Improving Scaling in Large-Scale Assessment: A Variable Selection Approach to Latent Regression Modelling

Yunxiao Chen, Motonori Oka
London School of Economics and Political Science
Matthias von Davier
Boston College

Abstract

This report concerns the construction of scaling models for large-scale assessments in education. A scaling model, which makes use of information from both responses to cognitive assessment and background survey items, produces plausible values for individual students. There are two major challenges when building a scaling model – (1) a large number of background variables and (2) many missing values in the background survey data. Traditionally, these challenges are tackled by a latent regression model, in which the missing values are handled by a missing indicator approach, and the high dimensionality of the background variables and their missing indicators is reduced by Principal Component Analysis (PCA). However, this approach has three drawbacks: (1) the missing indicator approach does not perform well under certain data missingness patterns, (2) PCA may introduce spurious dependence between the achievement traits and background variables, and (3) the resulting model lacks interpretability due to the involvement of hard-to-interpret principal component scores. To overcome these drawbacks, we propose a variable selection approach to latent regression modelling. The proposed approach handles missing data by iterative imputation and controls variable selection error by a data-splitting procedure. Simulation studies are conducted to evaluate the performance of the proposed method and also compare the proposed method with the traditional one. Finally, the proposed method is applied to the Progress in International Reading Literacy Study (PIRLS) 2016 data, from which sensible results are obtained while limitations are also identified.

Keywords: Latent regression, large-scale assessment, variable selection, missing data, imputation, PIRLS

1 Introduction

Scaling plays a key role in large-scale assessments in education. Making use of information from students' responses to cognitive assessment and background survey items, a scaling procedure produces plausible values for student achievement (von Davier, 2013). Most scaling procedures in large-scale assessments rely on a latent regression model (Mislevy, 1984, 1991), a latent variable model combining an Item Response Theory (IRT) measurement model and a linear regression structural model. Thanks to the IRT measurement model, the latent regression model naturally handles the matrix sampling design in large-scale assessments in which different students receive different but overlapping sets of cognitive items (Mislevy et al., 1992). This design is favoured as it enables a cognitive assessment to cover an extensive content domain while avoiding giving

each student too many items. In addition, the linear regression structural model borrows information from non-cognitive background variables to compensate for the potential shortage of cognitive information.

Due to the complexity of large-scale assessment data, there are two challenges when building such a latent regression model. First, surveys typically collect a large number of background variables: without dimension reduction techniques, a latent regression analysis including all these variables likely suffers from the curse of dimensionality – the estimation and prediction will become inaccurate when the model involves too many parameters. Second, there are often many missing values in the background variables (some variables may have up to 30% to 40% missing). A sensible treatment of missing data is non-trivial. A PCA-based approach (Martin et al., 2016, 2017; OECD, 2017, 2019) is the state-of-the-art approach in large-scale assessments for building scaling models. In this approach, missing values are treated using a missing indicator approach (Cohen and Cohen, 1975). This approach treats a missing value as an "extra category" and creates an additional indicator (i.e. dummy) variable to code it. With the dummy-coded predictor matrix, PCA is applied to reduce the dimensionality of data, from which the principal component scores of the top principal components are retained. Finally, a latent regression model is built based on the cognitive assessment data and the retained principal component scores.

Although the PCA-based approach provides a mechanism to handle both the high dimensionality of the predictors and the missing values, it has several drawbacks. First, the missing indicator approach lacks theoretical guarantees even when data are missing at random and, thus, has been discouraged in the statistical literature (e.g., Jones, 1996; Schafer and Graham, 2002). As shown in a simulation study in Section 5, the missing indicator approach, when combined with PCA, can perform very poorly if a large block of the background data matrix is missing – a situation likely to happen in large-scale assessments (e.g., when many parents fail to return the home survey questionnaire). Second, the PCA-based approach does not directly characterise the relationship between students' achievement traits and the background variables. To study the relationship, applied researchers need to run further regression analyses using the plausible values of achievement traits and the background variables. However, since the plausible values are produced using the principal component scores, spurious dependence may be introduced in this process, leading to less valid inference results in the secondary analyses. Fundamentally, this is due to the inconsistency between the data imputation model (i.e., the PCA-based latent regression model) and the analysis model researchers use (e.g., a linear regression). Finally, the final latent regression model lacks interpretability due to the involvement of hard-to-interpret principal component scores. That is, it is hard to tell how each background variable contributes to the prediction of achievement traits. Consequently, it becomes difficult to communicate to the public about the scaling model.

This report proposes a new method for constructing a scaling model. The new method addresses the issues in the state-of-the-art PCA-based approach. Our approach relies on several technical tools. First, to handle the missing values, we introduce a joint model for the predictors, which is based on an exponential family graphical model (Chen et al., 2015; Tsao, 1967; Yang et al., 2015) – also known as a Second-Order Exponential (SOE) model in the case of multivariate binary data. Under this exponential family graphical model, we introduce an Iterative Imputation (II) algorithm that simultaneously (1) imputes the missing values and the latent variables and (2) samples the unknown parameters of the latent regression model and the exponential family graphical model under a Bayesian framework. Second, we propose a variable selection procedure to reduce the dimensionality of the background variables and enhance the interpretability of the scaling model. To solve the multiple comparison problems with variable selection, we propose to control the False Discovery Rate (FDR). We apply a recently proposed Data Splitting (DS) method (Dai et al., 2022)

to control FDR. Combining the exponential family graphical model for predictors, the II algorithm and the DS approach, we propose a variable selection method for latent regression models.

We note that the proposed estimation method based on iterative imputation is closely related to a fully conditional specification (FCS) approach proposed in Grund et al. (2021), which allows for joint treatment of plausible values and missing data. However, Grund et al. (2021) did not tackle the high-dimensionality of background variables. We also note that Yamaguchi and Zhang (2023) proposed a Bayesian variable selection method for latent regression and applied it to the 2018 Programme for International Student Assessment (PISA) data. However, their method does not handle missing data, and thus, they performed listwise deletion when analysing PISA data.

The rest of the report is organised as follows. In Section 2, we describe the problem setting and the proposed model. In Section 3, we introduce an iterative imputation algorithm and an associated estimation method. In Section 4, we introduce a data-splitting for controlling the FDR of variable selection. Simulation studies are conducted in Section 5, and an application to data from PIRLS 2016 is given in Section 6. We conclude with discussions in Section 7.

2 Problem Setup

Consider data collected from N students, where the data are independent across students. For each student i, the data can be divided into two parts – (1) responses to cognitive items and (2) non-cognitive predictors. We use a random vector $\mathbf{Y}_i$ to denote student i's cognitive responses. Due to the matrix sampling design for cognitive items in ILSAs, the items different students receive can be different. More precisely, we use $\mathcal{B}_i$ to denote the set of cognitive items that student i is assigned. Then $\mathbf{Y}_i = \{Y_{ij} : j \in \mathcal{B}_i\}$. Depending on how each item is scored, Y_{ij} may be a binary variable taking value in $\{0, 1\}$ or an ordinal variable taking value in $\{0, 1, ..., K_j\}$, $K_j \geqslant 2$. In addition, consider p predictors collected via non-cognitive survey questionnaires. Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ denote the complete predictor vector for student i. Often, there are missing values in $\mathbf{Z}_i$. Let $\mathcal{A}_i$ denote the set of observed predictors for student i, and let $\mathbf{Z}_i^{\text{obs}} = \{Z_{ij} : j \in \mathcal{A}_i\}$ and $\mathbf{Z}_i^{\text{mis}} = \{Z_{ij} : j \notin \mathcal{A}_i\}$. The predictors are of mixed types. In the current study, binary, categorical (ordinal/nominal), and continuous predictors are considered. Note that an ordinal variable will be treated as a nominal one here for simplicity. Further modelling can be done to incorporate the information in the category order for ordinal variables, though we believe that the improvement will be very marginal in terms of the prediction power of the resulting model.

2.1 Measurement Model

We introduce a latent variable θ_i as the latent construct, which is measured by the cognitive items. The measurement model is an IRT model which specifies the conditional distribution of $\mathbf{Y}_i$ given θ_i. More specifically, this model assumes local independence, an assumption that is commonly adopted in IRT models (Lord and Novick, 1968; Embretson and Reise, 2000). That is, Y_{ij}, $j \in \mathcal{B}_i$, are conditionally independent given θ_i. For a dichotomously scored non-multiple-choice item j, the conditional distribution of Y_{ij} given θ_i is assumed to follow a Bernoulli distribution with probability $\mathbb{P}(Y_{ij} = 1|\theta_i)$ that follows a two-parameter logistic model (2PL, Birnbaum, 1968). That is,

$$\mathbb{P}(Y_{ij} = 1|\theta_i) = \frac{\exp(a_j\theta_i + b_j)}{1 + \exp(a_j\theta_i + b_j)},$$

where a_j and b_j are two item-specific parameters. For a dichotomously scored multiple-choice item j, the conditional distribution of Y_{ij} given θ_i is assumed to follow a three-parameter logistic model (3PL, Birnbaum, 1968). That is,

$$\mathbb{P}(Y_{ij} = 1|\theta_i) = c_j + (1 - c_j)\frac{\exp(a_j\theta_i + b_j)}{1 + \exp(a_j\theta_i + b_j)},$$

where a_j, b_j and c_j are item-specific parameters. For an ordinal item j, the generalised partial credit model (Muraki, 1992) is assumed, for which

$$\mathbb{P}(Y_{ij} = k|\theta_i) = \frac{\exp(a_j\theta_i + b_{jk})}{1 + \sum_{l=1}^{K_j}\exp(a_j\theta_i + b_{jl})},\ k \geqslant 1,$$

and

$$\mathbb{P}(Y_{ij} = 0|\theta_i) = \frac{1}{1 + \sum_{l=1}^{K_j}\exp(a_j\theta_i + b_{jl})},$$

where a_j and b_{jl}, $l = 1, ..., K_j$ are item-specific parameters.

In the analysis below, we assume that all the item parameters are known – they are fixed at pre-calibrated values. This is not a constraint or limitation of the method. This is an assumption that follows operational practice, where the IRT measurement model is estimated first, and then the latent regression is estimated. This is the case for PIRLS, as well as the Trends in International Mathematics and Science Study (TIMSS), PISA, the Programme for the International Assessment of Adult Competencies (PIAAC), the National Assessment of Educational Progress (NAEP) and other studies using the general latent regression and plausible values approach.

2.2 Structural Model

The structural model regresses the latent construct θ_i onto the complete-data predictors Z_{i1}, ..., Z_{ip}. A linear regression model is assumed for θ_i given Z_{i1}, ..., Z_{ip}. More specifically, for each variable j, we introduce a transformation $g_j(Z_j)$. When Z_j is an ordinal or categorical variable with categories $\{0, ..., K_j\}$, the transformation function g_j creates K_j dummy variables, i.e., $g_j(Z_j) = (\mathbb{I}(\{Z_j = 1\}), ..., \mathbb{I}(\{Z_j = K_j\}))^\top$. For continuous and binary variables, g_j is an identity link, i.e., $g_j(Z_j) = Z_j$. We assume $\theta_i|\mathbf{Z}_i \sim N(\beta_0 + \boldsymbol{\beta}_1^\top g_1(Z_{i1}) + \cdots + \boldsymbol{\beta}_p^\top g_p(Z_{ip}), \sigma^2)$, where β_0 is the intercept, $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_p$ are the regression coefficients, and σ^2 is the residual variance. Note that $\boldsymbol{\beta}_j$ is a scalar when predictor j is continuous or binary and is a vector when the predictor is ordinal or categorical. Here, β_0, $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_p$, and σ are unknown and will be estimated from the model. The main goal of our analysis is to find predictors of θ for which $\|\boldsymbol{\beta}_j\| \neq 0$.

2.3 Predictor Model

To handle missing values in Z_{ij}s, we impose a joint model for the predictors. Although different models may be imposed here, we assume an exponential family graphical model, under which missing data imputation and parameter estimation can be carried out in a computationally efficient way (see Section 3 for more details). More precisely, we let $(\theta_i, \mathbf{Z}_i)$ be (independent and identically distributed) i.i.d., following an exponential family graphical model. This model includes the SOE model for multivariate binary data as a special. More specifically, the joint distribution for $\mathbf{X} = (\theta_i, g_1(Z_{i1}), ..., g_p(Z_{ip}))^\top$ satisfies

$$f(\mathbf{x}|\boldsymbol{\gamma}, \boldsymbol{\Lambda}) \propto \exp\left(\boldsymbol{\gamma}^\top\mathbf{x} + \mathbf{x}^\top\boldsymbol{\Lambda}\mathbf{x}\right), \tag{1}$$

where the notation $\propto$ means that the two sides of (1) differ by a constant that does not depend on $\mathbf{x}$, and $\mathbf{\Lambda}$ and $\boldsymbol{\gamma}$ denote unknown parameters of the model. More specifically, $\mathbf{\Lambda}$ is a symmetric matrix, and the sth entry of $\boldsymbol{\gamma}$ is zero if x_s is a dummy variable.

Under this model, the conditional distribution of θ_i given $\mathbf{Z}_i$ is the linear regression model in the above structural model. The conditional distribution of Z_{ij} given $(\theta_i, Z_{i,-j})$ takes the following forms:

- A linear regression model (with normal residual), if variable j is continuous;
- A logistic regression model, if variable j is binary;
- A multinomial logistic regression model if variable j is categorical.

These conditional distributions involve unknown parameters that depend on $\mathbf{\Lambda}$ and $\boldsymbol{\gamma}$ in the joint model. These conditional distributions will be used later for missing data imputation and parameter estimation. We remark that except for the parameters of the structural model, the rest of the parameters in the exponential family graphical model can be viewed as nuisance parameters, as they are not of interest to us. The predictor model and these nuisance parameters are introduced to handle the missing values in the predictors.

3 Estimation Procedure

In what follows, we introduce an Iterative Imputation (II) algorithm for estimating the parameters in the structural model, i.e., β_0, $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_p$, and σ^2. Recall that our setting assumes that the item parameters in the measurement model are known, and thus, these parameters are not estimated. This approach also estimates the parameters in the predictor model, though we do not explicitly list them as an output from the II algorithm.

Algorithm 1 (II Algorithm)**.**

> **Input:** *Total number of iterations T, burn-in size $B < T$, an initial imputation of the latent variables and the missing predictors, denoted by $\theta_i^{(0)}$ and $\mathbf{Z}_i^{(0),mis}$, respectively. Observed responses $\mathbf{Y}_i$ and observed predictors $\mathbf{Z}_i^{\text{obs}}$. We use $\mathbf{Z}_i^{(0)}$ to denote the complete predictor vector for student i, with the missing entries replaced by $\mathbf{Z}_i^{(0),mis}$. The priors for parameters in each conditional model of Z_j given $\mathbf{Z}_{-j}$ and θ, and those for the parameters in the linear regression model of Z_j given $\mathbf{Z}_{-j}$ and θ. Initial values for the parameters in the logistic and multinomial logistic conditional models (to be used for Gibbs sampling).*
>
> **For iterations** $t = 1, ..., T$**,** *we alternate between the following two steps in each iteration t.*
>
> > **1. Sampling (S) Step:** *For each of the variables $j = 1, ..., p$, sample the parameters in the conditional model of Z_j given $\mathbf{Z}_{-j}$ and θ. Exact or approximate samples are obtained from the posterior distribution of the regression coefficient parameters given imputed data $\mathbf{Z}_i^{(t-1)}$ and latent variables $\theta_i^{(t-1)}$ and the corresponding input prior distribution for these parameters. That is, for each variable j, we sample the parameters of the corresponding Bayesian linear, logistic, or multinomial logistic model (depending on the type of variable j). Finally, we sample the parameters of the linear regression model of θ given $\mathbf{Z}$ based on the imputed data and the input priors for these parameters. The sampled parameters from this model are denoted by $(\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, (\sigma^{(t)})^2)$.*

2. Imputation (I) Step: *Impute the missing values using the estimated conditional models from the S-step of the same iteration. We first update missing values in the predictors using a random-scan Gibbs-type procedure. That is, we first randomly generate the updating order* $j_1, ..., j_p$. *Then, we update the missing values in predictors* j_1, ..., j_p, *one predictor at a time. When imputing the missing values in variable* j_l, *we plug in* $\theta_i^{(t-1)}$ *and the most recently imputed data for* $\mathbf{Z}_{i,-j_l}$ *(i.e., the missing values in variables* j_1, ..., j_{l-1} *are from this iteration,* t, *and the missing values in variables* $j_{l+1}, ..., j_p$ *are from the previous iteration,* $t-1$*). After all the missing values in predictors have been imputed, we denote the recently updated complete data by* $\mathbf{Z}_i^{(t)}$, $i = 1, ..., N$. *Finally, we impute the latent variables based on the conditional distribution of* θ_i *given* $\mathbf{Y}_i$ *and* $\mathbf{Z}_i^{(t)}$, *where the conditional distribution is calculated under the known measurement model and the estimated structural model from the S-step. We denote the imputed latent variables by* $\theta_i^{(t)}$, $i = 1, ..., N$.

Output: *The estimate of the structural parameters*

$$\widehat{\beta}_0 = \frac{\sum_{t=B+1}^{T} \beta_0^{(t)}}{T-B}, \ \widehat{\boldsymbol{\beta}} = \frac{\sum_{t=B+1}^{T} \boldsymbol{\beta}^{(t)}}{T-B}, \ \text{and } \widehat{\sigma} = \frac{\sum_{t=B+1}^{T} \sigma^{(t)}}{T-B}.$$

We provide a few remarks on the II algorithm.

Remark 1. *In our implementation, weakly informative priors are used for the parameters of the Bayesian models in the S-step of the algorithm. Specifically, for the linear regression models, we assume that the regression coefficients are i.i.d., each following a normal distribution with mean zero and variance 100, and the residual variance follows an inverse gamma distribution, with scale and shape parameters both being 0.1. For the logistic and multinomial logistic regression models, we assume that the regression coefficients are i.i.d., each following a normal distribution with mean zero and variance 100.*

Under these priors, the sampling in the S-step is straightforward. More specifically, the priors for the linear regression models are conjugate, and thus, their parameters can be drawn from the normal and inverse gamma distributions. For the logistic and multinomial logistic regression models, we sample their parameters using a Pólya-Gamma Gibbs sampler, following Polson et al. (2013). This sampler is a state-of-the-art sampling technique for Bayesian logistic models that converges very fast. We note that the Pólya-Gamma Gibbs sampler only involves sampling from a Pólya-Gamma distribution and sampling from a normal distribution, both of which can be carried out efficiently. Sampling from the Pólya-Gamma distribution is implemented using the "pgdraw" package in R.

Remark 2. *The latent variables* θ_i *are sampled using an Adaptive Rejection Metropolis Sampling (ARMS) algorithm (Gilks and Wild, 1992) in the I-step. This sampling step is implemented using the R package "armspp".*

Remark 3. *The proposed method falls under the theoretical framework of Liu et al. (2014) for iterative imputation.*

Remark 4. *The II algorithm is very similar to the FCS method proposed in Grund et al. (2021) for the simultaneous imputation of plausible values of the latent traits and the missing predictors. However, Grund et al. (2021) used predictive mean matching to impute missing values in each iteration of their FCS method. The predictive mean matching method is a popular hot deck imputation method with good empirical perfor-*

mance. However, this method lacks a theoretical guarantee. In particular, it is not covered by the theoretical framework of Liu et al. (2014).

4 Variable Selection with FDR Control

4.1 FDR Control

To tackle the high dimensionality of background variables, we propose to perform variable selection and then use the selected variables in the scaling model. When performing variable selection, we face a trade-off between the type I error (selecting a null variable as non-null) and the type II error (selecting a non-null variable as null). Most statistical procedures consider controlling a certain type-I error rate, such as the family-wise type-I error rate and the False Discovery Rate (FDR). Since the FDR is an error metric that better scales with the high-dimensionality of data, we consider variable selection with a controlled FDR. Let S^* denote the set of non-null predictors, i.e., the predictors for which $\boldsymbol{\beta}_j^* \neq 0$. Let $\widehat{S}$ be the selected set of non-null variables given by a variable selection method. Then the FDR is defined as

$$\text{FDR} = E\left((\widehat{S} \backslash S^*)/|\widehat{S}|\right),$$

i.e., the expected proportion of null predictors among the selected ones. We want to control the FDR to be below an acceptable level α to ensure that most of the selected variables are relevant. In practice, we may use $\alpha = 0.05$ or 0.1.

4.2 Data Splitting Method

We apply a recently proposed Data Splitting (DS) method (Dai et al., 2022) to control the FDR of variable selection. We describe this method in Algorithm 2 below.

Algorithm 2 (DS Method)**.**

> **Input:** *Total number of iterations T, burn-in size $B < T$, an initial imputation of the latent variables and the missing predictors, denoted by $\theta_i^{(0)}$ and $\mathbf{Z}_i^{(0),mis}$, respectively. Observed responses $\mathbf{Y}_i$ and observed predictors $\mathbf{Z}_i^{\text{obs}}$. We use $\mathbf{Z}_i^{(0)}$ to denote the complete predictor vector for student i, in which we replace $\mathbf{Z}_i^{(0),mis}$. A Target FDR level $\alpha \in (0,1)$.*
>
> **Step 1:** *Randomly split the data into two equal-size datasets.*
>
> **Step 2:** *For each dataset, use the II algorithm (Algorithm 1) and obtain an estimate of the structural parameters. We denote the estimated regression coefficients from the two datasets by $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$, respectively.*
>
> **Step 3:** *For each variable j, we calculate a mirror statistic, M_j. When predictor j is a continuous or binary variable, the corresponding regression coefficient is a scalar. Let the estimates of the parameter from Step 2 be $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$. Then its mirror statistic can be constructed as*
>
> $$M_j = sign(\widehat{\beta}_j^{(1)}\widehat{\beta}_j^{(2)})(|\widehat{\beta}_j^{(1)}| + |\widehat{\beta}_j^{(2)}|).$$

When predictor j is a categorical variable with K_j categories, the corresponding regression coefficient is a vector. Let the estimates of $\boldsymbol{\beta}_j$ from Step 2 be $\widehat{\boldsymbol{\beta}}_j^{(1)} = (\widehat{\beta}_{j1}^{(1)}, ..., \widehat{\beta}_{j,K_j}^{(1)})^\top$ and $\widehat{\boldsymbol{\beta}}_j^{(2)} = (\widehat{\beta}_{j1}^{(2)}, ..., \widehat{\beta}_{j,K_j}^{(2)})^\top$. Its mirror statistic is constructed as

$$M_j = \sum_{k=1}^{K_j} sign(\widehat{\beta}_{jk}^{(1)} \widehat{\beta}_{jk}^{(2)})(|\widehat{\beta}_{jk}^{(1)}| + |\widehat{\beta}_{jk}^{(2)}|).$$

For a non-null variable, its regression coefficients from two splits tend to be near their non-zero true value. Thus, they tend to be large in absolute value and of the same sign, yielding a large and positive mirror statistic. In contrast, the regression coefficients from two splits tend to be near zero, and thus, its mirror statistic may be negative or close to zero.

Step 4: *Calculate the cutoff τ_α as*

$$\tau_\alpha = \min\left\{t > 0 : \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \leqslant \alpha\right\}$$

Output: *The selected variables $\{j : M_j > \tau_\alpha\}$.*

Under reasonable regularity conditions, this procedure can control the FDR as the number of variables goes to infinity. See Dai et al. (2022) for the intuitions and theoretical results.

4.3 Aggregating Multiple Data Splittings

The algorithm above only involves only one data splitting, which involves some randomness. To minimise such randomness, we further adapt the aggregation procedure of Dai et al. (2022) to aggregate the results from multiple data splittings. With additional regularity conditions, this aggregation procedure asymptotically controls the FDR.

Algorithm 3 (DS Aggregation)**.**

Input: *Total number of iterations T, burn-in size $B < T$, an initial imputation of the latent variables and the missing predictors, denoted by $\theta_i^{(0)}$ and $\mathbf{Z}_i^{(0),mis}$, respectively. Observed responses $\mathbf{Y}_i$ and observed predictors $\mathbf{Z}_i^{\text{obs}}$. We use $\mathbf{Z}_i^{(0)}$ to denote the complete predictor vector for student i, with the missing entries replaced by $\mathbf{Z}_i^{(0),mis}$. A target FDR level $\alpha \in (0, 1)$. The number of independent data splittings m.*

Step 1: *Run the DS method (Algorithm 2) m times, each time with an independent data splitting. Let $\widehat{S}^{(k)}$ be the selected predictors in each run, $k = 1, ..., m$.*

Step 2: *Calculate the empirical inclusion rate $\widehat{I}_j$ as*

$$\widehat{I}_j = \frac{1}{m} \sum_{k=1}^{m} \frac{1_{\{j \in \widehat{S}^{(k)}\}}}{|\widehat{S}^{(k)}| \vee 1}.$$

Step 3: *Sort the features with respect to their empirical inclusion rates in increasing order. Denote the sorted empirical inclusion rates as $0 \leqslant \widehat{I}_{(1)} \leqslant \widehat{I}_{(2)} \leqslant \cdots \leqslant \widehat{I}_{(p)}$.*

Step 4: *Find the largest $l \in 1, ..., p$ such that $\widehat{I}_{(1)} + \cdots \widehat{I}_{(l)} \leqslant \alpha$.*

Output: *Selected predictors $\widehat{S} = \{j : \widehat{I}_j > \widehat{I}_{(l)}\}$.*

5 Simulation Study

5.1 Checking the Properties of the II Estimator

We check the properties of the II estimator, with a focus on the estimation of the regression coefficients in the structural model. We consider a setting with $N = 2000$, $J = 60$ and $p = 60$. Each student is assumed only to receive 20 items randomly selected from the 60 items. The predictors are missing completely at random, with about 40% of the data entries missing. The 60 variables include 20 continuous variables, 20 binary variables, and 20 categorical variables. For each variable type, there are only five non-null variables. As it is difficult to specify the true parameters for the exponential family graphical model (as setting the parameters arbitrarily leads to unreasonable distributions), we generate the background variables using a Gaussian copula model. Therefore, the predictor model is slightly misspecified.

For the DS procedure to work, we need the estimate of the zero coefficients in $\boldsymbol{\beta}$ to be (asymptotically) symmetric about zero. We check this property graphically. In Figure 1, we show the histograms of the estimates of six zero coefficients based on 100 independent simulations. We have also checked the estimates of the other zero coefficients, for which the results are similar. We see that the symmetric assumption roughly holds. We further check the accuracy of the point estimation for the non-zero coefficients. In Figure 2, we show the histograms of the estimates of six non-zero coefficients based on 100 independent simulations, where the true values of the coefficients are indicated by vertical red lines. We have also checked the estimates of the other non-zero coefficients. We see that the estimates are reasonably accurate but have some bias. We believe that the bias might be due to two reasons: (1) model misspecification and (2) overfitting (as the number of predictors is large). The biases of the zero and nonzero coefficients are given in Figure 3. We also notice that the variances of the parameter estimates are not very small. This variance includes two parts – (1) statistical error determined by the sample size and (2) Monte Carlo error that can be further reduced by increasing the number of iterations T in Algorithm 1 (in this analysis, we set $T = 1000$ and burn-in size $B = 200$). If the aim is to estimate a final model, then we may need to increase T. See results in Section 5.2 below for the model selection results.

5.2 Checking the Results of Variable Selection

We now check the performance of Algorithm 3 under the same setting as above. We set the number of data splittings to be $m = 20$ and the target FDR level $\alpha = 0.1$. The results below are based on 100 replications (1 replication takes about 30 hours on one CPU core without parallel computing). The estimated FDR based on 100 replications is 9.98%, which is very close to the nominal level of 10%. The histogram of False Discovery Proportion (FDP) based on the 100 replications is given in Figure 4, where we see that most of the FDPs are around the nominal level, and there are 17 cases (out of 100) for which the FDP is above 0.2. The True Positive Rate for variable selection is 98.3%, and the False Positive Rate is 4.1%. With this relatively high TPR and low FPR, the latent regression model using the selected variables should perform similarly to the true model in terms of scaling.

5.3 Examining PCA-based Approach

In this subsection, we give constructed examples in which the PCA approach does not perform well. We consider three examples. The first example concerns the selection of non-null variables, and the second and third examples concern the prediction of the latent trait.

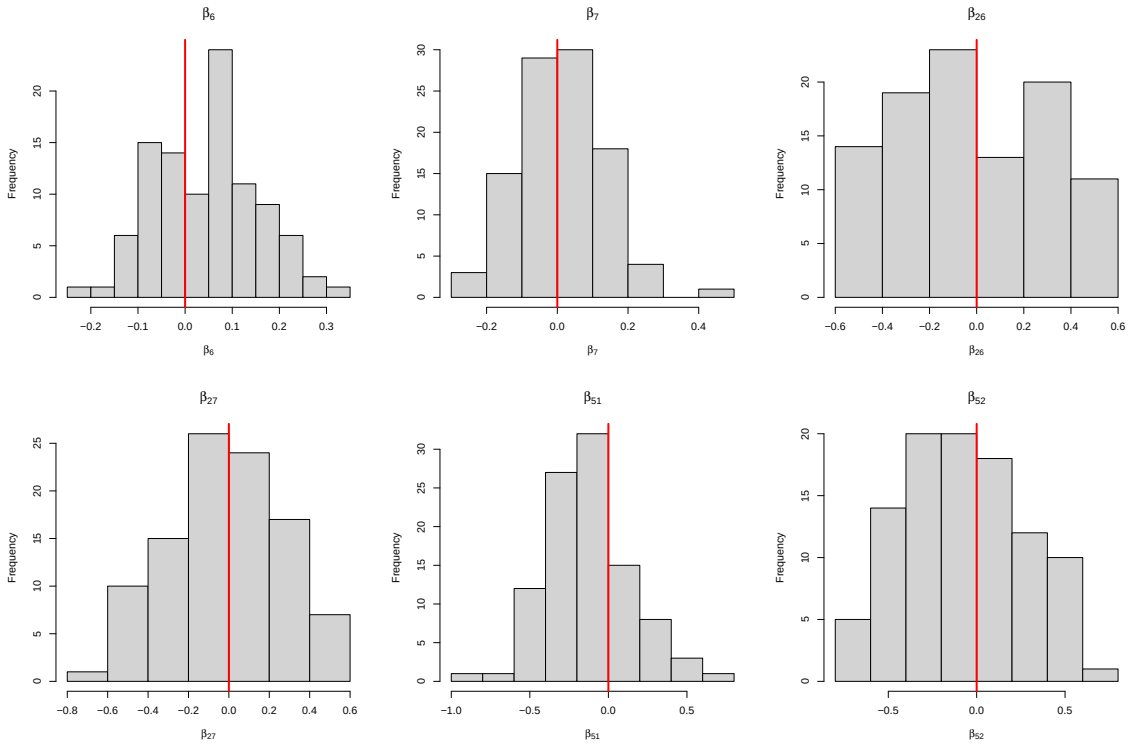

Figure 1: The histograms of the estimates for four zero coefficients, where β_6 and β_7 are the coefficients for two continuous variables, β_{26} and β_{27} correspond to two binary variables, and β_{51} and β_{52} correspond to a categorical variable with two categories.

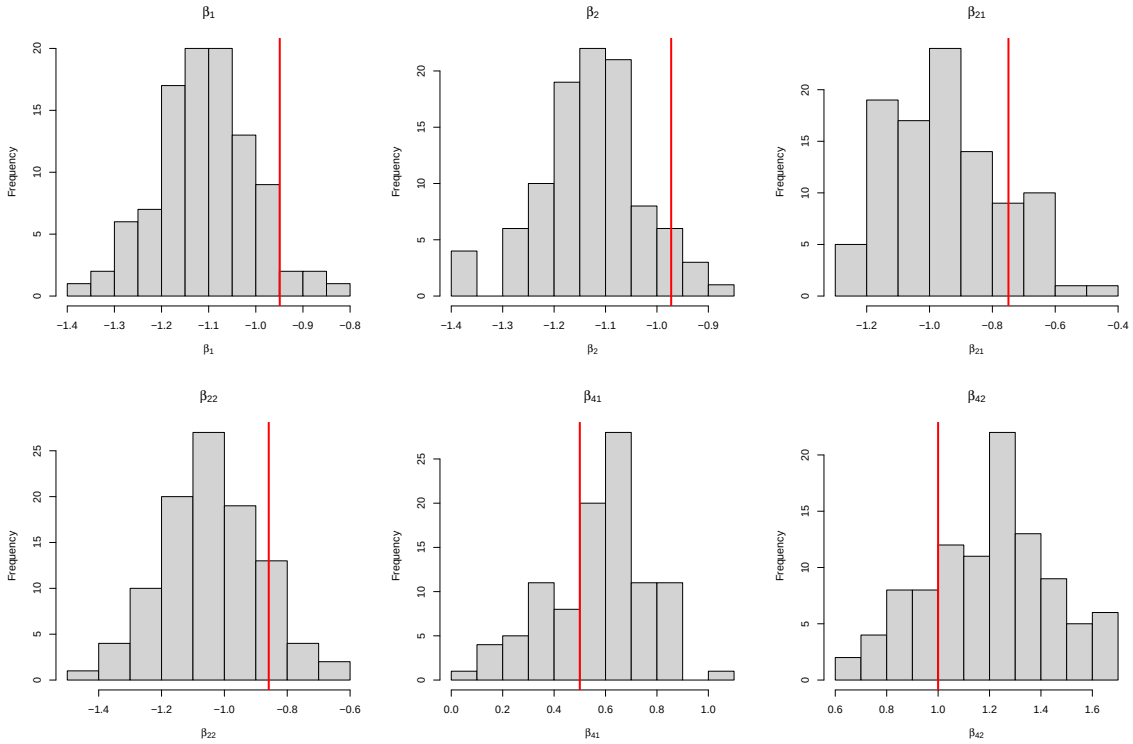

Figure 2: The histograms of the estimates for four non-zero coefficients, where β_1 is the coefficient for a continuous variable, β_{21} corresponds to a binary variable, and β_{41} and β_{42} correspond to a categorical variable with two categories.

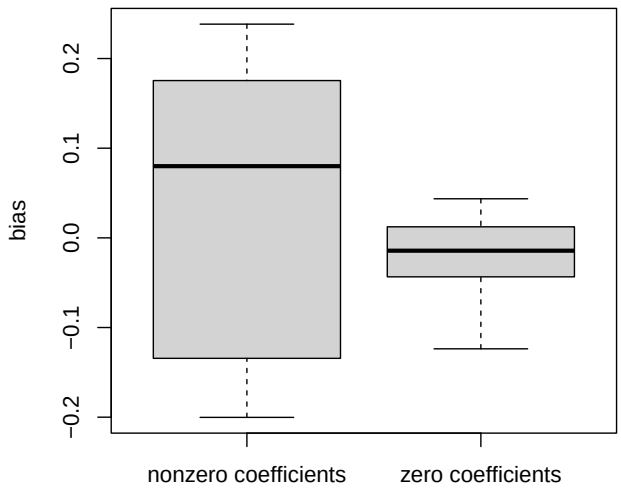

Figure 3: Biases of the 80 latent regression coefficients calculated based on 100 replications. Among the 80 coefficients, 60 have a true value of zero, and 20 have non-zero true values. The non-zero true parameters take values between 0.5 and 1.

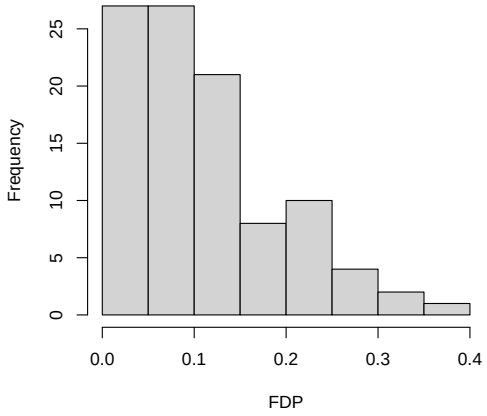

Figure 4: The histogram of false discovery proportion based on the 50 replications.

Example 1. For simplicity, we assume all the background variables are continuous, and there is no missing data, but similar examples can be constructed for categorical variables and when there are missing values. Consider the following setting. We generate data from a sparse latent regression model. Let $N = 20,000$, $J = 20$ and $p = 60$. Assume there is no missingness in the responses and background variables. The background variables are generated by a two-dimensional factor model

$$\mathbf{Z} = A\mathbf{F} + \boldsymbol{\epsilon},$$

where $\mathbf{F}$ follows a two-dimensional standard normal distribution, $A = (a_{ij})_{p\times 2}$ is the loading matrix, and $\boldsymbol{\epsilon}$ is the vector of the independent errors. We generate θ by adding a Gaussian noise to $\beta_0 + \boldsymbol{\beta}^\top \mathbf{Z}$, where only β_1 through β_{10} are non-zero. By our construction, the top two PCs account for 80% of the total variance.

1. We compare the true latent regression model and the misspecified latent regression model in terms of the R^2 value – the proportion of total posterior variance of θ that is explained by the predictors (in the true model, the predictors are Z_1 to Z_{10}; in the PCA model, the predictors are the PC scores).

2. We use the plausible values from the misspecified model for subsequent analysis. In particular, we regress the plausible values on $\mathbf{Z}$. We look at the p-values of the null variables. We also check the significance of the non-null variable.

We see the following results.

1. The R^2 for the true model is 56% while the R^2 for the misspecified model is 12%. This result is due to that the non-null variables (Z_1 to Z_{10}) have small loadings on the factors. Thus, the PC scores do not contain much information about the non-null variables.

2. We note that the non-null variables are very significant that can hardly be missed. However, we also note that there might be a problem with the inference about the null variables in the sense that their p-values do not follow a uniform distribution. Figure 5 provides a histogram for the p-values of null variables (based on 200 independent simulations). We see that the distribution is not uniform, and it

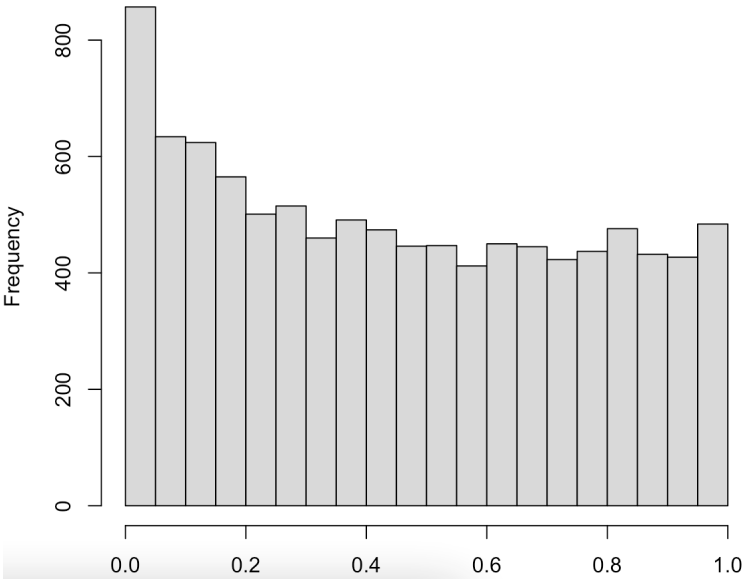

Figure 5: A histogram for the p-values of null variables (based on 200 independent simulations).

skews towards 0. It means that the type I error cannot be controlled with these p-values. This is due to the use of a misspecified model - the effects of the null variables are introduced into the plausible values through the PCs. This issue seems only to become serious when the sample size is large (we have also tried some small sample size settings, where this issue is less severe).

Example 2. In this example, we compare the proposed method and PCA approach based on missing indicators. We consider the same simulation setting as in Section 5.1. For each method, we calculate the correlation between the plausible value of the latent trait and the true latent trait value. For the PCA approach, we use the top PCs that account for 80% of the total variance.

We run 100 independent simulations. The average correlation for the proposed method is 87.3%. And the average correlation for the PCA approach is 85.7%. The proposed method predicts the latent trait slightly better.

Example 3. This example considers the same setting as in Example 2, except for the missing data pattern. In this example, instead of assuming that the data entries are uniformly missing, we assumed that for the first 600 observations (recall that the sample size is 2000), their data on 30 variables are completely missing. This mimics situations, for example, when many parents fail to return the home survey questionnaire. We run 100 independent simulations. The average correlation for the proposed method is 89.0%. And the average correlation for the PCA approach is 67.4%. The proposed method predicts the latent trait much better.

6 Application to PIRLS 2016 Data

6.1 PIRLS 2016 Data

The PIRLS is an international assessment of students' reading achievement for fourth-grade students conducted by the TIMSS & PIRLS International Study Center at Boston College for the International Association for the Evaluation of Educational Achievement (IEA). The assessment is implemented in a five-year cycle, and its data are used widely to inform educational policymaking in the participating countries of the PIRLS.

In this study, we used the PIRLS 2016 dataset for fourth-graded students in the USA to illustrate the use of the proposed method. This dataset contains the responses from three versions of the PIRLS assessment: PIRLS, PIRLS Literacy, and ePIRLS (Mullis and Martin, 2015, Chapter 3). The PIRLS comprehensively evaluates the reading literacy of fourth-grade students. The PIRLS Literacy shares a similar scope to the PIRLS; however, its content is aimed at measuring the reading literacy of the students at the lower end of the achievement scale in the PIRLS. The ePIRLS is an extension of the PIRLS to assess reading skills in online settings. Here, we focused on the dataset from the regular PIRLS.

6.2 Data Description and Handling

The dataset from the PIRLS is divided into two main parts: achievement and background datasets. In the following sections, we summarise the data description and procedure of data handling for each dataset.

6.2.1 Achievement data

For the achievement data, the PIRLS adopted a booklet design to implement its assessment of reading literacy because conducting all the sets of the PIRLS assessment items is infeasible in a limited testing time for fourth-grade students. A booklet is a set of test items carefully chosen from the PIRLS assessment items, and 16 booklets were constructed. Every student was assigned to one of these booklets. Thus, each student was exposed only to one booklet so that the responses to the items in the other booklets that were not administered to the student were all missing by this test-implementation design. We treated such responses as "not implemented" and did not include further analysis for missing-value imputations. However, students also have missing responses in an assigned booklet where some responses are missing because they could not reach the corresponding items during testing time. We treated such missing responses as incorrect ones by following the same procedure of data handling in the PIRLS 2016 for generating proficiency scores (Martin et al., 2017, Chapter 12, p. 7).

In addition, we need to specify a measurement model to generate a latent variable θ_i in the Imputation step of the II algorithm. For a dichotomously scored item without a multiple-choice question, we use a 2PL model. For a dichotomously scored item with a multiple-choice question, we use a 3PL model to account for the probability of a correct response by a random guess. For a polytomous item with ordered scoring (e.g., incorrect, partially correct and correct), we use a GPCM. Since all the item-specific parameters of these three measurement models were calibrated and provided beforehand in the dataset from the PIRLS, we treated these parameters as known. It should be noted that we included only the items whose item-specific parameters are provided.

Lastly, we present the number of students and test items in each booklet. The total number of students is 4425. The number of students and items in each booklet ranges from 237 to 714 and 25 to 34, respectively.

For details on the specifications of the booklets, refer to Mullis and Martin (2015, Chapter 3).

Table 1: The number of students and items in each booklet

Booklet ID	Number of students	Number of items
1	249	28
2	241	29
3	249	31
4	237	29
5	257	34
6	251	25
7	246	25
8	247	27
9	240	26
10	251	28
11	242	26
12	252	30
13	251	25
14	243	30
15	255	30
16	714	32

6.2.2 Background data

The background data consist of four types of PIRLS 2016 context questionnaires: home, teacher, school, and student. Since more than 90% of the responses in the questionnaire about home are missing, we excluded the variables relevant to this questionnaire and used the responses from the other three types of questionnaires for data analysis. However, as we illustrated in Example 3 of Section 5.3, the proposed method may still perform well when including the highly missing home survey data.

We now provide brief descriptions of the other three types of questionnaires quoted from Mullis and Martin (2015, Chapter 3, p. 67–68).

- *Home*: The home questionnaire, which is named the Learning to Read Survey, was given to parents or primary caregivers of students participating in the PIRLS assessment. It asks about the home context, including the languages spoken at home, the parents' attitudes toward reading, and their education and occupation. The questionnaire also asks about the students' educational activities outside of school, such as early childhood education and early literacy and numeracy activities, as well as the child's reading readiness at the beginning of primary school. As mentioned previously, this questionnaire is excluded in the analysis.

- *Teacher*: The teacher questionnaire asks about the classroom context for reading instruction, including class characteristics, reading instructional time, and approaches to instruction. It also gathers information about the teacher's characteristics, such as their career satisfaction, education, and recent professional development activities.

- *School*: The school questionnaire was given to the principal of each school. It asks about the school's characteristics, including student demographics, the school environment, and the availability of resources and technology. It also includes items that focus on the principal's leadership role, education, and experience.

- *Student*: The student questionnaire collects information on their home environment, such as the languages spoken at home, the availability of books, and other resources for learning. It also gathers information on their experiences in school, including their sense of belonging and whether they have experienced bullying. Additionally, the questionnaire gathers data on their out-of-school reading habits and attitudes toward reading, including their level of interest in reading, their confidence in reading, and their engagement in reading lessons.

The procedure of data handling for background variables is as follows. For continuous variables, we treated the variables with more than ten unique values as continuous ones, and they were all standardised. For a categorical variable, if a certain category does not have any observations, we remove that category. In addition, all the ordinal and nominal variables were treated as nominal ones, and they were transformed into dummy variables. It should be noted that the PIRLS 2016 dataset provides context questionnaire scales that summarise a set of items in one latent dimension. For instance, the dataset contains the "Sense of School Belonging" scale that was constructed from the questions such as "I like being in school" and "I feel safe when I am at school" (Mullis and Martin, 2015, Chapter 14, Appendix A). We included such scales in the analysis and removed the variables that were used to construct these scales. In the end, the number of background variables was reduced from 364 to 157 in the analysis. The number of dummy-transformed background variables becomes 345. Lastly, we computed the missing rates of background variables. These missing rates range from 0.043 to 0.541; see Figure 6 for a visualisation.

6.3 Details of Estimation

We first performed the variable selection with FDR control to identify important background variables. Subsequently, we fitted a latent regression model with the II algorithm given the selected background variables to obtain a final model.

6.3.1 Variable Selection with FDR Control

We set a target FDR level to $\alpha = 0.1$ and the number of independent data splittings to $m = 20$. The number of iterations and burn-in iterations was set to $T = 200$, and $B = 50$, by which the convergence of structural parameters was confirmed through a visual check of trace plots and the convergence diagnostics called the rank-normalised split-$\widehat{R}$ (Vehtari et al., 2021). In particular, all the structural parameters satisfied $\widehat{R} < 1.25$.

An initial estimation procedure is performed to obtain the initial values for the II algorithm. For the missing values of background variables, we imputed these entries by the generated samples from the empirical distribution of a corresponding variable. For latent variables, we estimated these values by a grid search on the likelihood value given the three measurement models and their item-specific parameters. A grid of latent variables was set to 0.005. With the complete dataset, we obtained the initial values of relevant parameters in the Sampling and Imputations steps of the II algorithm by using linear regression and logistic and multinomial logistic regression with a ridge penalty, where the ridge penalty was specified to $\lambda = 1500/N$. Here, N denotes the sample size.

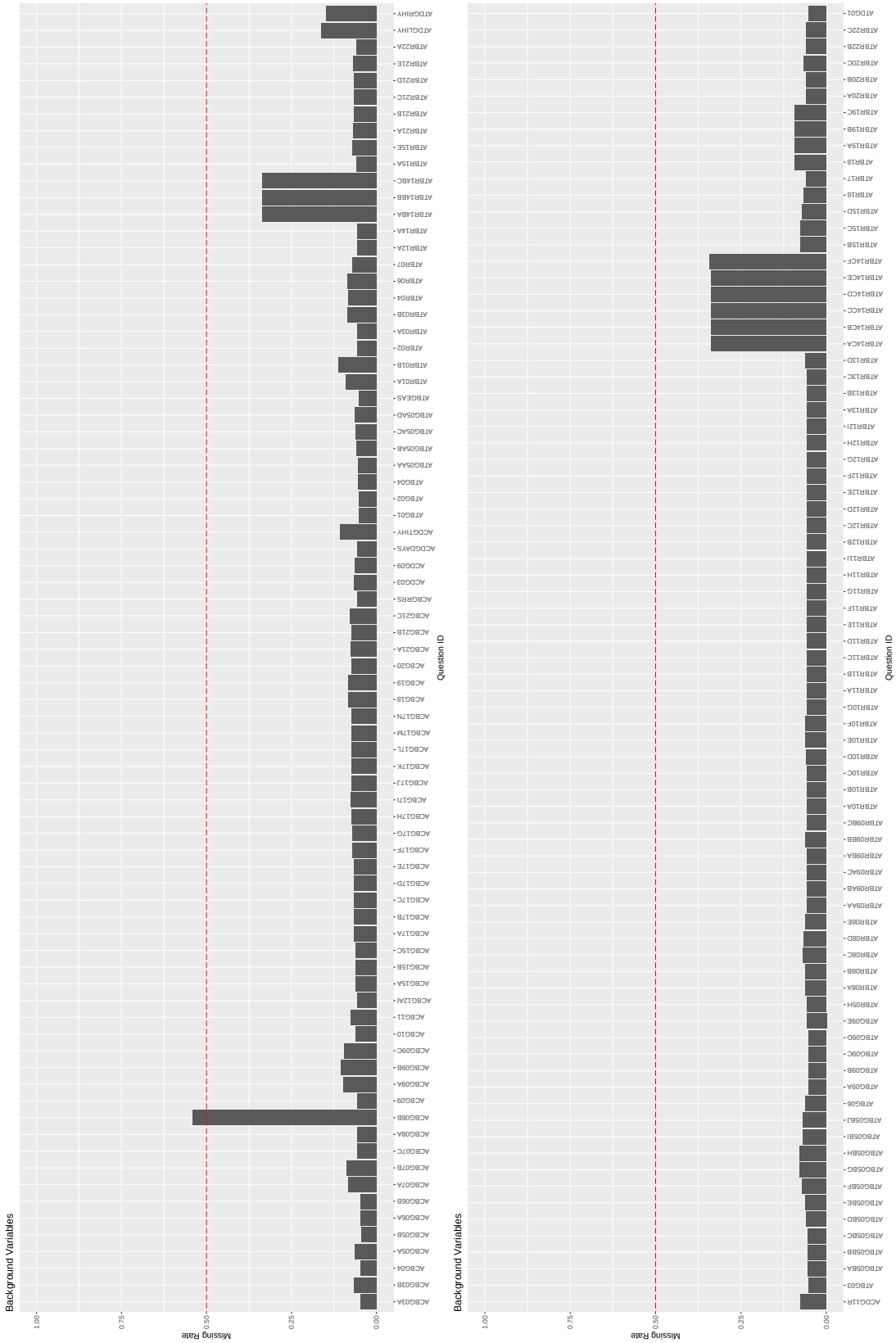

Figure 6: Missing rates of background variables.

6.3.2 Latent Regression Model with Selected Background Variables

For the latent regression model with selected background variables, we also set the number of iterations and burn-in iterations to $T = 200$ and $B = 50$. We also confirmed the convergence of structural parameters through a visual check of trace plots and the convergence diagnostics called the rank-normalised split-$\widehat{R}$ (Vehtari et al., 2021), where all the structural parameters satisfied $\widehat{R} < 1.1$. The same initial values of the II algorithm as in the variable selection with FDR control were used for parameter estimation.

6.4 Result

With the proposed method, twelve variables were selected, and these variables consisted of questions relevant to teachers and schools, while no variable from the student survey was selected. The selected background variables with their inclusion rates, question ID, and contents are given in Table 2. A latent regression model is fitted with the twelve selected variables. Table 3 presents the estimates of the expected a posteriori (EAP) and posterior standard deviation (SD) for the partial regression coefficients in this model.

We note that the selected variables are ordered by their inclusion rates defined in Step 2 of Algorithm 3, as the inclusion rate can be viewed as a measure of variable importance. For example, we see that the top six variables are ATBR12E (teacher asking students to compare what they have read with other things they have read), ACBG03B (the percentages of students in school who come from economically affluent homes), ACBG17N (the grade at which school put emphasis on determining the author's perspective or intention), ACBG07C (the number of days that school is open for instruction), ATBR18 (teachers' expectation on how much time students spend on homework involving reading), and ACBG11 (the number of computers at the school).

The estimates of most parameters in Table 3 are intuitive. For example, for ATBR12E, the negative signs of the parameter estimates, except for the last category "Never or almost never", indicate that a higher frequency of applying this teaching strategy is associated with better reading performance (with the rest of the variables controlled). The exception with the last category "Never or almost never" is likely due to a high uncertainty with this parameter estimate, as the sample size for the last category is relatively small (the sample size for this category is 46). The result for variable ACBG03B suggests that the higher proportion of students from economically affluent homes is associated with better reading performance (with the rest of the variables controlled). The result for ACBG17N further suggests that the earlier the school puts emphasis on determining the author's perspective or intention, the better the reading performance (with the rest of the variables controlled). The estimated coefficients for ACBG07C suggest that the number of days the school is open for instruction matters – the more days the school is open, the better the reading performance (with the rest of the variables controlled). Moreover, teachers' expectations on the amount of time students spending on homework involving reading is positively associated with students' reading performance (with the rest of the variables controlled) – the longer the expected time, the better the performance. Furthermore, the number of computers at school (for students in grade 4) is positively associated with students' reading performance (with the rest of the variables controlled). These findings may help identify the key factors associated with students' reading performance.

The results also reveal some limitations of the current analysis. Possible solutions to these problems are discussed in Section 7. We notice that the method only selected teacher- and school-level variables, while no student-level variables were selected. This result is a little surprising. It is likely due to that we did not introduce school or class-specific random effects into the model, and thus, the teacher- and school-level

variables are selected to compensate for such random effects. Or it could be due to that the individual-level data are reported with larger measurement errors. Consequently, the estimated regression coefficients suffer from attenuation (Chapter 3, Carroll et al., 2006), making them less likely to be selected.

In addition, for certain ordinal predictors, we expect the coefficients of the corresponding dummy variables to have a monotone ordering so that the effect increases or decreases with the ordered categories. However, such monotonicity relationships were not imposed. For example, we may expect the coefficients for the ordinal variable ATBR12E to be negative and monotone decreasing. However, it is not the case when moving from category C3 (-0.159) to C4 (0.216), which is likely due to the relatively smaller sample size for the last category. It may be a good idea to impose some monotonicity constraints in the estimation to achieve better interpretability.

Finally, the result of the current model may also suffer from the issue of collinearity. In particular, both variables ATBG01 (the number of teaching years) and ATBG03 (teacher age) are included in the current model. These two variables have a polyserial correlation of 0.72. Although the signs of the estimated coefficients are consistent with our expectation, the estimated parameter values may be inaccurate due to the high collinearity.

Table 2: Contents of the selected background variables and their inclusion rates

Order	QuestionID	Inclusion Rate	Type	Content
1	ATBR12E	0.224	Teacher	How often do you ask the students to do the following things to help develop reading comprehension skills or strategies? Compare what they have read with other things they have read
2	ACBG03B	0.066	School	Approximately what percentage of students in your school have the following backgrounds? Come from economically affluent homes
3	ACBG17N	0.050	School	At which grade do the following reading skills and strategies first receive a major emphasis in instruction in your school? Determining the author's perspective or intention
4	ACBG07C	0.042	School	For the (fourth grade) students in your school: In one calendar week, how many days is the school open for instruction?
5	ATBR18	0.032	Teacher	In general, how much time do you expect students to spend on homework involving reading (for any subject) each time you assign it?
6	ACBG11	0.025	School	How many computers (including tablets) does your school have for use by (fourth grade) students?
7	ATBG01	0.025	Teacher	By the end of this school year, how many years will you have been teaching altogether?
8	ACBG05B	0.025	School	Which best describes the immediate area in which your school is located?
9	ATBG03	0.025	Teacher	How old are you?
10	ACBG09A	0.017	School	Approximately how many books (print) with different titles does your school library have (exclude magazines and periodicals)?
11	ATBG09B	0.017	Teacher	How often do you have the following types of interactions with other teachers? Observe another classroom to learn more about teaching
12	ATBR08C	0.016	Teacher	When you have reading instruction and/or do reading activities, how often do you organize students in the following ways? I create mixed-ability groups

7 Discussion

In this report, we proposed a variable selection approach to latent regression modelling. The new method tackles the challenges of high dimensionality and data missingness in background variables. In the meantime, it avoids the issues with the state-of-the-art PCA-based approach, including the issues with the missing indicator approach and the PCA. Our method and its implementation are designed under the most flexible setting that allows for different types of measurement items (2PL, 3PL, and GPCM) and different types of background variables (binary, continuous, categorical). Our simulation results confirm that the proposed

Table 3: The estimates of the EAP and posterior SD for the partial regression coefficients in the latent regression model

Order	Question ID	C1	C2	C3	C4	C5	C6
1	ATBR12E	1: Every day or almost every day	2: Once or twice a week	3: Once or twice a month	4: Never or almost never	-	-
		-	-0.044(0.035)	-0.159(0.067)	0.216(0.156)	-	-
2	ACBG03B	1: 0 to 10%	2: 11 to 25%	3: 26 to 50%	4: More than 50%	-	-
		-	0.203(0.044)	0.382(0.051)	0.399(0.059)	-	-
3	ACBG17N	1: First grade or earlier	2: Second grade	3: Third grade	4: Fourth grade	5: Not in these grades	-
		-	-0.052(0.048)	-0.038(0.046)	-0.227(0.068)	-0.004(0.2)	-
4	ACBG07C	1: 6 days	3: 5 days	4: 4 1/2 days	5: 4 days	-	-
		-	-0.224(0.176)	-0.479(0.26)	-0.255(0.216)	-	-
5	ATBR18	1: 15 minutes or less	2: 16–30 minutes	3: 31–60 minutes	-	-	-
		-	0.234(0.062)	0.244(0.075)	-	-	-
6	ACBG11	The number of computers	-	-	-	-	-
		0.003(0.017)	-	-	-	-	-
7	ATBG01	how many years will you have been teaching altogether?	-	-	-	-	-
		0.008(0.027)	-	-	-	-	-
8	ACBG05B	1: Urban–Densely populated	2: Suburban–On fringe or outskirts of urban area	3: Medium size city or large town	4: Small town or village	5: Remote rural	-
		-	0.374(0.057)	0.256(0.053)	0.497(0.06)	0.102(0.095)	-
9	ATBG03	1: Under 25	2: 25–29	3: 30–39	4: 40–49	5: 50–59	6: 60 or more
		-	0.335(0.091)	0.289(0.085)	0.376(0.089)	0.272(0.101)	0.554(0.122)
10	ACBG09A	2: 251–500	3: 501–2,000	4: 2,001–5,000	5: 5,001–10,000	6: More than 10,000	-
		-	0.156(0.125)	-0.186(0.109)	0.098(0.106)	0.044(0.108)	-
11	ATBG09B	1: Very often	2: Often	3: Sometimes	4: Never or almost never	-	-
		-	-0.21(0.087)	-0.076(0.079)	-0.165(0.083)	-	-
12	ATBR08C	1: Always or almost always	2: Often	3: Sometimes	4: Never	-	-
		-	0.039(0.061)	0.043(0.065)	0.071(0.107)	-	-

Note. The estimates of he EAP and posterior SD for the partial regression coefficients are presented in the manner as "EAP (posterior SD)." Additionally, "Order" denotes the rank of selected background variables sorted by the magnitude of their inclusion rates in Table 2. Since the partial regression coefficient associated with the base category of dummy-transformed categorical variables is not estimated, the cells for categorical variables in the "C1" column are blank.

method can control the FDR under a reasonable target level (set to be 0.1 in this report) while maintaining high power. We also applied the proposed method to PIRLS 2016 data. The results from this application look promising, though we also identified some limitations of the new approach to be discussed below.

We did not fully investigate the choice of FDR target level α in this report, though some simulations have been conducted with $\alpha = 0.05$ and 0.2 (results not included in this report). The simulation results under different α levels show that the FDR can be reasonably controlled by the proposed method under these target levels. However, note that FDR control is in an average sense (aggregated over many datasets) since FDR is defined as an expectation. When looking at a specific dataset, we often do not observe a monotone relationship between the α level and the size of the selected set. That is, as α increases, which means a less strict constraint on the FDR, the number of selected variables may sometimes decrease, which is opposite to our expectation (even though the procedure is still asymptotically valid in an average sense). The decrease in the selection variable set (as α increases) happened in the analysis of PIRLS 2016 data. We believe this is due to the current DS aggregation algorithm, and the issue can be fixed if we modify this aggregation procedure based on a method recently proposed in Ren and Barber (2022). We leave it for future investigation.

The analysis of PIRLS 2016 suggests several directions that are worth future investigation. First, we notice that all the selected variables are at the teacher and school levels, whereas no student-level variables are selected. We believe that it does not mean these individual-specific factors are not important to students' reading achievement. Instead, it is likely due to that we did not include school- or class-specific random effects in the latent regression model, and the teacher- and school-level variables are selected to compensate for such random effects, or it could be the individual-level data being reported with larger measurement errors and, thus, suffering from regression attenuation. If the former reason, then we should add these random effects into our model. If the latter, one possible solution is to consider separate FDR measures for the student-level variables and teacher- and school-level variables and select separately from the two sets of variables. Second, we notice that the proposed method may select variables with collinearity. To tackle this issue, we may pre-process the data by removing or merging some variables. Third, for certain ordinal covariates, we may expect the coefficients of the corresponding dummy variables to have a monotone ordering so that the effect increases or decreases with the ordered categories. The estimated model may lack interpretability if the estimated coefficients do not have a monotone relationship. Such monotone constraints are currently not imposed in our variable selection and estimation procedures, making some coefficients in the final model hard to interpret. In the future, we may incorporate such monotone constraints into the estimation and variable selection procedures through tailored prior specifications. Finally, the current estimation procedure – Algorithm 1 – may be further improved by imposing sparsity-inducing priors, such as spike-and-slab priors (Ishwaran and Rao, 2005), on the regression coefficients and parameters of the exponential family graphical model. These priors will be helpful when the sample size is relatively small compared with the number of variables, which may be the case for certain countries in an international large-scale assessment.

We point out that we currently consider a latent regression model with a unidimensional latent trait. This model and the computational algorithms can be easily extended to a multidimensional latent regression model that can be applied to tests with multiple subjects, such as TIMSS. We leave this extension for future investigation.

Admittedly, the proposed method is computationally more demanding than the PCA-based method due to the iterative sampling steps in the proposed method. However, we believe that the computation of the current method is affordable for operational use, thanks to the use of conjugate priors and efficient samplers

like the Pólya-Gamma sampler. The method is currently implemented in the statistical software R, which is convenient but not very suitable for large-scale computation. We believe that the computational time can be substantially reduced if we optimise the implementation in other programming languages that are more suitable for large-scale computation. Furthermore, most steps of the proposed method can be substantially speeded up by parallel computing. We will explore this option to speed up the computation.

Acknowledgement

We thank IEA for supporting this research with a Research and Development Fund. We also thank Prof. Irini Moustaki for providing valuable comments on this research.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman and Hall/CRC, Boca Raton, FL.

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64.

Cohen, J. and Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum, Mahwah, NJ.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, 00(0):1–18. Retrieved from `https://doi.org/10.1080/01621459.2022.2060113`.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum, Mahwah, NJ.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348.

Grund, S., Lüdtke, O., and Robitzsch, A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, 46(4):430–465.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433):222–230.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173.

Lord, F. and Novick, M. (1968). *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA.

Martin, M. O., Mullis, I. V., and Hooper, M., editors (2016). *Methods and procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: `https://timssandpirls.bc.edu/publications/timss/2015-methods.html`.

Martin, M. O., Mullis, I. V. S., and Hooper, M., editors (2017). *Methods and Procedures in PIRLS 2016*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: `https://timssandpirls.bc.edu/publications/pirls/2016-methods.html`.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49(3):359–381.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2):177–196.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2):131–154.

Mullis, I. V. S. and Martin, M. O., editors (2015). *PIRLS 2016 Assessment Framework (2nd ed.)*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: `https://timssandpirls.bc.edu/pirls2016/framework.html`.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176.

OECD (2017). *PISA 2015 Technical Report*. OECD, Paris, France.

OECD (2019). *Technical report of the survey of adult skills (PIAAC)(3rd ed.)*. OECD, Paris, France.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ren, Z. and Barber, R. F. (2022). Derandomized knockoffs: Leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147 – 177.

Tsao, R. (1967). A second order exponential model for multidimensional dichotomous contingency tables with applications in medical diagnosis. *Unpublished doctoral thesis, Harvard University, Department of Statistics*.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In Rutkowski, L., von Davier, M., and Rutkowski, D., editors, *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, pages 175–201. Chapman Hall/CRC, Boca Raton, FL.

Yamaguchi, K. and Zhang, J. (2023). Fully Gibbs sampling algorithms for Bayesian variable selection in latent regression models. *Journal of Educational Measurement*, 60(2):202–234.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847.