable as the scale of the abilities is not specified. Identifiability is achieved by setting $\beta_0 = 0$ and $\sigma_e^2 = 1$.

2.1 Modelling the covariates and their missingness

In order to define the joint distribution $\pi(X)$ of the covariates, we set $X = (X_{\text{obs}}, X_{\text{mis}})$, where X_{obs} and X_{mis} represent the observed and missing values of the covariates, respectively. Recall that X is a $J \times Q$ matrix with X_{jq} being the response given by individual j to question q, which may or may not be missing. We define another $J \times Q$ matrix R as

$$R_{jq} = \begin{cases} 1 & \text{if } X_{jq} \text{ is missing;} \\ 0 & \text{if } X_{jq} \text{ is not missing.} \end{cases}$$

Matrix R describes the missing pattern and, depending on the case, ought to be modelled jointly with X. In particular, (Rubin, 1976) classifies the missingness process into three categories:

- *MCAR (missing completely at random)*: R is independent of $(X_{\text{mis}}, X_{\text{obs}})$.
- *MAR (missing at random)*: R is not independent of X_{obs}, but it is independent of X_{mis}.
- *MNAR (missing not at random)*: R is not independent of X_{mis}.

If we assume that the distribution of R has no common parameters with the remainder of the model and those two sets of parameters are independent *a priori*, the missingness process may be ignored to perform inference in the first two cases above. Under the Bayesian approach, this is clear by looking at the following equation. Suppose, for a moment, that X_{obs} and X_{mis} represent all the observed and missing data, respectively, from the model under consideration. Now let ϕ be the set of parameters indexing the distribution of R and let θ be all the other parameters in the model. We have that

$$\begin{aligned}
&\pi(X_{\text{mis}}, \theta, \phi | X_{\text{obs}}, R) \\
&\quad \propto \pi(R | X_{\text{mis}}, \theta, \phi, X_{\text{obs}}) \pi(\phi) \pi(X_{\text{mis}}, \theta | X_{\text{obs}}) \\
&\quad = \pi\big(R | \phi, (X_{\text{obs}})\big) \pi(\phi) \pi(X_{\text{mis}}, \theta | X_{\text{obs}}),
\end{aligned} \tag{2.4}$$

where the last equality is obtained under a MCAR or MAR scenario.

Equation (2.4) states that the posterior distribution of (X_{mis}, θ) is independent of R and, therefore, the latter can be ignored in the inference process. We assume throughout this paper that the missingness process is MCAR or MAR. Nevertheless, it is straightforward to extend our results for the MNAR case, whenever information is available to suitably model R. Furthermore, one may infer about possible MNAR structures based on the missing data estimates provided by our methodology. For example, substantial differences between $(X|R = 0)$ and $(X|R = 1)$ indicate the existence of a MNAR structure.

Different dependence structures may be considered when specifying the joint distribution of X. In particular, they may differ for different questions in the questionnaire. We consider the following possibilities:

$$X_{jq} \sim \begin{cases} \pi_\varphi(\cdot), & \\ \pi_\varphi(\cdot | X_{jq^*}), & q^* \neq q, \\ \pi_\varphi(\cdot | X_{j^*q}), & j^* \neq j, \\ \pi_\varphi(\cdot | X_{jq^*}, X_{j^*q}), & q^* \neq qej^* \neq j, \end{cases}$$

where φ are possible unknown parameters indexing the distributions. In the first specification above, X_{jq} is independent from any other response. A reasonable example would be modelling the indicator variable of the response male in a gender question as a $\text{Ber}(p^*)$ r.v. For the second specification, the distribution of X_{jq} depends on the responses of the same individual to the other questions. For example, questions like "family income" and "type of school" would reasonably admit such a dependence. For the third specification, the distribution of X_{jq} depends on

the responses given to the same question by other students. Finally, in the fourth specification, the two types of dependence are present.

In order to specify the joint distribution of X, one should consider the probabilistic features of the covariates—categorical, latent or not; the information from specialists, if this is available; and previous studies.

3 Bayesian inference

Model specification and inference is performed under the Bayesian approach. The full model specification combines the model proposed in the previous section with the prior distribution of the remaining unknown quantities. Standard prior distributions are adopted in a way to facilitate the computation in the MCMC algorithm. In particular, we set:

$$(a, b)' \sim N_2\big((\mu_a, \mu_b)', \operatorname{diag}(\sigma_a^2, \sigma_b^2)\big), \tag{3.1}$$

$$c \sim \text{Beta}(\alpha_c, \beta_c), \tag{3.2}$$

$$\beta \sim N_Q(\mu_\beta, \Sigma_\beta), \tag{3.3}$$

$$\varphi \sim \cdots, \tag{3.4}$$

where the prior for φ is suitably chosen in a case-by-case basis.

For computational reasons, we introduce two sets of auxiliary variables, as proposed in Gonçalves, Dias and Soares (2018), which allow us to sample directly from all the full conditional distributions of the Gibbs sampler to be devised. Define Z_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, where $Z_{ij} \sim \text{Bernoulli}(c_i)$, and V_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, where $(V_{ij}|Z_{ij} = 0) \sim N(a_i\theta_j - b_i, 1)$ and $P(V_{ij} = 0|Z_{ij} = 1) = 1$. We get that:

$$Y_{ij} = \begin{cases} 1 & \text{if } (V_{ij} = 0, Z_{ij} = 1) \text{ or } (V_{ij} \geq 0, Z_{ij} = 0); \\ 0 & \text{if } (V_{ij} < 0, Z_{ij} = 0). \end{cases}$$

Note that this preserves the original marginal model for the data Y.

Under the Bayesian paradigm, inference is based on the posterior distribution of all the unknown quantities of the model, defined as $\psi = (Z, V, a, b, c, \theta, \beta, X_{\text{mis}}, \varphi)$, omitting the respective indexes for cleanness of notation. The posterior density of ψ is given by

$$\begin{aligned} \pi(\psi|\cdot) \propto & \prod_{i=1}^{I}\prod_{j=1}^{J} \pi(Y_{ij}|V_{ij}, Z_{ij})\pi(Z_{ij}|c_i)\pi(V_{ij}|Z_{ij}, a_i, b_i, \theta_j) \\ & \times \prod_{i=1}^{I} \pi(a_i)\pi(b_i)\pi(c_i) \prod_{j=1}^{J} \pi(\theta_j|X_{j\cdot}, \beta)\pi(X|\varphi)\pi(\beta)\pi(\varphi). \end{aligned} \tag{3.5}$$

This is a complex and highly dimensional distribution which cannot be analytically explored. Instead, we draw from this distribution via MCMC and compute Monte Carlo (MC) estimates to explore its properties, like means, variances, quantiles, marginal densities, etc. We propose a Gibbs sampling algorithm with blocks

$$(Z, V), c, (a, b), \theta, \beta, X_{\text{mis}}, \varphi.$$

It is feasible to sample directly from all the full conditional distributions. The algorithms to do so are presented in the Appendix.

4 Examples

We present some simulated and real examples. The simulated example illustrates the efficiency of the proposed inference methodology to recover the parameters and missing values. Several datasets from two large scale educational assessment exams in Brazil—Saeb and Enem, are analysed to illustrate the applicability of the proposed methodology. We highlight the possibility to investigate possible MNAR structures in the missingness process.

4.1 Simulated example

We consider 5000 thousand individuals responding 30 items each. The contextual questionnaire contains 3 items, in which item 1 has three alternatives and items 2 and 3 have two alternatives.

The real values of the item parameters are drawn from $U(0.5, 3)$, $U(-9, 8)$ and $U(0, 0.15)$, for a, b and c, respectively. We also set $\beta = (-2, 2.5, -2.5, 2)'$ and X as a binary matrix such that the first two columns refer to the first item and the other two columns to items 2 and 3, respectively. Parameters c_i were fixed in their respective real values for the analysis.

The missing values were randomly chosen to have around 10% of missingness. As a result, we had 503 missing responses from 487 individuals. The observed responses were uniformly chosen among the alternatives in each of the three items.

The fitted model assumes Bernoulli and multinomial priors for the responses of the questions with two and three alternatives, respectively, with uniform priors (Beta(1,1) and Dihichlet(1,1,1)) for the respective probability vectors.

The MCMC chain runs for 20 thousand iterations with a burn-in of 5 thousand. The following priors are used: $a_i \sim N_{(0,\infty)}(1, 2^2)$, $b_i \sim N(0, 4^2)$, $\forall i = 1, \ldots, 30$, $\beta \sim N_4(0, \text{diag}(10{,}000))$.

Figure 1 shows the good recovery of the item parameters and abilities. Figure 2 shows the posterior probability of the real response for each of the missing values and highlights the ones correctly identified by the posterior mode. Finally, Table 1 shows the estimated proportions to each alternative of each item.

The posterior mean of the regression coefficients β are -2.052, 2.523, -2.502 and 1.999, respectively, with standard deviations 0.0425, 0.0507, 0.0392 and 0.0408. All the results suggest that the proposed methodology is efficient to recover the unknown quantities of the models.

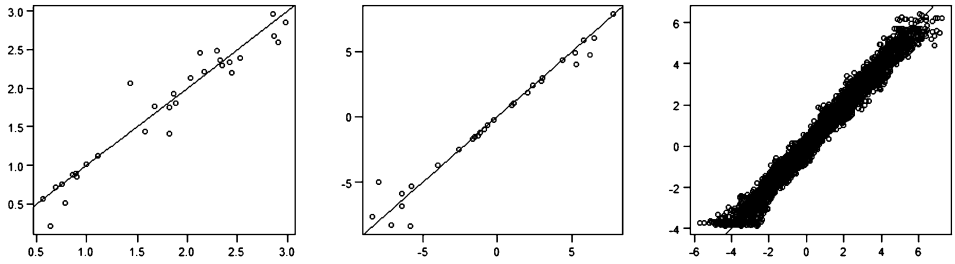

Figure 1 *Real* (x*-axis*) *versus estimate* (y*-axis—posterior mean*) *of the item parameters and abilities. From left to right*: a, b *and* θ.

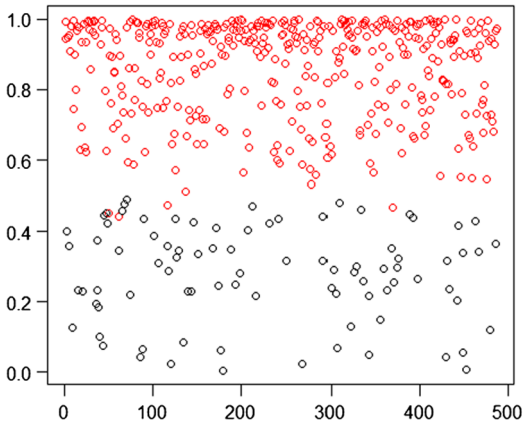

Figure 2 *Posterior probability of the real response for each of the missing values. Values in red refer to the responses correctly identified by the posterior mode.*

Table 1 *Real and estimated proportions of the missing responses*

Item	Real	Estimated
1	(0.306, 0.341, 0.353)	(0.331, 0.344, 0.325) − 173
2	(0.545, 0.455)	(0.558, 0.442) − 176
3	(0.442, 0.558)	(0.476, 0.524) − 154

4.2 Applications

We apply the proposed methodology to datasets from two large scale educational assessment exams in Brasil. The *Sistema Nacional de Avaliação da Educação Básica* (Saeb) assesses the basic educational system in Brazil. It consists of Portuguese and Mathematics exams applied every other year to students in 5th and 9th year of Elementary School (ES) and 3rd year of High School (HS). We analyse the datasets from Saeb 2015 shown in Table 2.

Table 2 *Datasets analysed*

Grade	Exams	Location
5th (ES)	Port/Math	Minas Gerais state
9th (ES)	Port/Math	Bahia state
9th (ES)	Port	Pará state
3rd (HS)	Port/Math	Brazil

Table 3 *Covariates used in at least one of the analysis. Options* (*a*) *and* (*b*) *of gender where swaped in the analysis of Enem*

Cavariate	Description	Alternatives
Gender	What is your gender?	(a) Male (b) Female
Ethnicity	How do you consider yourself?	(a) Black or Brown (b) Indigenous or Yellow (c) White
LikeMaths	Do you like to study Maths?	(a) Yes (b) No
Talk	Do your parents talk to you about school?	(a) Yes (b) No
Work	Do you work out of your home?	(a) Yes (b) No
PC	Do you have a computer at home?	(a) No (b) Yes, 1 (c) Yes, 2 or more
Net	Do you have internet access at home?	(a) No (b) Yes
Encour	Do your parents encourage you to study?	(a) Yes (b) No
Read	Can you mother read and write?	(a) Yes (b) No

Table 4 *Covariates used in each analysis*

Exame	Grade	Exams	Population	Covariates
Saeb	5°	Math/Port	Minas Gerais state	Ethnicity, Talk, Work
Saeb	5°	Port	Pará state	Ethnicity, Read, PC
Saeb	9°	Math/Port	Bahia state	PC, Work, Encour
Saeb	3°	Math/Port	Brazil	Ethnicity, Gender, LikeMaths
Enem		Math	Belo Horizonte city	Ethnicity, Gender, Net

The *Exame Nacional do Ensino Médio* (Enem) is annually applied to students in the 3rd year of High School or who have finished it and is used as an admission criterion by most of the universities in Brasil. We consider the Math test applied in 2015 and restrict the analysis to applicants from Belo Horizonte city.

Several items from the respective contextual questionnaires are considered in the analysis. They are presented in Table 3 and assigned as shown in Table 4.

Since all the covariates are categorical, they are introduced in the regression model using dummy variables, with the first one referring to option (a) and the second one to option (b) (when there are three options). The interpretation of the

Table 5 *Observed and estimated proportions for each response with the respective number of responses. The estimated refers to the proportions of the responses that were estimated for the missing cases*

Covariate	Observed	Estimated
Ethnicity	$(0.634, 0.068, 0.298) - 5167$	$(0.459, 0.248, 0.293) - 896$
Talk	$(0.814, 0.186) - 5679$	$(0.633, 0.367) - 384$
Work	$(0.104, 0.896) - 5597$	$(0.473, 0.527) - 466$

results ought to take into account the scale of the regression model, that is, unit variance error.

In all the analyses considering both the Portuguese and Maths test, the first was fit separately under the 3PNO model with scale $N(0, 1)$ and the estimates (posterior mean) of the abilities used as covariates in the regression model to explain the the ability in Math. A more elaborated analysis should consider the joint modelling of the two abilities to fit both tests jointly as it is proposed in the extension presented in Section 5.2. The implication of adopting the first approach is that the uncertainty about the ability in Portuguese is ignored when estimating the ability in math and, therefore, the uncertainty about the relation between the two abilities is underestimated. Nevertheless, the analysis presented here is still useful to investigate this relation.

The analysis regarding missing patterns is performed by comparing the observed and estimated proportions of each response. We use the posterior mean of the proportions as an estimate. Given that the relation between the respective question and the abilities is the only source of information about missing patterns, conclusions about possible MNAR structures can only be drawn for question with significant regression coefficients. In those cases, larger differences between observed and estimated proportions indicate the presence of a MNAR structure.

The regression coefficients are ordered according to the order that the respective covariates are presented in Table 4. For questions with two options, the respective covariate is the indicator of alternative (a). For questions with three options, the respective first covariate is the indicator of alternative (a) and the second one is the indicator of alternative (b).

4.2.1 *Saeb*, *5th grade*, *Math/Port*, *Minas Gerais*. The dataset consists of 6063 students and each test has 11 items. 1070 students did not respond one of the three questions in the questionnaire, 206 did not respond two of them and 88 all of them, leading to a total of 1746 missing values. In particular 896, 384 and 466 did not answer the first, second and third questions, respectively. Results are shown in Tables 5 and 6 and Figure 3.

Results suggest a possible MNAR structure associated to the first and third questions, in which Indigenous or Yellow students would be more likely not to declare

Table 6 *Estimates of the regression coefficients—posterior mean and* 99% *credibility interval*

Coefficient	Estimate
β_{11}	−0.198 (−0.313, −0.078)
β_{12}	−0.218 (−0.440, −0.0004)
β_2	0.047 (−0.076, 0.164)
β_3	−0.301 (−0.478, −0.130)
β_4	1.082 (1.003, 1.167)

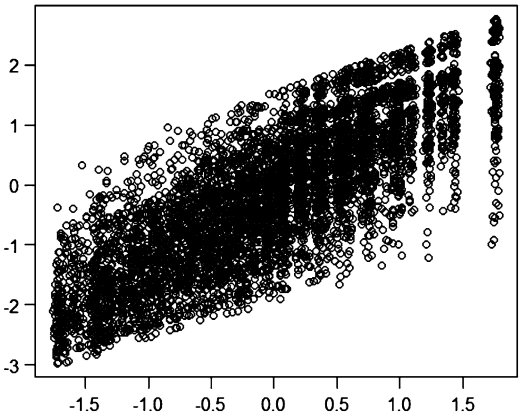

Figure 3 *Estimated abilities in Portuguese* (x*-axis*) *and Mathematics*. *The correlation is* 0.804.

their ethnicity, followed by White students and then Black and Brown students. Also, students who work are more likely not to declare their work status than students who do not work. The fitted regression model indicates a positive association between the abilities in Portuguese and Mathematics. It also indicates a lightly worse performance by non-white students and by student who work. The regression coefficient of the covariate "Talk" is the least significant one, therefore, the difference between observed and estimated proportions in Table 5 only weakly suggests that students who do not talk to their parents about school are more likely not to answer that question.

4.2.2 *Saeb*, 5*th grade*, *Port*, *Pará*. The dataset consists of 5395 students and 11 items. 918 students did not respond one of the three questions in the questionnaire, 145 did not respond two of them and 104 all of them, leading to a total of 1520 missing values. In particular 864, 352 and 304 did not answer the first, second and third questions, respectively. Question PC was dichotomised by merging options (b) and (c). Results are shown in Tables 7 and 8.

Results do not suggest the presence of a MNAR structure in any of the three questions. The fitted regression model indicates a lightly better performance by

Table 7 *Observed and estimated proportions for each response with the respective number of responses. The estimated refers to the proportions of the responses that were estimated for the missing cases*

Covariate	Observed	Estimated
Ethnicity	$(0.761, 0.068, 0.194) - 4531$	$(0.757, 0.045, 0.198) - 864$
Read	$(0.934, 0.066) - 5043$	$(0.927, 0.073) - 352$
PC	$(0.644, 0.356) - 5091$	$(0.654, 0.346) - 304$

Table 8 *Estimates of the regression's coefficients—posterior mean and 99% credibility interval*

Coefficient	Estimate
β_{11}	0.193 (0.072, 0.315)
β_{12}	0.201 (−0.046, 0.452)
β_2	0.652 (0.466, 0.830)
β_3	−0.231 (−0.329, −0.137)

Table 9 *Observed and estimated proportions for each response with the respective number of responses. The estimated refers to the proportions of the responses that were estimated for the missing cases*

Covariate	Observed	Estimated
PC	$(0.972, 0.027, 0) - 6035$	$(0.486, 0.179, 0.335) - 144$
Work	$(0.152, 0.847) - 5830$	$(0.359, 0.641) - 349$
Encour	$(0.982, 0.018) - 5964$	$(0.718, 0.282) - 215$

non-white students and a considerably better performance by students whose mother can read and write. A lightly worse performance by students who do not have a computer at home is also suggested.

4.2.3 *Saeb*, 9*th grade*, *Math/Port*, *Bahia*. The dataset consists of 6179 students and each test has 13 items. 311 students did not respond one of the three questions in the questionnaire, 104 did not respond two of them and 63 all of them, leading to a total of 708 missing values. In particular 144, 349 and 215 did not answer the first, second and third questions, respectively. Results are shown in Tables 9 and 10 and Figure 4.

Results suggest a possible MNAR structure associated to the first question, in which students with at least one computer at home would be more likely not to provide that information. The fitted regression model indicates a strong positive association between the abilities in Portuguese and Mathematics. It also indicates

Table 10 *Estimates of the regression's coefficients—posterior mean and* 99% *credibility interval*

Coefficient	Estimate
β_{11}	−0.339 (−0.532, −0.148)
β_{12}	−0.256 (−0.452, −0.067)
β_2	−0.013 (−0.168, 0.144)
β_3	0.108 (−0.386, 0.511)
β_4	0.994 (0.909, 1.101)

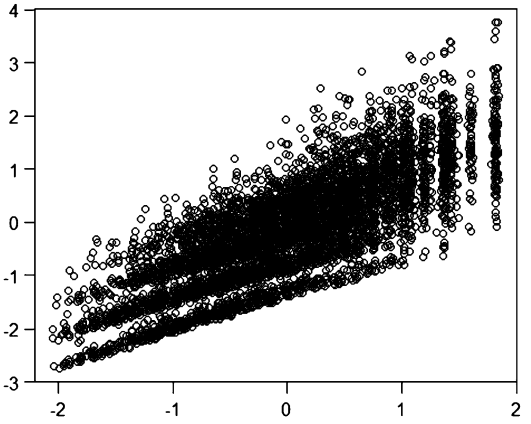

Figure 4 *Estimated abilities in Portuguese* (*x-axis*) *and Mathematics*. *The correlation is* 0.804.

a lightly worse performance by students with 1 or 0 computer at home compared to those with 2 or more.

4.2.4 *Saeb*, 3*rd grade*, *Math/Port*, *Brazil*. The dataset consists of 8000 students and each test has 13 items. 530 students did not respond one of the three questions in the questionnaire, 80 did not respond two of them and 9 all of them, leading to a total of 717 missing values. In particular 349, 183 and 205 did not answer the first, second and third questions, respectively. Results are shown in Tables 11 and 12 and Figure 5.

Results suggest a possible MNAR structure associated to the third question, in which students who like Maths would be more likely not to respond this question. The fitted regression model indicates a positive association between the abilities in Portuguese and Mathematics. It also indicates a lightly worse performance by Indigenous or Yellow and moderate worse performance of Black or Brown students when compared to White students. Finally, a better performance by male students and by students who declared that they like Maths is strongly suggested.

Table 11 *Observed and estimated proportions for each response with the respective number of responses. The observed refers the proportions of the responses that were observed, and the estimated refers to the proportions of the responses that were estimated for the missing cases*

Covariate	Observed	Estimated
Ethnicity	$(0.617, 0.055, 0.327) - 7651$	$(0.617, 0.055, 0.327) - 349$
Gender	$(0.449, 0.551) - 7817$	$(0.453, 0.547) - 183$
LikeMaths	$(0.561, 0.439) - 7779$	$(0.738, 0.262) - 211$

Table 12 *Estimates of the regression's coefficients—posterior mean and 99% credibility interval*

Coefficient	Estimate
β_{11}	-0.370 (-0.462, -0.287)
β_{12}	-0.178 (-0.392, 0.044)
β_2	0.548 (0.449, 0.646)
β_3	0.873 (0.769, 0.977)
β_4	1.300 (1.213, 1.384)

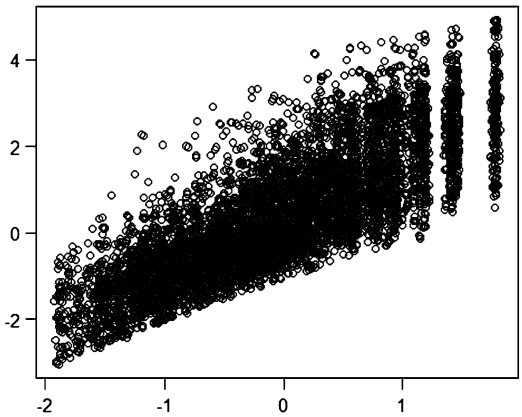

Figure 5 *Estimated abilities in Portuguese* (x*-axis*) *and Mathematics. The correlation is* 0.810.

4.2.5 *Enem*, *Math*, *Belo Horizonte*. The dataset consists of 4470 students and each test has 45 items. 129 students did not respond the first question in the questionnaire and 4 did not respond third one. Results are shown in Tables 13 and 14 and Figure 6.

Results do not suggest the presence of a MNAR structure in any of the two questions with missing values. The fitted regression model indicates a lightly worse performance by Black and Brown students and a considerably worse performance by Indigenous or Yellow students when compared to White ones. It also indicates

Table 13 *Observed and estimated proportions for each response with the respective number of responses. The estimated refers to the proportions of the responses that were estimated for the missing cases*

Covariate	Observed	Estimated
Ethnicity	$(0.626, 0.027, 0.346) - 4341$	$(0.606, 0.025, 0.368) - 129$
Gender	$(0.617, 0.383) - 4470$	-0
Net	$(0.160, 0.840) - 4466$	$(0.128, 0.872) - 4$

Table 14 *Estimates of the regression's coefficients—posterior mean and 99% credibility interval*

Coefficient	Estimate
β_{11}	-0.384 $(-0.491, -0.286)$
β_{12}	-0.770 $(-1.097, -0.448)$
β_2	-0.279 $(-0.376, -0.182)$
β_3	-0.701 $(-0.838, -0.560)$

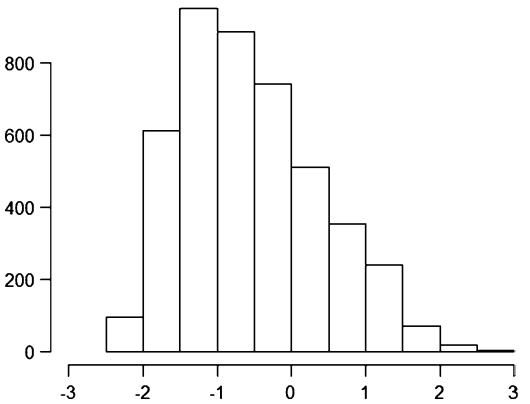

Figure 6 *Estimated abilities in Math for the Enem dataset from Belo Horizonte.*

a lightly worse performance by female students and a considerably worse performance by students who do not have internet access at home.

5 Extensions of the regression model for the abilities

We present three possible extensions of the regression model for the abilities presented in Section 2.

5.1 Multilevel model

Fox (2005) proposes a multilevel regression model to explain the abilities using covariates from different levels, for example, school, teachers, students. The two level model, say for school and student, is given by:

$$\begin{aligned} \theta_{jk} &= \mathbf{X}_{j\cdot k}\beta_k + e_{jk}, \\ \beta_q &= \mathbf{W}_q\gamma_q + u_q, \end{aligned} \tag{5.1}$$

where jk refers to student j from school k, $\beta_k = (\beta_{0k}, \ldots, \beta_{Qk})$, $\mathbf{X}_{j\cdot k}$ is the jth row of the design matrix X_k from school k, $\beta_q = (\beta_{q1}, \ldots, \beta_{qK})'$, $\mathbf{W}_q$ are the covariates in the second level to explain the variability of β_q among schools, $\gamma_q = (\gamma_{0q}, \ldots, \gamma_{Sq})$ and, finally, $u_q = (u_{q1}, \ldots, u_{qK})'$ are independent random error with $u_{qk} \sim N(0, \sigma_u^2)$.

Identifiability is achieved by making $\beta_{0k} = 0$, for all k and $e_{jk} \sim N(0, 1)$, for all jk.

5.2 Latent covariates

The use of latent covariates may be reasonable in several examples. These may include socioeconomic and cultural status, or any other ability. Each of these covariates is a latent factor explaining responses in (part of) a questionnaire or test. Let λ represent a latent covariate and $\dot{X}$ be the responses, from a questionnaire or another test, modelled by this covariate.

A latent covariate may be inserted in the model under two different approaches. In the first one, it is previously estimated, based on the responses that it models, and the point estimates are used as a fixed covariate. This approach is appropriate when:

- $(\lambda|\theta, \dot{X}) \approx (\lambda|\dot{X})$;
- it is reasonable to ignore the variability/uncertainty about λ.

One possible example is the socioeconomic status.

The second approach to include a latent covariate in the model consists of modelling this jointly with the remaining components of the model. This is reasonable, for example, when the latent covariate is a factor summarising (via factor analysis or IRT) the variability among (some of) the items in the questionnaire. This could allow the use of the information in the questionnaire without having multicollinearity problems. Another example is when the latent covariate is in fact another ability, measured in a different test.

In order to present the general formulation of the model with latent covariates, we consider that the latent covariates treated under the first approach are included in the design matrix X. The model is as following:

$$\pi(Y|\theta, \xi)\pi(\xi)\pi(\dot{X}|\lambda, \zeta)\pi(\zeta)\pi(\theta|\lambda, X, \beta, \alpha)\pi(\lambda)\pi(X)\pi(\beta)\pi(\alpha), \tag{5.2}$$

where:

- $\pi(Y|\theta, \xi)$ is the 3PNO model in (2.1);
- $\pi(\xi)$ is the prior on the item parameters;
- $\pi(\dot{X}|\lambda, \zeta)$ is the factor model or IRT model for the responses modelled by the latent covariates. ζ are loadings or item parameters;
- $\pi(\zeta)$ is the prior on ζ;
- $\pi(\theta|\lambda, X, \beta, \alpha)$ defines the regression model $\theta_j = X_{j\cdot}\beta + \lambda_{j\cdot}\alpha + e_j$;
- $\pi(\lambda)$ is a $N(0, 1)$ prior;
- $\pi(X)$ is the prior on X, essential to model and estimate the missing data;
- $\pi(\beta)\pi(\alpha)$ are the prior on the regression coefficients, with $\alpha \in (0, 1)$.

Model identifiability is achieved by setting, for example, $e_j \sim N(0, 1 - \alpha^2)$.

5.2.1 *Graded response IRT model.* Given that the questions in the questionnaire are typically polytomous, responses $\dot{X}$ are often suitably modelled by a graded response IRT model. We consider the model proposed by Samejima (1969), which assumes a natural graduation of the possible responses, in the sense of being monotonically related to the latent covariate. Defining p_{hkj} as the probability that individual j gives a response k to item h, we have that:

$$p_{hkj} = p^{+}_{h(k-1)j} - p^{+}_{hkj}, \tag{5.3}$$

$$p^{+}_{hkj} = \Phi(\dot{a}_h\lambda_j - \dot{b}_{hk}), \tag{5.4}$$

where $-\infty = \dot{b}_{h0} < \dot{b}_{h1} < \dot{b}_{h2} < \cdots < \dot{b}_{hK} = \infty$.

5.3 Mixtures

Mixtures of distributions play an increasingly important role in statistical modelling, specially for regression models. At least two nice features may be introduced by mixtures in a regression context.

Traditional linear regression models assume the errors to be normally distributed. Although reasonable in many cases, more flexible structures may sometimes need to be considered. Important features like multimodality, skewness and heavy tails can be efficiently accommodated by mixtures of Normal distributions in a parsimonious way (see Richardson and Green, 1997). Gonçalves, Dias and Soares (2018) propose a flexible approach to model the abilities through a mixture of normals in the 3PNO model that is able to properly accommodate the features described above and still guarantee model identifiability. The same structure may be used to model the errors in the regression model for the abilities proposed in this paper. In particular, this can be a way to model possible latent sources of heterogeneity which are not captured by the covariates. We assume

$$e_j \sim \sum_{l=1}^{L} r_l N\big(\mu_l, \sigma_l^2\big), \quad r_l > 0, \sum_{l=1}^{L} r_l = 1. \tag{5.5}$$

Gonçalves, Dias and Soares (2018) provide the required conditions to achieve model identifiability as well as the details to adapt our MCMC algorithm.

Another interesting use of mixtures in a regression context is to perform variable selection by considering a point-mass mixture prior for the regression coefficients. This is a well known modelling technique that works well for a reasonable number of potential covariates. More specifically, we assume

$$\beta_q \sim r N\left(\mu_l, \sigma_l^2\right) + (1-r)\delta_0, \quad r \in (0,1), \tag{5.6}$$

where δ_0 is a point-mass at zero. This approach allows the variable selection procedure to be performed under the Bayesian Paradigm, based on the posterior probability of each model.

6 Final remarks

This paper addressed the problem of using covariates from the contextual questionnaire to explain the ability of students with data modelled via the 3PNO model. The possibility of having missing values in the covariates was considered by modelling the missing data jointly with the other components of the model. An efficient MCMC algorithm was proposed to perform inference under the Bayesian approach. The efficiency of the algorithm was investigated in a simulated example which also illustrated the possibility of identifying MNAR structures in the missingness process. Finally, the analysis of some real datasets concerning two large scale educational assessment exams in Brazil illustrated the applicability of the proposed methodology and led to some interesting conclusions. Some extensions of the proposed model were also discussed by considering a hierarchical regression model for the abilities, the use of latent covariates and mixtures distributions.

Finally, we highlight the fact that the estimation of the abilities is not affected by the regression prior, in the sense that information about a student's ability comes from its performance in the test and is not influenced by the group it belongs to w.r.t. its covariates' values. In order to see that, note that, if we integrate out the regression coefficients β by assuming a zero mean normal prior for these, the resulting model is the traditional 3PNO model with a zero mean normal prior for all the abilities and different variances for students from different groups. Given that the joint posterior of all the abilities is the same whether or not we integrate out the regression coefficients, the regression structure should not affect the estimation of the abilities.

Appendix

We present all the full conditional distributions of the Gibbs sampling algorithm proposed in Section 3.

- (Z, V)

All pairs (Z_{ij}, V_{ij}) are conditionally independent with

$$\pi(Z_{ij}, V_{ij}|\cdot)\begin{cases}\propto \phi(v_{ij}-m)I_{(Z_{ij}=0)}I_{(V_{ij}<0)} \\ \quad \text{if } Y_{ij}=0, \\ = wI_{(Z_{ij}=1)}I_{(V_{ij}=0)} \\ \quad +(1-w)\dfrac{\phi(v_{ij}-m)}{\Phi(m)}I_{(Z_{ij}=0)}I_{(V_{ij}>0)} \\ \quad \text{if } Y_{ij}=1,\end{cases} \tag{A.1}$$

where $m = a_i\theta_j - b_i$ and $w = \frac{c_i}{c_i+(1-c_i)\Phi(m)}$.

- c

All the c_i's are conditionally independent with

$$(c_i|\cdot) \sim \text{Beta}\left(\sum_{j=1}^{J} Z_{ij} + \alpha_c, J - \sum_{j=1}^{J} Z_{ij} + \beta_c\right) \tag{A.2}$$

- (a_i, b_i)

All pairs (a_i, b_i) are conditionally independent with

$$(a_i, b_i|\cdot) \sim N_2(\mu, \Sigma), \tag{A.3}$$

where $\mu = \left[\begin{smallmatrix}\mu_a{}^* \\ \mu_b{}^*\end{smallmatrix}\right]$, $\Sigma = \left[\begin{smallmatrix}\sigma_a^{2*} & \gamma \\ \gamma & \sigma_b^{2*}\end{smallmatrix}\right]$ and $\sigma_a^{2*} = \frac{\sigma_a^2}{(\sigma_a^2\sum_{j=1}^{L_i}\theta_j^2+1)(1-\gamma^2)}$, $\sigma_b^{2*} = \frac{\sigma_b^2}{(\sigma_b^2 J+1)(1-\gamma^2)}$, $\gamma = \frac{\sigma_a\sigma_b\sum_{j=1}^{L_i}\theta_j}{[(\sigma_a^2\sum_{j=1}^{L_i}\theta_j^2+1)(\sigma_b^2 J\theta_j^2+1)]^{\frac{1}{2}}}$, $\mu_a{}^* = \sigma_a^{2*}(\sum_{j=1}^{J} v_{ij}\theta_j + \mu_a\sigma_a^{-2}) - \sigma_a{}^*\sigma_b{}^*\gamma(\sum_{j=1}^{L_i} v_{ij} - \mu_b\sigma_b^{-2})$, $\mu_b{}^* = \sigma_a{}^*\sigma_b{}^*\gamma(\sum_{j=1}^{J} v_{ij}\theta_j + \mu_a\sigma_a^{-2}) - \sigma_b^{2*}(\sum_{j=1}^{L_i} v_{ij} - \mu_b\sigma_b^{-2})$, $L_i = \{j; z_{ij} \neq 0\}$.

- θ

All the θ_j's are conditionally independent with

$$(\theta_j|\cdot) \sim N\left(m_\theta, \sigma_\theta^2\right), \tag{A.4}$$

where $m_\theta = \frac{\sigma_e^2\sum_{i=1}^{L_j} a_i(v_{ij}+b_i)+Xj\cdot{}^*\beta^*}{\sigma_e^2\sum_{i=1}^{L_j} a_i^2+1}$, $\sigma_\theta^2 = \frac{\sigma_e^2}{\sigma_e^2\sum_{i=1}^{L_j} a_i^2+1}$, $L_j = \{i; z_{ij} \neq 0\}$.

- β^*

$$\left(\beta^*|\cdot\right) \sim N_{Q+P+1}\left(\mu_\beta, \Sigma_\beta^*\right), \tag{A.5}$$

where $\mu_\beta = \Sigma_\beta(\Sigma_\beta^{-1}\mu_\beta + X'\theta)$ and $\Sigma_\beta^{*-1} = \Sigma_\beta^{-1} + X'X$.

- X_{mis}

Assuming that the prior on X_{mis} is discrete, we have

$$P(X_{\text{mis}} = x_{\text{mis}}|\cdot) \propto \pi(\theta|X\beta)P(X_{\text{mis}} = x_{\text{mis}}), \tag{A.6}$$

where the missing values of matrix X in the first term of the rhs are set as x_{mis}. If, for example, the prior of X is independent for different individuals, that is also true for the full conditional distribution.

Acknowledgments

The authors would like to thank Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos—CEBRASPE, for financing this project. The authors also thank the anonymous referee for many useful comments that led to a much more improved version of the paper. The first author would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq, for financial support.

References

Alves, M. T. G., Xavier, F. P., Barbosa, L. E., de Figueiredo Caldeira, B., Silva, C. A. S. and Soares, J. F. (2015). Fatores contextuais das escolas de educação básica brasileiras: Dados, métodos e aplicações. *Reuniões da ABAVE* **8**, 57–76.

da Silva Fernandes, N., Soares, T. M., Pena, A. C. and Cunha, I. C. (2010). O conhecimento do professor em avaliação educacional e a proficiência do aluno. *Estudos em Avaliação Educacional* **21**, 569–590.

Fox, J. (2005). Multilevel irt using dichotomous and polytomous response data. *British Journal of Mathematical & Statistical Psychology* **58**, 145–172.

Gonçalves, F. B., Dias, B. C. C. and Soares, T. M. (2018). Bayesian item response model: A generalised approach for the abilities' distribution using mixtures. *Journal of Statistical Computation and Simulation* **88**, 967–981.

Laros, J. A. and Marciano, J. L. (2008). Índices educacionais associados à proficiência em língua portuguesa: Um estudo multinível. *Avaliação Psicológica* **7**, 371–389.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement.*

Soares, J. F. and Alves, M. T. G. (2003). Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa* **29**, 147–165.

Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika* **56**, 589–600.

Universidade Federal de Minas Gerais
Av. Antônio Carlos
6627-DEST/ICEx/UFMG-Belo Horizonte
Minas Gerais, 31270-901
Brazil
E-mail: fbgoncalves@est.ufmg.br