

## Plausible Values and Multilevel Models in Large-Scale Assessments

**Khurrem Jehangir** 

*Mohammed bin Rashid University, Dubai*

**Jean-Paul Fox** 

*University of Twente*

*Institutional datasets of large-scale assessments (LSAs) typically contain a set of multiple imputations known as plausible values (PVs) that serve as a proxy for measures of latent proficiency. These are a set of draws from the posterior distribution of latent proficiency that account for measurement error. The PVs are typically drawn from a conditioning model that also includes information from covariates. The advantage of using PVs is that they can be readily used to provide estimates of population quantities of interest at the individual and group levels in secondary analysis of institutional datasets. In this study, the suitability of the PV methodology for secondary multilevel analyses is investigated. It is theoretically shown that consistent estimates for multilevel regression effects are obtained, even when using a single-level conditioning model for the PVs. However, when ignoring the hierarchical structure in the data in constructing PVs, simulation studies showed that the statistical inference is biased and that Type-1 errors, standard errors, and confidence intervals are invalid. The implications for school-level analyses in LSAs are discussed in light of the results.*

**Keywords:** *plausible value methodology; single-level conditioning model; multilevel conditioning model*

### Introduction

The goal of (international) large-scale assessments (LSAs) like National Assessment of Educational Progress (NAEP), Programme for International Student Assessment (PISA) PISA, and Trends in International Mathematics and Science Study (TIMSS) TIMSS are not only to measure the skills assessed but also to collect background information of examinees so that it can be linked to proficiency. LSAs typically achieve broad coverage of the targeted content domain by partitioning the total pool of test items into blocks or clusters of items. Then, each examinee receives only one or more of these blocks that constitute the total assessment pool.

Traditional estimates of student proficiency exclude relationships with student background variables, and subsequently relating those estimates to examinee explanatory variables will result in bias of relevant statistics (Adams et al., 1997; Mislevy, 1991; Von Davier et al., 2009). In an LSA, the estimation procedure of examinee proficiency includes its relationship with covariate information. Therefore, in any secondary analysis, the resulting proficiency estimates can be used to explore relationships with background variables. This estimation technique is based on a latent regression model in which the proficiency variable (also referred to as an ability or person parameter) is related to background variables. In a simultaneous estimation procedure (e.g., marginal maximum likelihood or Markov chain Monte Carlo [MCMC]) the latent regression parameters are estimated together with the item response theory (IRT) model parameters. The method is further formalized by using plausible values (PVs), which are drawn from the posterior distribution of examinee proficiency given item response data and background variables. The PVs are supposed to capture, for instance, the most likely level of proficiency, the standard error, and the relationship with background variables—a set of PVs represents draws of the marginal posterior distribution of a proficiency parameter. PVs were first developed for the analyses of NAEP, by Mislevy and colleagues (Mislevy, 1991; Mislevy, Beaton, et al., 1992; Mislevy, Johnson, & Muraki, 1992) based on Rubin’s work on multiple imputations (Rubin, 1978, 1987).

The PV methodology is based on a conditioning model, which is the posterior distribution of the proficiency variable and combines the IRT measurement component and the latent regression of the proficiency variable. This conditioning model determines the information that can be expected to be included in the PVs. It is well known through simulation studies (Mislevy, 1991) that biased estimates of regression effects are found for explanatory variables that are not included in the conditioning model for constructing the PVs. Obviously, the PVs do not contain information about the relationship between proficiency and any missing predictor variables in the latent regression model. In this paper, it is shown theoretically under a (nonlinear) IRT model that unbiased estimates of regression effects can be expected when the corresponding explanatory variables are included in the conditioning model through the latent regression model. Furthermore, it is shown that when the conditioning model includes the predictor variables but ignores a hierarchical structure of the data (i.e., multilevel data design), unbiased estimates of the regression effects can still be expected.

### *The Optimal Conditioning Model*

Thus, even within a multilevel design, unbiased estimators of regression effects can be constructed using PVs from a single-level conditioning model (SLCM) given that the corresponding predictor variables are included. This

result may indicate that an SLCM, which does not account for any hierarchy in the data, is sufficient for the construction of PVs to analyze relationships between proficiency and background variables in a secondary analysis. However, it has been reported in the literature that the statistical inference can be biased and that Type-1 errors, standard errors, and confidence intervals are invalid. Rutkowski et al. (2010) discussed the most important features of international LSAs. They remarked that under a stratified multistage sampling design—most international LSAs use this design—simple random sampling assumptions for calculating standard errors do not apply. PVs constructed from an SLCM will not account for the hierarchical structure of the data and, therefore, assume independently distributed observations. Those PVs cannot be used to assess, for instance, standard errors of the regression effects in the presence of clustering effects in the data, and therefore misinterpret the information in the data about the uncertainty of the regression effects.

### *Performance of Multilevel Conditioning Models*

Recently, Monseur and Adams (2009) investigated the importance of taking the multilevel structure of the data into account when generating PVs for use in secondary analyses. Their study explored the bias resulting from using point estimates such as the maximum likelihood estimate and expected posterior estimate versus PVs drawn from a multilevel conditioning model (MLCM). However, their simulation study was limited to an intra-class correlation coefficient (ICC) of .4 and a cluster size of 25. Furthermore, they compared point estimates of proficiency with PVs drawn from an MLCM but did not make comparisons with PVs drawn from an SLCM. They showed the superiority of PVs from an MLCM in recovering within-cluster and between-cluster variances, but in comparison to regular proficiency estimates without conditioning. Furthermore, given MLCM PVs, they did not examine the validity of the statistical inference of regression effects in a secondary (multilevel) analysis. In the study of Laukaityte and Wiberg (2017), SLCM PVs are used for a multilevel analysis in which the level-2 variance showed a downward bias and the level-1 variance upward bias (Laukaityte & Wiberg, 2017, Tables 4 and 5). However, they fail to provide an explanation of the result and do not provide an alternative MLCM PV analysis. Zheng (2024) considered different methods to generate MLCM PVs by expanding an SLCM. Dummy-coded variables for cluster membership are included in an SLCM to account for the multilevel design. The disadvantage is that each second-level cluster requires a dummy variable, which can lead to overfitting the latent regression model. Another expansion of the SLCM is to use cluster mean scores—regular proficiency estimates are aggregated to the cluster level—as covariates in the conditioning model. However, the cluster mean scores are estimates, and their uncertainty is ignored in the

construction of the PVs. Both methods showed bias in the estimates of the variance components in the multilevel analysis of the PVs. Their MLCM showed adequate performance but did not outperform the other methods for all considered conditions. Results from the simulation study are limited to the considered sample size, cluster size, and ICC.

### *Objective and Method*

In this study, the performance of the MLCM in comparison to the SLCM is thoroughly investigated through a simulation study, in which the number of test items, the sample and cluster sizes, and the ICC are varied. Bias in statistical inferences is assessed by comparing the statistical results of a secondary analysis using PVs from an SLCM with those from an MLCM. This study also shows that PVs from an MLCM are appropriate for secondary multilevel analysis and that those PVs capture clustering effects in the data. This relates to the observation of Lüdtke et al. (2017) that imputations for missing values should include the multilevel structure of the data to ensure valid statistical inferences of a linear multilevel analysis with imputations for the missing data. However, in the current study, the extension is made to examine the suitability of PVs from a nonlinear multilevel analysis (i.e., multilevel IRT model), in which an IRT model is used to relate the proficiency variable to discrete response patterns.

A Bayesian modeling approach is used to construct PVs. The posterior distribution of the proficiency variable is analytically intractable, and therefore, a data augmentation method is used in which latent response data is sampled (Fox, 2010). Conditional on the latent response data, the posterior distribution of the proficiency variable for a multilevel latent regression model (MLCM) is a normal distribution (Fox, 2010, Chapter 6). This makes the sampling of MLCM PVs straightforward and computationally efficient, which supports a large-scale simulation study regarding the performance of MLCMs.

The paper is organized as follows. A brief overview is given of the limitations of various statistical procedures to adjust standard errors of model-based estimators. Then, a description is given of the IRT models used in this study, which is followed by a mathematical representation of the construction of SLCM and MLCM PVs. It is theoretically shown that SLCM and MLCM PVs produce unbiased estimators for the regression parameters of a multilevel regression model when all predictor variables are included in the conditioning model. Then, the simulation study is presented to compare the performance, in particular the validity of the statistical inferences, of SLCM and MLCM PVs, where different conditions are considered by varying the ICC, the cluster size, and the number of items and persons. After that, empirical examples from an international LSA are provided. Then, the results are discussed, and conclusions are given concerning the conditioning model for constructing PVs.

## **The SLCM in Practice**

In most international LSAs, a stratified multistage sampling design is used (Rutkowski et al., 2010), and this is usually a two-level design (Mang et al., 2021). Therefore, sampling weights can be used in the construction of PVs to get a proper representation of the randomization under the multistage sampling design, instead of the (model-based) randomization under the SLCM. Note that model-based inferences will fail if the sample selection process is ignored. However, there are several reasons why to avoid the use of sampling weights in the construction of PVs.

First, to adjust the SLCM by including sampling weights to achieve exact design unbiasedness is difficult to prove for regression estimators under a non-linear model. At best, the incorporation of weights leads to more efficient estimators. However, using sampling weights unnecessarily can inflate the standard error of parameter estimates, and in that case, the efficiency of estimators is improved by not using weights (Bollen et al., 2016). Furthermore, the statistical inference of a regression analysis can also be poor when the sampling weights differ widely (Holt et al., 1980).

Second, including sampling weights in a multilevel analysis is not straightforward. There are different methods for applying weights to regression effects (Pfeffermann, 1993), but these procedures are not very flexible (Gelman, 2007), and there are no recommendations on which method is to be preferred (Mang et al., 2021). Computing sampling weights is also a complicated procedure, which requires subjective choices (Gelman, 2007).

Third, the most complicating feature of PVs from an SLCM using weights is that they are generated under a multilevel measurement model by weighting the lower-level response observations. The PVs are defined at a higher level than the (weighted) response observations. Therefore, there are no sampling weights defined for the PVs. Weights are only assigned to the response observations from which PVs are constructed. As a result, in a secondary analysis, it is not possible to extract sampling weights applicable to the PVs.

Fourth, when PVs are generated from a conditioning model using sampling weights, the computation of standard errors of regression estimates using the PVs as outcomes also requires sampling weights. Methods for standard error computation, such as the balanced repeated replication, the jackknife, and the bootstrap, are based on resampling from the original sample to estimate the sampling variability. The resampling requires the sampling weights to mimic the design-based randomization. Obviously, in practice, this is not possible when the sample consists of PVs, for which sampling weights are not defined. Furthermore, there is no statistical theory for complex models in which model-based standard errors are adjusted by design-based information. For instance, unbiased estimators of fixed effect parameters in a secondary analysis can be based on PVs constructed under an SLCM. However, the corresponding

standard errors are known to be biased due to ignoring the hierarchical design of the sampling procedure, and statistical theory to adjust the standard error estimates using sampling design features is just evolving (Abadie et al., 2023).

A more fruitful approach avoids the use of sampling weights in the construction of PVs. Therefore, design variables characteristic of the sampling design are incorporated into the (conditioning) model (Pfeffermann, 1989). For instance, by including dummy-coded cluster identifiers or cluster-level proxies in the SLCM, features of a multistage sampling design can be approximated under the SLCM. However, a fixed effect modeling approach to describe multi-level design characteristics is theoretically limited in comparison to a random effects approach. The fixed effects approach requires many more parameters to be estimated, leading to computational challenges (Zheng, 2024), and the borrowing-strength principle is ignored, creating less bias but more (explained) variance by capitalizing on the within-cluster information. This can have the negative effect of over-representing the multilevel design properties. The PVs will simply represent too much between-cluster variance.

The challenges in using sampling weights for secondary analysis and the limitations of incorporating multistage sampling design characteristics in the SLCM motivates the use of an MLCM for the construction of PVs. The MLCM includes a relationship between the sample selection probabilities and the sample data through a random effects specification and is also straightforward to generalize to more complex sampling designs. In the next section, the construction of the PVs under an MLCM is discussed.

## **The Computation of PVs**

### *The Measurement Model*

The (two-parameter) normal-ogive IRT model for dichotomous data (2PNO; Albert, 1992) is used as the measurement model for examinee proficiency. In this 2PNO, an item is characterized by a difficulty parameter  $b_k$  and a discrimination parameter  $a_k$ . The ability parameter of individual  $i$  is represented by  $\theta_i$ . The probability of a correct response to an item  $k$  of person  $i$  (denoted by  $Y_{ik} = 1$ ) is given by

$$\begin{aligned} P(Y_{ik} = 1 | \theta_i, a_k, b_k) &= \Phi(a_k \theta_i - b_k) \\ &= \int_{-\infty}^{a_k \theta_i - b_k} \phi(t) dt, \end{aligned} \tag{1}$$

where  $\Phi()$  and  $\phi()$  are the normal cumulative and normal density function (Fox, 2010, p. 76).

Prior distributions are assumed for the item parameters. For the 2PNO, a normal prior is assumed for the discrimination parameter and difficulty parameters

with vague hyperpriors for the prior's mean and variance parameters. The discrimination parameters are also restricted to be positive.

### *Latent Regression Model*

A (multilevel) latent regression model is assumed for the person parameter (latent variable),  $\theta_{ij}$ , which is considered to be an outcome of a linear regression equation:

$$\begin{aligned}\theta_{ij} &= \boldsymbol{\beta}'\mathbf{X}_{ij} + \boldsymbol{\gamma}'\mathbf{W}_j + u_j + e_{ij} \\ u_j &\sim N(0, \tau^2) \\ e_{ij} &\sim N(0, \sigma^2),\end{aligned}\tag{2}$$

where index  $j$  represents the group in which person  $i$  is located,  $\mathbf{X}_{ij}$  represents (person-specific) predictor variables, and  $\mathbf{W}_j$  represents group-specific variables for person  $i$  in group  $j$ . The random error component  $u_j$  represents the group-specific error and  $e_{ij}$  the person-specific error. Without any random errors at the level of groups, the latent regression model represents a linear regression model for persons in which the error variance at the group level equals zero ( $\tau^2 = 0$ ). Common priors for the linear regression parameters are assumed with normal priors, with vague hyper priors for the regression parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$ , and an inverse-gamma prior for the error variance  $\sigma^2$  and  $\tau^2$ . This latent regression model is also known as the conditioning model for  $\theta_{ij}$  since it represents the distribution for  $\theta_{ij}$  while conditioning on the explanatory variables.

### *Plausible Values*

The PVs are drawn from the posterior distribution of the ability parameter given the item response data and covariates included in the latent regression model. The PVs are random draws from the posterior distribution of the ability parameter, where the draws represent the information about the point estimate for the person parameter but also the uncertainty associated with this point estimate. The object is to use the PVs for secondary analysis without needing the observed item response data since the PVs are assumed to include the relevant information about the person parameters from the item response data.

The posterior distribution of the person parameter can be defined through the model description. That is, for a person  $i$  in cluster  $j$ , given an item response pattern  $\mathbf{y}_{ij}$  and covariates  $\mathbf{X}_{ij}$  and  $\mathbf{W}_j$ , the posterior distribution of the person parameter  $\theta_{ij}$  is defined from the IRT model (Equation 1) and the latent regression model (Equation 2). Then, the posterior distribution is given by

$$h(\theta_{ij} | \mathbf{y}_{ij}, \mathbf{X}_{ij}, \mathbf{W}_j) = \frac{\prod_k P(y_{ijk} | \theta_{ij}) P(\theta_{ij} | \mathbf{X}_{ij}, \mathbf{W}_j)}{\int \prod_k P(y_{ijk} | \theta_{ij}) P(\theta_{ij} | \mathbf{X}_{ij}, \mathbf{W}_j) d\theta_{ij}}, \quad (3)$$

where the other model parameters are already integrated to emphasize the distribution of  $\theta_{ij}$  conditional response data and explanatory variables. In general, the combination of a latent variable (measurement) model and a population model for the latent variable leads via Bayes' theorem to the posterior distribution of the latent variable given observed data and model parameters (Mislevy, 1991). The conditioning model in Equation 2 can also be referred to as the population model for the latent variable.

The posterior distribution in Equation 3 is analytically intractable, which makes it difficult to sample from this distribution. Therefore, a data augmentation algorithm is used to facilitate direct sampling from the posterior distribution of the latent variable given the augmented data. Following Albert (1992) and Fox (2010), an augmented latent response variable  $Z_{ijk}$  is defined, which is assumed to be normally distributed with a mean  $a_k \theta_{ij} - b_k$  and variance one for binary response data  $Z_{ijk} > 0$  if  $Y_{ijk} = 1$  and  $Z_{ijk} \leq 0$  if  $Y_{ijk} = 0$ . Then, the posterior distribution of  $\theta_{ij}$  given that the augmented data is normal with variance

$$\Omega_0 = (\mathbf{a}'\mathbf{a} + \sigma^{-2})^{-1} \quad (4)$$

and mean

$$E(\theta_{ij} | \mathbf{Z}_{ij}, \mathbf{X}_{ij}, \mathbf{W}_j, \mathbf{a}, \mathbf{b}, u_j, \sigma^2) = \Omega_0 ((\mathbf{a}'\mathbf{a})\hat{\theta}_{ij} + \sigma^{-2}(\boldsymbol{\beta}'\mathbf{X}_{ij} + \boldsymbol{\gamma}'\mathbf{W}_j + u_j)), \quad (5)$$

where  $\hat{\theta}_{ij} = (\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}'(\mathbf{Z}_{ij} + \mathbf{b})$ . Using a Gibbs sampling algorithm (Fox, 2010, p. 159), posterior samples can be drawn of the  $\theta_{ij}$ , and these draws will have the distribution according to Equation 3.

The draws are considered MLMC PVs and are independently drawn values  $\theta_{ij}$  from the posterior distribution  $h(\cdot)$ . Historically, 5 or 10 PVs have been drawn in different international LSAs to enable a secondary analysis using these PVs as dependent variables.

The mean term in Equation 5 already reveals the relevance of the (level-2) random intercept  $u_j$ . By conditioning on the random intercept, the PV includes the effect of cluster mean differences in ability. Without the random intercept, an SLCM PV is sampled, which does not account for cluster mean differences. The generalization to random slope effects is straightforward but is not further discussed in this study.

### Errors in PVs

It is theoretically shown that an unbiased estimator for the regression effects can be constructed from the PVs when the corresponding explanatory variables

are included in the SLCM. Subsequently, the same result is shown for PVs constructed from an MLCM. Then, it is shown that even when ignoring the hierarchical structure in the data, an unbiased estimator for the regression effects is still obtained when the SLCM includes the explanatory variables. However, in the latter case, when ignoring the hierarchy in the data in constructing the PVs, inflated Type-1 errors, invalid standard errors, and confidence intervals can be expected when using those PVs in a secondary regression analysis.

Consider (latent) responses  $\mathbf{Z}$  such that the item response model can be translated to a normal linear regression model. When assuming a latent regression model  $\theta_i$ , the model for the latent response data  $\mathbf{Z}$  with a latent regression is expressed as

$$\begin{aligned} Z_{ik} &= a_k \theta_i - b_k + \varepsilon_{ik} \\ \theta_i &= \boldsymbol{\beta}' \mathbf{X}_i + e_i \\ \varepsilon_{ik} &\sim N(0, 1) \\ e_i &\sim N(0, \sigma^2). \end{aligned} \tag{6}$$

Then, the conditional expectation of  $\theta_i$  given  $\mathbf{Z}_i$  can be expressed as

$$\begin{aligned} E(\theta_i | \mathbf{z}_i, \mathbf{X}_i) &= \boldsymbol{\beta}' \mathbf{X}_i + \sigma^2 \mathbf{a}' (\mathbf{I}_K + \sigma^2 \mathbf{a}' \mathbf{a})^{-1} (\mathbf{z}_i - (\mathbf{a} \boldsymbol{\beta}' \mathbf{X}_i - \mathbf{b})) \\ &= \left( \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \boldsymbol{\beta}' \mathbf{X}_i + \left( 1 - \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \hat{\theta}_i \end{aligned} \tag{7}$$

where  $\hat{\theta}_i = (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\mathbf{z}_i + \mathbf{b})$  (see, for more details, Fox, 2010, p. 151). When PVs are sampled from the conditional distribution of  $\theta_i$  given  $\mathbf{z}_i$  and  $\mathbf{X}_i$ , it is shown that the PVs lead to an unbiased estimator for the regression effects  $\boldsymbol{\beta}$ . Computing the expectation of  $E(\theta_i | \mathbf{z}_i)$  over the latent response data, it follows that

$$\begin{aligned} E(E(\theta_i | \mathbf{z}_i, \mathbf{X}_i)) &= \left( \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \boldsymbol{\beta}' \mathbf{X}_i + \left( 1 - \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) E \left( (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\mathbf{Z}_i + \mathbf{b}) \right) \\ &= \left( \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \boldsymbol{\beta}' \mathbf{X}_i + \left( 1 - \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \boldsymbol{\beta}' \mathbf{X}_i \\ &= \boldsymbol{\beta}' \mathbf{X}_i. \end{aligned} \tag{8}$$

Thus, PVs from the conditional distribution of  $p(\theta_i | \mathbf{z}_i, \mathbf{X}_i)$  will lead to correct estimates of the latent regression effects, given that the SLCM contains all predictor variables  $\mathbf{X}_i$ .

In many situations, data is not collected according to simple random sampling, but a stratified and/or cluster sampling approach is taken. Furthermore, the effects of explanatory variables can vary across groups, which leads to

heterogeneity in the data between clusters. When ignoring this, observations cannot be assumed to be independently distributed. The hierarchical structure in the data should also be reflected in the PVs when it leads to violations of independently distributed PVs. For an MLCM, a random effect can be included in the latent regression model to model dependencies between clustered persons. This leads to the so-called multilevel IRT (mlirt) model. The ability of person  $i$  in cluster  $j$  is represented by  $\theta_{ij}$ . The mlirt model for the latent response data with a (multilevel) latent regression model is represented by

$$\begin{aligned} Z_{ijk} &= a_k \theta_{ij} - b_k + \varepsilon_{ijk} \\ \theta_{ij} &= \boldsymbol{\beta}' \mathbf{X}_{ij} + u_j + e_{ij} \\ \varepsilon_{ijk} &\sim N(0, 1) \\ e_{ij} &\sim N(0, \sigma^2) \\ u_j &\sim N(0, \tau^2). \end{aligned} \quad (9)$$

In the same way as in Equation 7, the posterior distribution of  $\theta_{ij}$  given  $\mathbf{z}_{ij}$ ,  $\mathbf{X}_{ij}$  and  $u_j$  is again normal, and its posterior expectation is given by

$$\begin{aligned} E(\theta_{ij} | \mathbf{z}_{ij}, \mathbf{X}_{ij}, u_j) &= \boldsymbol{\beta}' \mathbf{X}_{ij} + u_j + \sigma^2 \mathbf{a}' (\mathbf{I}_K + \sigma^2 \mathbf{a}' \mathbf{a})^{-1} (\mathbf{z}_{ij} - (\mathbf{a} (\boldsymbol{\beta}' \mathbf{X}_{ij} + u_j) - \mathbf{b})) \\ &= \left( \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) (\boldsymbol{\beta}' \mathbf{X}_{ij} + u_j) + \left( 1 - \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \hat{\theta}_{ij} \\ &= \Lambda (\boldsymbol{\beta}' \mathbf{X}_{ij} + u_j) + (1 - \Lambda) \hat{\theta}_{ij}, \end{aligned} \quad (10)$$

where  $\hat{\theta}_{ij} = (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\mathbf{z}_{ij} + \mathbf{b})$  and  $\Lambda = (\mathbf{a}' \mathbf{a})^{-1} / ((\mathbf{a}' \mathbf{a})^{-1} + \sigma^2)$ . The posterior expectation is taken of the random effect  $u_j$  to determine the posterior expectation  $E(\theta_{ij} | \mathbf{z}_{ij}, \mathbf{X}_{ij})$ . In Fox (2010, pp. 188–189) an expression is derived for the (marginal) posterior expectation

$$E(u_j | \mathbf{z}_j, \mathbf{X}_j) = \left( \frac{n_j \tau^2}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2 + n_j \tau^2} \right) \hat{u}_j,$$

where  $\hat{u}_j = (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\bar{\mathbf{z}}_j - (\mathbf{a} \overline{\boldsymbol{\beta}' \mathbf{X}_j} - \mathbf{b}))$ , and  $\bar{\mathbf{z}}_j = \sum_{i=1}^{n_j} \mathbf{z}_{ij} / n_j$ . This is used to determine the posterior expectation  $E(\theta_{ij} | \mathbf{z}_{ij}, \mathbf{X}_{ij})$ . Therefore, this expression for the posterior expectation is integrated into Equation 10, and it follows that

$$\begin{aligned} E(\theta_{ij} | \mathbf{z}_j, \mathbf{X}_{ij}) &= \Lambda (\boldsymbol{\beta}' \mathbf{X}_{ij} + E(u_j | \mathbf{z}_j, \mathbf{X}_j)) + (1 - \Lambda) \hat{\theta}_{ij} \\ &= \Lambda \boldsymbol{\beta}' \mathbf{X}_{ij} + (1 - \Lambda) \hat{\theta}_{ij} + \Lambda \left( \frac{n_j \tau^2}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2 + n_j \tau^2} \right) \\ &\quad (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\bar{\mathbf{z}}_j - (\mathbf{a} \overline{\boldsymbol{\beta}' \mathbf{X}_j} - \mathbf{b})). \end{aligned} \quad (11)$$

When taking the expectation over the latent response data  $\mathbf{Z}_{ij}$ , it follows that

$$\begin{aligned}
 E(E(\theta_{ij}|\mathbf{Z}_{ij}, \mathbf{X}_{ij})) &= \Lambda \boldsymbol{\beta}' \mathbf{X}_{ij} + (1 - \Lambda) E(\hat{\theta}_{ij}|\mathbf{Z}_{ij}, \mathbf{X}_{ij}) + \Lambda \Omega(\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' \\
 &\quad \left( E(\bar{Z}_j) - (\mathbf{a} \boldsymbol{\beta}' \bar{\mathbf{X}}_j - \mathbf{b}) \right) \\
 &= \Lambda \boldsymbol{\beta}' \mathbf{X}_{ij} + (1 - \Lambda) \boldsymbol{\beta}' \mathbf{X}_{ij} \\
 &= \boldsymbol{\beta}' \mathbf{X}_{ij},
 \end{aligned} \tag{12}$$

since  $E(\bar{Z}_j) = \sum_{i=1}^{n_j} E(Z_{ij})/n_j = \overline{\mathbf{a} \boldsymbol{\beta}' \mathbf{X}_{ij}} - \mathbf{b}$ . Thus, also for the mlirt model in Equation 9, PVs that are sampled from the conditional distribution of  $\theta_{ij}$  given  $\mathbf{z}_{ij}$  and  $\mathbf{X}_{ij}$ , lead to an unbiased estimator for the regression effects  $\boldsymbol{\beta}$ . The integration over the latent response data  $\mathbf{Z}$  with respect to the posterior distribution  $p(\mathbf{z}_{ij}|\mathbf{y}_{ij})$  does not change this property of unbiasedness of the PVs for the regression parameters  $\boldsymbol{\beta}$ . The posterior expectation of the  $\theta_{ij}$  is not affected by the integration of the latent response data.

Mislevy (1991) showed bias in regression effects when using PVs from a conditioning model with missing predictor variables. Bias appeared in the estimated regression effects corresponding to the missing predictor variables, and the bias depended on the reliability of the PVs. This bias is zero when the PVs are perfectly measured, when the predictor has no relationship with the dependent variable, or when other predictor variables perfectly correlate with the missing predictor variables. The bias is shown under a structural linear errors-in-variables regression model (Gleser, 1992), in which it is assumed that the predictor variables for each person are assumed to be (multivariate) normally distributed with a common mean and common covariance matrix.

#### *More Errors in PVs*

It was already shown in Equation 8 that when the PVs are sampled from an SLCM, given the explanatory variables, unbiased estimates of the fixed effect parameters are obtained. Furthermore, in Equation 12 it was shown that unbiased estimates of the fixed effects are also obtained under a cluster sampling design when PVs are sampled from an MLCM given the explanatory variables. Even when ignoring the multilevel nature of the sampling design, with  $E(u_j|\mathbf{z}_j, \mathbf{X}_j) = 0$  in Equation 11, the estimates of the regression effects will still be unbiased.

However, when the hierarchical structure in the data is ignored when constructing the PVs, it can be expected that the standard errors of the fixed effect estimates will be biased when doing a (secondary) multilevel analysis on those PVs. To identify the (multilevel) errors in the PVs, the following latent multilevel model is considered:

$$\begin{aligned}
 \theta_{ij} &= \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + u_j + e_{ij} \\
 u_j &\sim N(0, \tau^2) \\
 e_{ij} &\sim N(0, \sigma^2),
 \end{aligned} \tag{13}$$

where  $\theta_{ij}$  represents the ability of subject  $i$  in group  $j$ . The  $W_j$  is the explanatory variable at the group level and the  $X_{ij}$  at the subject level. The ICC is represented by the ratio  $\tau^2/(\sigma^2 + \tau^2)$  and represents the proportion of explained variance by the clustering of subjects in groups. The abilities are not directly observed, instead, response patterns are observed. Then, according to a two-parameter IRT model and assuming latent response data  $\mathbf{Z}_{ij}$ , the measurement model for the  $\theta_{ij}$  is given by

$$\begin{aligned}
 \mathbf{Z}_{ij} &= \mathbf{A}\theta_{ij} - \mathbf{B} + \boldsymbol{\varepsilon}_{ij} \\
 \text{var}(\mathbf{Z}_{ij}|\theta_{ij}) &= \mathbf{A}\Sigma_{\theta}\mathbf{A}^t + \mathbf{I}_K,
 \end{aligned} \tag{14}$$

where the covariance matrix  $\Sigma_{\theta}$  includes the error variance  $\sigma^2$  but also the variance of the group effect  $u_j$  according to Equation 13. The item discrimination and difficulty parameters are stored in matrices  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. Assume an approximate normal distribution for the PV,  $\theta_{PV}$  with mean  $\theta$  and variance  $\Sigma_{PV}$ , which is the marginal posterior distribution of  $\theta_{PV}$  given data. The measurement errors  $\boldsymbol{\varepsilon}$  in Equation 14 are assumed to be independent of the measurement errors of the PVs,  $\text{cov}(\boldsymbol{\varepsilon}, \mathbf{e}_{PV}) = 0$ . The distribution of the measurement errors and the PV errors is multivariate normal with mean zero and a diagonal covariance matrix with components  $\Sigma_{\boldsymbol{\varepsilon}}$  and  $\Sigma_{PV}$ .

The multivariate normal distribution of  $\mathbf{Z}_{ij}$  and  $\theta_{PV}$  has mean

$$E(\mathbf{Z}_{ij}, \theta_{PV}) = (\mathbf{A}(\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij}) - \mathbf{B}, \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij}) \tag{15}$$

and covariance matrix

$$\text{var}(\mathbf{Z}_{ij}, \theta_{PV}) = \begin{pmatrix} \mathbf{A}\Sigma_{\theta}\mathbf{A}^t + \Sigma_{\boldsymbol{\varepsilon}} & \mathbf{A}\Sigma_{\theta} \\ \Sigma_{\theta}\mathbf{A}^t & \Sigma_{PV} + \Sigma_{\theta} \end{pmatrix}. \tag{16}$$

Then, the reliability matrix  $\Lambda = (\Sigma_{\theta} + \Sigma_{PV})^{-1}\Sigma_{\theta}$  plays an essential role in the estimation of the regression parameters  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\boldsymbol{\gamma}$ . Its importance follows from the conditional distribution of  $\mathbf{Z}_{ij}$  given  $\theta_{PV}$ , which has mean

$$\begin{aligned}
 E(\mathbf{Z}_{ij}|\theta_{PV}) &= \mathbf{A}\theta_{ij} - \mathbf{B} + \mathbf{A}\Lambda(\theta_{PV} - \theta_{ij}) \\
 &= \mathbf{A}\theta_{ij}(\mathbf{I} - \Lambda) - \mathbf{B} + \mathbf{A}\Lambda\theta_{PV} \\
 &= \mathbf{A}(\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij})(\mathbf{I} - \Lambda) - \mathbf{B} + \mathbf{A}\Lambda\theta_{PV}
 \end{aligned} \tag{17}$$

and covariance matrix

$$\text{var}(\mathbf{Z}_{ij}|\theta_{PV}) = \mathbf{I}_K + \mathbf{A}\Sigma_{\theta}\Lambda\mathbf{A}^t. \tag{18}$$

From the regression follows that estimates of  $\gamma$  are unbiased only when  $\Lambda = \mathbf{I}$ . However, another restriction is that the conditional distribution of  $\mathbf{Z}_{ij}$  given the  $\theta_{PV}$  is assumed to be independent from the other response patterns in group  $j$  since the PV is assumed to include the information from the group component  $u_j$ . Thus, incorrectly assuming independently distributed PVs in the regression of  $\mathbf{Z}_{ij}$  on the explanatory variables in Equation 17 will lead to an underestimation of the standard errors of the fixed effect estimates (assuming positive correlation between the PVs).

Ignoring a small positive or negative clustering leads to the incorrect assumption that the observations are independently distributed. Nielsen et al. (2021) demonstrated that any violation of the independence assumption (positive and negative) results in inaccurate Type-I errors, which increase the risk of accepting an incorrect hypothesis. Barcikowski (1981) quantified the effects of ignoring small positive correlations in clustered observations in a two-level study design (with a group and an individual level). For instance, when having 10 observations per group, even the ignorance of an ICC as small as .01 can lead to an inflation of making a Type-I error: A regression effect will be assumed to be significant with a significance level of 5% although the true significance level equaled 6%. Furthermore, by increasing the number of observations per group, the Type-I error is even more inflated (the findings of Barcikowski are in line with those of many others (see e.g., Clarke, 2008; Dorman, 2008)).

### **Simulation Study**

In a simulation study, various types of errors were assessed in a secondary multilevel analysis when using PVs from an SLCM in comparison to using PVs from an MLCM. Item response data was simulated according to a mlirt model, consisting of a two-parameter IRT model, Equation 1, with an MLCM for the person parameters, according to Equation 13. The PVs were simulated from an MLCM and an SLCM, including the explanatory variables  $X_{ij}$  and  $W_j$ . Without missing covariate information, the results from the secondary multilevel analysis of Equation 13 using the MLCM and SLCM PVs were expected to show unbiased estimates for the regression effects  $\gamma_{00}$  (intercept),  $\gamma_{10}$  (level-1 effect of  $X_{ij}$ ), and  $\gamma_{01}$  (level-2 effect of  $W_j$ ). However, when ignoring the hierarchy in the data, the estimation results based on the SLCM PVs were expected to show incorrect Type-I errors, invalid standard errors, and confidence intervals.

In this study, the number of test items ( $K = 10, 20, 30$ ), the number of subjects ( $N = 500, 1,000, 3,000$ ), and the ICC ( $ICC = .1, .35, .6$ ) were varied. The cluster size was also varied for the different number of subjects, and for  $N = 500, 1,000$ , and  $3,000$ , the cluster sizes were  $n = 5, 10$ , and  $30$ , respectively. The motivation for selecting these simulation conditions was to try and cover the range encountered in real-life international surveys. For instance, the number of items in the

PISA 2006 Reading test ranged between 14 and 28 across the various PISA 2006 test booklets. The ICC values in the simulation study varied from .1 to .6 because, for instance, in the PISA 2009 Reading test the ICC values across countries varied between .09 and .67 (see PISA [2009] International Report Volume 4, Table 4.2.2a, p. 160). The number of students per country in PISA 2009 varied between 329 students to 38,250 students, and the average cluster size varied between 20 and 30 across most countries (see PISA 2009 International Report, Volume 4, Table A2.3, p. 133) while the variation in cluster size within individual countries was sometimes large. For each simulation condition, a total of 500 data replications were performed and reported results were averaged over the 500 replications. For each data set, a total of 10 independent PVs were drawn under an MLM and an SLM. The average across the independent parameter estimates given the PVs was considered the final estimate. For each parameter, the posterior variance was estimated by the sum of the within-imputation and between-imputation variance (see Fox, 2010, pp. 168–169). The complete algorithm of the simulation study is described in Appendix A.

The *MLIRT* software of Fox (2007) was used, which was adapted to make it suitable for generating PVs from an MLM and an SLM. A (standard) Gibbs sampler was used to fit the multilevel model in Equation 13 given PVs (Zeger & Karim, 1991). Proper inverse-gamma priors were used for the variance components of the multilevel model with shape and scale parameters equal to .01. The priors for the mlirt model are described in Fox (2010, chapter 6). The number of MCMC iterations for each simulation case was 5,000 with a burn-in of 500 iterations. The convergence and effective sample size for each data set were examined to ensure that reliable estimation results were obtained. The PVs under the MLM and the SLM were rescaled such that the mean and variance were equal to those of the simulated scale of the person parameters. Therefore, the simulation results of the secondary analysis under both PVs were directly comparable to each other and to the true simulated values.

The simulation results were reported for the regression effects ( $\gamma_{00} = -.25$ ,  $\gamma_{01} = .4$ ,  $\gamma_{10} = .1$ ) and the variance parameters at level 1 ( $\sigma^2 = 1.8$ ) and level 2 ( $\tau^2$ ). The effect of the intercept is represented by  $\gamma_{00}$ , for the level-1 effect by  $\gamma_{10}$ , and for the level-2 effect by  $\gamma_{01}$ . For each data set, a level-1 explanatory variable ( $X_{ij}$ ) and a level-2 explanatory variable ( $W_j$ ) were each simulated from a normal distribution with a standard deviation of 0.5. The  $\tau^2$  was varied to .2, .97 and 2.7, which led to an ICC of .1, .35 and .60, respectively.

In Table 1, the Root Mean Squared Error (RMSE) results for the regression effects are reported for the 500 replications. It was verified that the bias for each regression parameter under MLM and SLM was approximately zero (i.e.,  $< .000$ ). Thus, without missing explanatory variables in the conditioning model, the (average) point estimates of the regression effects were equal to the (true) simulated values. Subsequently, the RMSEs under both PVs were also

TABLE 1.  
*RMSE for the Fixed Effect Estimates Given PVs From MLCM and SLCM*

K	ICC	N	MLCM			SLCM		
			$\gamma_{10}$	$\gamma_{01}$	$\gamma_{00}$	$\gamma_{01}$	$\gamma_{00}$	
10	.1	500	.12	.14	.06	.13	.14	.06
		1,000	.09	.11	.05	.09	.11	.05
		3,000	.05	.08	.04	.05	.08	.04
	.35	500	.14	.21	.08	.14	.21	.08
		1,000	.09	.18	.08	.09	.18	.08
		3,000	.05	.17	.08	.05	.17	.08
	.6	500	.15	.32	.15	.16	.32	.15
		1,000	.11	.31	.14	.11	.30	.14
		3,000	.05	.32	.13	.06	.30	.13
20	.1	500	.12	.13	.06	.12	.13	.06
		1,000	.08	.11	.05	.09	.11	.05
		3,000	.04	.08	.04	.04	.08	.04
	.35	500	.13	.19	.09	.13	.19	.09
		1,000	.08	.19	.08	.08	.19	.08
		3,000	.05	.17	.08	.05	.17	.08
	.6	500	.14	.30	.13	.14	.30	.13
		1,000	.10	.30	.14	.10	.30	.14
		3,000	.05	.31	.12	.05	.29	.12
30	.1	500	.11	.13	.06	.11	.13	.06
		1,000	.08	.11	.05	.08	.11	.05
		3,000	.04	.08	.04	.04	.08	.04
	.35	500	.13	.19	.10	.13	.19	.10
		1,000	.08	.19	.08	.08	.19	.08
		3,000	.04	.17	.08	.04	.17	.08
	.6	500	.13	.30	.13	.14	.30	.13
		1,000	.09	.30	.13	.09	.29	.13
		3,000	.05	.32	.14	.05	.31	.14

*Note.* ICC = intra-class correlation coefficient; MLCM = multilevel conditioning model; PV = plausible value; SLCM = single-level conditioning model.

comparable, which followed from the fact that the PVs were drawn on the same scale. However, for an ICC of .60 and *N* greater than 500, the RMSE for the level-2 effect was slightly higher for the MLCM PV than that of the SLCM. The number of items did not lead to differences in RMSEs of the regression effects.

The 95% highest posterior density (HPD) interval for each fixed effect under the MLCM PV and the SLCM PV was computed. For each fixed effect parameter, structural differences in 95% HPDs between MLCM and SLCM were examined. Therefore, the proportion across replicated data sets was computed

TABLE 2.  
SE and 95% HPD Interval Comparison for PVs Under MLCM and SLCM

95% HPD			$L_{.95}(\text{MLCM} > \text{SLCM})$			$SE(\text{MLCM} > \text{SLCM})$			
$K$	ICC	$N$	$\gamma_{10}$	$\gamma_{01}$	$\gamma_{00}$	$\gamma_{10}$	$\gamma_{01}$	$\gamma_{00}$	
10	.1	500	.51	.36	.30	.37	.80	.73	
		1,000	.56	.20	.05	.22	.98	.96	
		3,000	.55	.04	.00	.14	1.00	1.00	
	.35	500	.64	.17	.01	.00	1.00	1.00	
		1,000	.67	.10	.00	.00	.99	1.00	
		3,000	.68	.11	.00	.00	.99	1.00	
	.6	500	.79	.17	.00	.00	.99	1.00	
		1,000	.81	.20	.00	.00	.99	1.00	
		3,000	.78	.30	.00	.00	.90	1.00	
	20	.1	500	.56	.35	.28	.34	.83	.75
			1,000	.56	.26	.11	.26	.97	.94
			3,000	.54	.13	.01	.24	1.00	1.00
.35		500	.68	.23	.05	.00	.99	.99	
		1,000	.64	.19	.01	.00	.99	1.00	
		3,000	.69	.24	.00	.00	.93	1.00	
.6		500	.78	.23	.02	.00	.98	1.00	
		1,000	.79	.26	.00	.00	.94	1.00	
		3,000	.75	.35	.00	.00	.79	1.00	
30		.1	500	.52	.39	.33	.35	.75	.72
			1,000	.50	.30	.19	.33	.94	.94
			3,000	.58	.19	.06	.31	.96	1.00
	.35	500	.62	.26	.08	.02	.97	.99	
		1,000	.68	.19	.03	.01	.96	1.00	
		3,000	.68	.24	.01	.01	.88	1.00	
	.6	500	.77	.29	.03	.00	.94	1.00	
		1,000	.76	.28	.01	.00	.89	1.00	
		3,000	.76	.44	.01	.00	.71	1.00	

Note.  $L_{.95}(\text{MLCM} > \text{SLCM})$  represents the proportion that the 95% HPD lower bound under the MLCM PV is greater than the one under the SLCM PV.  $SE(\text{MLCM} > \text{SLCM})$  represents the proportion that the SE under the MLCM PV is greater than that of the SLCM PV. HPD = highest posterior density; ICC = intra-class correlation coefficient; MLCM = multilevel conditioning model; PV = plausible value; SLCM = single-level conditioning model; SE = Standard error.

that the lower bound of the 95% HPD under the MLCM was greater than the one under the SLCM. This statistic was computed for each regression effect and represented in Table 2 under the header  $L_{.95}(\text{MLCM} > \text{SLCM})$ . When the proportion was around .50, the width of the 95% HPDs under both conditioning models was, on average, equal. When the  $L_{.95}(\text{MLCM} > \text{SLCM})$  equaled zero, the lower bound under the MLCM of the 95% HPD was consistently lower than that under the SLCM across data replications. As a consequence, the 95% HPD

intervals were wider under the MLCM than under the SLCM. When the  $L_{.95}(MLCM > SLCM)$  equaled one, the 95% HPD interval was consistently tighter under the MLCM than under the SLCM. When the  $L_{.95}(MLCM > SLCM)$  was greater (lower) than .50, the 95% HPD interval was on average tighter (wider) under the MLCM than under the SLCM, but not for all replicated data sets due to sampling error.

For the level-2 regression effect, the average proportion is around .50 for a low ICC. However, when increasing the ICC, the lower bound under the MLCM is, on average, lower than under the SLCM. This is also true for the intercept. As a result, the width of the 95% HPD interval for the level-2 regression effect and the intercept is on average wider under the MLCM than under the SLCM.

The standard error (SE)  $SE$  of the regression effects was also estimated under both PVs. To compare the size of the  $SE$ s, the proportion was computed that the  $SE$  under MLCM was greater than that of the SLCM across data replications for all fixed effect parameters, and reported under the label  $SE(MLCM > SLCM)$ . It follows that the  $SE$  for the intercept and level-2 effect are consistently higher under the MLCM PV in comparison to the SLCM PV. Thus, ignoring the hierarchy in the data in constructing the PVs leads to incorrect Type-I errors and  $SE$ s of the level-2 regression effect and intercept. The PVs sampled under the SLCM assume independence and do not take into account the correlation between the response patterns of cluster members. This leads to smaller  $SE$ s of the intercept and level-2 regression effect than for PVs sampled under the MLCM. Furthermore, for the same reason, the 95% HPD interval for the intercept and level-2 effect is tighter given SLCM PVs in comparison to MLCM PVs.

For the level-1 regression effect, an opposite effect occurred. For the MLCM PVs, the  $SE$ s are on average smaller than those of the SLCM PVs. The lower bound of the 95% HPD is also on average somewhat higher for an increasing ICC under the MLCM PVs. The SLCM PVs did not account for dependence between response patterns within each cluster. This led to more unexplained residual variance within a cluster—the total variance in each vector of PVs was fixed to the simulated total variance—which led to inflated  $SE$ s and a wider 95% HPD interval for the level-1 regression effect.

Next to the difference in the inference of the regression effects between the use of MLCM and SLCM PVs, there was also a difference in the variance component estimates  $\sigma^2$  and  $\tau^2$ . The analysis using the MLCM PVs was expected to produce the true variance estimates since the MLCM PVs were generated under the multilevel design. Subsequently, the MLCM PVs were expected to carry the true ICC. In Table 3, the bias (%bias) is reported for the variance components, where the bias represents the average difference between the estimate and the true value (%bias is the bias in percentage divided by the true value). It follows for the residual (level-1) variance that the bias is very small (approximately

TABLE 3.  
*Assessment of Bias of the Variance Component Estimates Using PVs From MLCM*

<i>K</i>	ICC	<i>N</i>	Bias $\sigma^2$	% Bias	Bias $\tau^2$	%Bias	Bias ICC	%Bias
10	.1	500	.05	3	-.05	-25	-.02	-20
		1,000	.01	1	.00	0	.00	0
		3,000	.00	0	.00	0	.00	0
	.35	500	.02	1	.00	0	.00	0
		1,000	.00	0	.02	2	.00	0
		3,000	.00	0	.02	2	.00	0
	.6	500	.01	1	.05	2	.00	0
		1,000	.01	1	.05	2	.00	0
		3,000	.00	0	.05	2	.00	0
20	.1	500	.05	3	-.04	-20	-.02	-20
		1,000	.01	1	.00	0	.00	0
		3,000	.00	0	.00	0	.00	0
	.35	500	.02	1	.01	1	.00	0
		1,000	.00	0	.02	2	.00	0
		3,000	.00	0	.02	2	.00	0
	.6	500	.01	1	.04	1	.00	0
		1,000	.00	0	.05	2	.00	0
		3,000	.00	0	.05	2	.00	0
30	.1	500	.05	3	-.04	-20	-.02	-20
		1,000	.01	1	.00	0	.00	0
		3,000	.00	0	.00	0	.00	0
	.35	500	.01	1	.01	1	.00	0
		1,000	.00	0	.02	2	.00	0
		3,000	.00	0	.02	2	.00	0
	.6	500	.01	1	.05	2	.00	0
		1,000	.00	0	.05	2	.00	0
		3,000	.00	0	.05	2	.00	0

*Note.* ICC= intra-class correlation coefficient; MLCM= multilevel conditioning model; PV= plausible value.

zero) unless the sample size is small ( $N=500$ ) when the ICC is small (ICC= .10), and when the ICC is large (ICC= .60). In both cases, the PVs from MLCM had difficulties in capturing the true level-1 variance. This relates to the feature that the bias for the level-2 estimates was also apparent in those scenarios.

In Table 4 the estimation results are presented for the variances and the intra-class correlations using PVs from the SLCM. A typical pattern is visible, which shows that the SLCM PVs lack a proper representation of the two-level structure in the data. The estimated level-1 residual variance shows upward bias, and the bias increases for an increasing ICC and decreasing cluster size. The estimated level-2 variance shows a downward bias, and the bias also increases for

TABLE 4.  
*Assessment of Bias of the Variance Component Estimates Using PVs From SLCM*

<i>K</i>	ICC	<i>N</i>	Bias $\sigma^2$	% Bias	Bias $\tau^2$	% Bias	Bias ICC	%Bias
10	.1	500	.11	6	-.10	-50	-.05	-50
		1,000	.09	5	-.09	-45	-.04	-40
		3,000	.08	4	-.07	-35	-.03	-30
	.35	500	.36	20	-.34	-35	-.13	-37
		1,000	.33	18	-.33	-34	-.12	-34
		3,000	.33	18	-.32	-33	-.11	-31
	.6	500	.91	51	-.88	-33	-.19	-32
		1,000	.88	49	-.87	-32	-.19	-32
		3,000	.88	49	-.86	-32	-.19	-32
20	.1	500	.10	6	-.09	-45	-.04	-40
		1,000	.07	4	-.06	-30	-.03	-30
		3,000	.05	3	-.04	-20	-.02	-20
	.35	500	.26	14	-.25	-26	-.09	-26
		1,000	.26	14	-.24	-25	-.09	-26
		3,000	.25	37	-.23	-65	-.08	-25
	.6	500	.69	38	-.66	-24	-.15	-25
		1,000	.66	37	-.63	-23	-.14	-23
		3,000	.66	37	-.63	-23	-.14	-23
30	.1	500	.09	5	-.08	-40	-.04	-40
		1,000	.06	3	-.05	-25	-.03	-30
		3,000	.04	2	-.03	-15	-.01	-10
	.35	500	.20	11	-.18	-19	-.07	-20
		1,000	.20	11	-.19	-20	-.07	-20
		3,000	.19	11	-.18	-19	-.06	-17
	.6	500	.58	32	-.54	-20	-.12	-20
		1,000	.57	32	-.52	-19	-.12	-20
		3,000	.52	29	-.49	-18	-.11	-18

*Note.* ICC = intra-class correlation coefficient; MLCM = multilevel conditioning model; PV = plausible value; SLCM = single-level conditioning model.

increasing ICC and decreasing cluster size. It follows that the estimated ICCs also show this negative bias. The multilevel analysis of the SLCM PVs shows the failure of those PVs to represent the true ICC. It leads to a systematic underestimation of the level-2 variance and the ICC, and an overestimation of the level-1 variance, specifically for a high ICC and a small cluster size.

### *Real Data Example*

The dataset used for the real data example was the PISA 2009 Mathematics dataset. The sample consisted of a total of 10 randomly selected countries. The PISA 2009 Mathematics dataset consisted of a total of 35 items, 32 dichotomous

items, and 3 polytomous items. The scoring scheme for the three polytomous items was not clearly specified in the dataset, therefore only the 32 dichotomous items were used for the analyses. The two-parameter normal-ogive IRT model was used to scale the data. The number of examinees varied from approximately 3,500 to 15,000 across different countries. Because of a matrix sampled design, each examinee answered only a sub-section of the Mathematics test. Each examinee received a test booklet that contained up to four test clusters consisting of Reading, Mathematics, and/or Science. There were six booklets that contained a single Mathematics cluster of around 10 to 15 items and there were three booklets with two Mathematic clusters each and containing between 20 and 25 answerable items in total.

For the latent multilevel regression (conditioning) model for this study, given in Equation 13, Mathematics proficiency was the latent outcome variable denoted as  $\theta_{ij}$ . The variable  $W_j$  represented the *School Mean Socio-economic status (Mean-SES)* with effect  $\gamma_{01}$ , variable  $X_{ij}$  the student's *Home Educational Resources (HedRes)* with effect  $\gamma_{10}$ , and  $\gamma_{00}$  represented the intercept. A total of 10 PVs were generated for each country. The SLCM PVs were generated ignoring the clustering of students in schools, and the MLCM PVs were generated accounting for the clustering effect of schools by including the random intercept,  $u_j$ , for school in the latent regression (conditioning) model. A mlirt model with the latent multilevel regression model in Equation 13 was fitted as a reference for the standard errors of the regression effects and the size of the variance components at the student and school levels.

The modified version of the MLIRT R-package (Fox, 2007) was used to fit the mlirt model (using Gibbs sampling with 5,000 iterations and a burn-in of 500 iterations), to draw MLCM PVs and SLCM PVs, and to re-fit the multilevel regression model on each of these PVs for each of the 10 countries. The assessment of the MCMC chains did not reveal any convergence issues and showed stable parameter estimates. The priors on the variance components were inverse-gamma distributions with shape and scale parameters each equal to .01. The other priors were also similar to those in the simulation study.

Table 5 presents the variance estimates of the mlirt model, and those using the MLCM PVs and SLCM PVs. For all 10 countries, the results from the SLCM PVs are very different from those of the MLCM PVs, where the latter closely correspond to the MLIRT results. The MLCM PVs took into account the clustering effect of students nested in schools—this led to a correlation between student performances of the same school—and showed a comparable partitioning of the variance in Mathematics proficiency in student and school-level components. Subsequently, the MLCM PVs also led to approximately similar ICC estimates as those retrieved with the mlirt model. As expected, the results of the multilevel analysis of the SLCM PVs showed an underestimation of the level-2 variance and an overestimation of the level-1 variance. The

TABLE 5.  
*Multilevel Variance Estimates for MLIRT, MLCM PVs, and SLCM PVs*

Country	MLIRT			MLCM PVs			SLCM PVs		
	$\sigma^2$	$\tau$	ICC	$\sigma^2$	$\tau$	ICC	$\sigma^2$	$\tau$	ICC
AER	.53	.18	.25	.53	.17	.24	.65	.04	.06
AUS	.76	.05	.06	.76	.05	.06	.79	.01	.01
DEU	.41	.13	.24	.41	.15	.27	.50	.03	.06
FIN	.93	.05	.05	.93	.04	.04	.96	.01	.01
FRA	.45	.19	.30	.43	.20	.32	.58	.05	.08
GBR	.71	.07	.09	.71	.07	.09	.76	.01	.01
ISL	.85	.10	.11	.86	.09	.09	.90	.02	.02
NLD	.40	.24	.38	.38	.24	.39	.58	.04	.06
POL	.84	.04	.05	.84	.04	.05	.87	.01	.01
USA	.74	.07	.09	.73	.08	.10	.79	.01	.01

*Note.* ICC = intra-class correlation coefficient; MLCM = multilevel conditioning model; MLIRT = multilevel item response theory; PV = plausible value; SLCM = single-level conditioning model.

multilevel analysis of the SLCM PVs revealed that the hierarchical structure in the data was not carried over in the PVs. The recovered level-2 variances with the SLCM PVs were (most likely) mainly driven by the inverse-gamma prior, which is known to lead to a small overestimation of the true variance when the true variance is (close to) zero (Gelman, 2006).

In Table 6, the standard errors are given for the multilevel regression effects for the mlirt model, and those from the multilevel regression on the MLCM PVs and the SLCM PVs. Again, the standard errors of the effects are approximately similar for the MLIRT and MLCM PVs. Those of the SLCM PVs are smaller for the intercept and the level-2 regression effect, which is caused by the underestimation of the level-2 variance. The estimated standard errors of the level-1 regression effect are similar to those of the MLIRT and MLCM PVs. Although not shown, it was verified that the estimates of the regression effects were approximately similar for the three models. However, the SLCM PVs led to a bias in the statistical inference of the intercept and level-2 regression effect, which was caused by ignoring the hierarchical structure of the data. This bias was identified by a clear pattern of underestimation of the standard errors in comparison to the standard errors for MLIRT and MLCM PVs.

### Discussion

The idea behind LSAs is not only to focus on estimating proficiencies in populations but also to inform policymakers on how to improve the quality and

TABLE 6.  
*Standard Errors for the Multilevel Regression Effects From MLIRT, MLCM PVs, and SLCM PVs*

Country	MLIRT			MLCM PVs			SLCM PVs		
	$\gamma_{00}$	$\gamma_{10}$	$\gamma_{01}$	$\gamma_{00}$	$\gamma_{10}$	$\gamma_{01}$	$\gamma_{00}$	$\gamma_{10}$	$\gamma_{01}$
AER	.02	.01	.05	.02	.01	.05	.01	.01	.03
AUS	.02	.01	.04	.02	.01	.04	.01	.01	.03
DEU	.03	.02	.07	.03	.01	.06	.01	.01	.03
FIN	.04	.02	.07	.03	.01	.06	.01	.01	.03
FRA	.04	.02	.07	.04	.01	.07	.02	.01	.04
GBR	.02	.01	.05	.02	.01	.04	.01	.01	.03
ISL	.07	.03	.09	.06	.02	.08	.04	.02	.06
NLD	.05	.02	.10	.05	.01	.09	.02	.01	.05
POL	.02	.02	.04	.02	.02	.04	.02	.02	.03
USA	.02	.02	.05	.03	.01	.05	.02	.01	.03

*Note.* ICC = intra-class correlation coefficient; MLCM = multilevel conditioning model; MLIRT = multilevel item response theory; PV = plausible value; SLCM = single-level conditioning model.

equity of educational systems. Equity is becoming an ever more important point of discussion in educational debates in many countries. The focus of LSAs therefore is not only on student-level analyses but also on school-level analyses. One important aspect in this regard is the partitioning of student performance variance into student and school-level components. The accuracy of the ICC that measures this ratio is important for informing equity-related educational policies across different countries. LSAs like PISA, TIMMS, and NAEP provide a set of PVs drawn from a conditioning model that contains a universe of covariates and a smaller set of contrast variables that describe the demographic characteristics of the sample. School identifier in the form of contrast codes is included as a regressor in the conditioning model to obtain correct variance estimates for the between-school variance. While an LSA database is constructed by trained experts, researchers around the world may not be aware of the necessity of constructing PVs that properly incorporate the hierarchical structure of the data. In the real data study, this point was illustrated.

In this study, the validity of the PV technology for multilevel analysis was assessed. Therefore, the results of multilevel analyses of PVs from a SLCM and MLCM were compared to each other. The simulation study results show a superiority of the MLCM PVs over the SLCM PVs across conditions concerning the validity of the statistical inferences of the multilevel parameters. It was shown for MLCM PVs and SLCM PVs that unbiased estimators for the (multi-level) regression effects can be expected when the predictor variables are

included in the conditioning model. This is in line with other work in which it was concluded that the PV method provides consistent estimates if the PVs are generated using a conditioning model compatible with subsequent analysis (Laukaityte & Wiberg, 2017; Lüdtke et al., 2017). However, the statistical inferences of the regression effects are invalid for SLCM PVs when the hierarchical structure of the data is ignored and clustering effects are present in the distribution for the examinee proficiencies. For the MLCM PVs, the bias in the variance estimates is negligible for sufficiently large sample sizes, which leads to valid statistical inferences of the estimated regression effects.

The level-1 and level-2 variance estimates based on SLCM PVs showed bias, where typically there was an upward bias in level-1 variance estimates and a downward bias in level-2 variance estimates. Subsequently, the standard error of the regression effects was biased, since they depend on the level-1 and level-2 variances. Typically, when the level-2 variance is underestimated, standard errors of the intercept and level-2 regression effects are underestimated (Nielsen et al., 2021). An underestimation of the level-2 (random intercept) variance has a smaller impact on the standard errors of level-1 predictor effects. However, they are overestimated due to an overestimation of the level-1 variance estimate.

The simulation study did not quantify the Type-1 errors of the regression effects for the SLCM PVs and the MLCM PVs. To accurately estimate Type-1 errors, the multilevel regression should have been performed on at least 1,000 SLCM PVs and 1,000 MLCM PVs for each drawn data set, with in total of 1,000 datasets per condition. The computational burden was too heavy to be executed in a reasonable amount of time. With 10 PVs per drawn data set, the computed standard errors and 95% HPD limits of the regression effects for MLCM PVs and SLCM PVs showed a pattern of differences. This showed the impact of ignoring the hierarchical structure of the data on the statistical inferences of the regression effects. However, by examining patterns in the width of 95% HPD intervals and the size of SEs across data replications under the MLCM and SLCM it was possible to quantify in which way and to what extent the inferences will be biased under the SLCM due to ignoring the hierarchical structure in the data.

In this study, bias caused by ignoring heterogeneity in slope effects across clusters was not investigated. However, it can be expected that the statistical inference of the fixed slope effects will be invalid when ignoring variance in slope effects between clusters in constructing the PVs. These SLCM PVs will ignore dependence in the data caused by random slope effects, which will lead to an underestimation of the standard errors (too-narrow confidence intervals) of the fixed slope effects. The fixed slope estimator is not biased, even when ignoring the hierarchical structure in the data. This is shown in Appendix B. However, inferences concerning the slope effects can be expected to be invalid.

The results showed the importance of taking the hierarchical nature of the data into account when drawing PVs for obtaining proper variance estimates at the level of students (level 1) and schools (level 2). As a consequence school-level analyses can be affected and that in turn can have implications for framing correct educational policies to address equity-related issues in the educational systems. National assessment endeavors that aim to construct datasets with PVs should properly take into account the hierarchical structure of the data in the PV conditioning model and also be aware of the bias that can still exist for certain characteristics of the data, such as a small cluster and sample size and a low ICC.

We have advocated to avoid the use of sampling weights in the construction of PVs. Randomization under a multistage sampling design can be described by an SLCM by including sampling weights, but this will seriously complicate secondary analyses based on corresponding PVs. A better approach is to include design variables responsible for the sampling design in the SLCM, which can lead, if necessary, to an MLCM. This will allow researchers to correctly use PVs in secondary analyses, even when making group comparisons. To avoid ending up with a highly parameterized conditioning model, a covariance structure modeling approach could be followed. Dependences among clustered observations can be modeled through covariance patterns, which avoids the use of random effects, but is sufficiently flexible to describe multilevel dependencies. Fox (2024) represents an mlirt model as a single-level IRT model with a structured covariance matrix, in which dependencies following from a higher-level clustering are described. Each type of clustering requires only one covariance parameter, independent of the number of clusters. Thus, this modeling approach should be able to support a (complex) multistage sampling design without having to include random intercept-slope effects. However, more research is needed to examine to what extent the modeling approach can describe a multistage sampling design without using sampling weights.

## **Appendix A**

The following is the simulation algorithm for assessing the performance of the SLCM PVs and the MLCM PVs.

*Step 0* The number of subjects  $N$ , the cluster size  $n$ , the number of items  $K$ , and true values for the item parameters ( $\mathbf{a}$  and  $\mathbf{b}$ ), regression effects ( $\boldsymbol{\gamma}$ ), and variance components ( $\sigma^2$  and  $\tau$ ) are set.

The following Steps 1 to 7 are repeated 500 times for each condition specified in Step 0.

*Step 1 (Simulate mlirt data)* Values for the predictor variables ( $X_{ij}$  and  $W_j$ ) are sampled from a normal distribution with mean zero and standard deviation 0.50. Item response data is sampled under a mlirt model. That is, sample random intercepts, sample latent proficiencies given random intercepts, and sample item response data given random intercepts and latent proficiencies:

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}|\tau^2) \\ \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \sigma^2, \mathbf{u}, \mathbf{X}, \mathbf{W}) . \\ \mathbf{y} &\sim p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}). \end{aligned}$$

*Step 2 (Fit mlirt model)* The mlirt model is fitted on the simulated data from Step 1. The mlirt model is represented by (using the latent response formulation):

$$\begin{aligned} Z_{ijk} &= a_k\theta_{ij} - b_k + \varepsilon_{ijk} \\ \theta_{ij} &= \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + u_j + e_{ij} \\ e_{ij} &\sim N(0, \sigma^2) \\ u_j &\sim N(0, \tau^2) \end{aligned}$$

A Gibbs sampling algorithm is used to fit the mlirt model and 5,000 MCMC iterations are made. Within this Gibbs sampling algorithm, the following Steps 3 to 5 are also executed. The algorithm can be found in Fox (2010, pp. 159–160).

*Step 3 (Draw MLCM PVs)* The MLCM PVs  $\left(\theta_{ij}^{(m)}\right)$  for  $m = 1, \dots, 10$  are drawn from the conditional distribution of

$$p(\theta_{ij}|\boldsymbol{\gamma}, \sigma^2, u_j, X_{ij}, W_j),$$

using (MCMC) sampled values for the  $\boldsymbol{\gamma}, \sigma^2, u_j$ . The 10 draws were randomly selected from the 5,000 MCMC iterations but after the burn-in period of 500 iterations. A description of the procedure is given in Fox (2010, p. 167).

*Step 4 (Draw SLCM PVs)* The SLCM PVs  $\left(\theta_{ij}^{*(m)}\right)$  for  $m = 1, \dots, 10$  are drawn from the conditional distribution of

$$p\left(\theta_{ij}^*|\boldsymbol{\gamma}, \sigma^2, X_{ij}, W_j\right),$$

using (MCMC) sampled values for the  $\boldsymbol{\gamma}, \sigma^2$  and excludes the random intercept. The 10 draws were randomly selected from the 5,000 MCMC iterations but after the burn-in period of 500 iterations.

Step 5 (*Re-scale drawn PVs*) The  $\theta_{ij}^{(m)}$  and  $\theta_{ij}^{*(m)}$  are rescaled to have their mean and variance equal to those of the simulated latent proficiency values  $\theta_{ij}$  from Step 1. The rescaling procedure of PVs can also be found in Gorter et al. (2015).

Step 6 (*Fit multilevel model on MLCM PVs  $\theta_{ij}^{(m)}$* ) The latent multilevel regression model, represented by

$$\begin{aligned}\theta_{ij}^{(m)} &= \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + u_j + e_{ij} \\ u_j &\sim N(0, \tau^2) \\ e_{ij} &\sim N(0, \sigma^2),\end{aligned}$$

is fitted using a Gibbs sampling method for each of the 10 MLCM PVs. The estimation results are stored and used to assess the performance of the MLCM PVs. The Gibbs sampler algorithm for multilevel models can be found in Levy and Enders (2023).

Step 7 (*Fit multilevel model on SLCM PVs  $\theta_{ij}^{*(m)}$* ) The latent multilevel regression model defined in Step 6 is fitted using a Gibbs sampler for each of the 10 SLCM PVs. The estimation results are stored and used to assess the performance of the SLCM PVs. The Gibbs sampler algorithm for multilevel models can be found in Levy and Enders (2023).

## Appendix B

It is shown that the posterior expectation of the latent variable  $\theta_{ij}$  is an unbiased estimator of the fixed effects, given the independent variables  $\mathbf{X}_{ij}$ , when a component  $\mathbf{X}_{ij}$  has a random slope effect. Consider a mlirt model for the latent response data with a (multilevel) latent regression model with a random slope effect:

$$\begin{aligned}Z_{ijk} &= a_k \theta_{ij} - b_k + \varepsilon_{ijk} \\ \theta_{ij} &= \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* u_j + e_{ij} \\ \varepsilon_{ijk} &\sim N(0, 1) \\ e_{ij} &\sim N(0, \sigma^2) \\ u_j &\sim N(0, \tau^2).\end{aligned}$$

where  $X_{ij}^*$  is the component of  $\mathbf{X}_{ij}$  with a random slope effect. It can be shown that the posterior expectation of  $\theta_{ij}$  is an unbiased estimator of the fixed regression effects  $\boldsymbol{\beta}$ . The posterior expectation can be expressed as

$$\begin{aligned}
 E(\theta_{ij} | \mathbf{z}_{ij}, \mathbf{X}_{ij}, u_j) &= \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* u_j + \sigma^2 \mathbf{a}' (\mathbf{I}_K + \sigma^2 \mathbf{a}' \mathbf{a})^{-1} \left( \mathbf{z}_{ij} - \left( \mathbf{a} \left( \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* u_j \right) - \mathbf{b} \right) \right) \\
 &= \left( \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \left( \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* u_j \right) + \left( 1 - \frac{(\mathbf{a}' \mathbf{a})^{-1}}{(\mathbf{a}' \mathbf{a})^{-1} + \sigma^2} \right) \hat{\theta}_{ij} \\
 &= \Lambda \left( \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* u_j \right) + (1 - \Lambda) \hat{\theta}_{ij},
 \end{aligned}$$

where  $\hat{\theta}_{ij} = (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' (\mathbf{z}_{ij} + \mathbf{b})$  and  $\Lambda = (\mathbf{a}' \mathbf{a})^{-1} / ((\mathbf{a}' \mathbf{a})^{-1} + \sigma^2)$ . Subsequently, the expectation is taken over the latent response data. It follows that

$$\begin{aligned}
 E(E(\theta_{ij} | \mathbf{Z}_j, \mathbf{X}_{ij}, u_j)) &= \Lambda \boldsymbol{\beta}' \mathbf{X}_{ij} + (1 - \Lambda) E(\hat{\theta}_{ij} | \mathbf{Z}_j, \mathbf{X}_{ij}) + E(X_{ij}^* u_j | \mathbf{Z}_j) \\
 &= \Lambda \boldsymbol{\beta}' \mathbf{X}_{ij} + (1 - \Lambda) \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* E(u_j | \mathbf{Z}_j) \\
 &= \boldsymbol{\beta}' \mathbf{X}_{ij} + X_{ij}^* E(u_j | \mathbf{Z}_j),
 \end{aligned}$$

where the last term,  $E(u_j | \mathbf{Z}_j)$ , is shown to equal zero, which proves the result. The posterior expectation of  $E(u_j | \mathbf{Z}_j)$  is derived. Following Fox (2010, pp. 190–191), the covariance matrix of  $\mathbf{Z}_j$  is given by

$$\mathbf{D} = \tau^2 \mathbf{X}_j^* \mathbf{X}_j^{*t} \otimes \mathbf{a} \mathbf{a}' + \mathbf{I}_{n_j} \otimes (\sigma^2 \mathbf{a} \mathbf{a}' + \mathbf{I}_K),$$

where  $\otimes$  denotes the Kronecker product. Subsequently, the posterior expectation  $E(u_j | \mathbf{Z}_j)$  is given by

$$E(u_j | \mathbf{Z}_j) = \left( \tau^2 \mathbf{X}_j^{*t} \otimes \mathbf{a}' \right) \mathbf{D}^{-1} \left( \mathbf{Z}_j - (\mathbf{X}_j \boldsymbol{\beta} \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b}) \right).$$

Then, taking the expectation over the latent data  $\mathbf{Z}_j$ , it follows that

$$\begin{aligned}
 E(E(u_j | \mathbf{Z}_j)) &= \left( \tau^2 \mathbf{X}_j^{*t} \otimes \mathbf{a}' \right) \mathbf{D}^{-1} \left( E(\mathbf{Z}_j) - (\mathbf{X}_j \boldsymbol{\beta} \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b}) \right) \\
 &= 0,
 \end{aligned}$$

since  $E(\mathbf{Z}_j) = (\mathbf{X}_j \boldsymbol{\beta} \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b})$ .


### **Declaration of Conflicting Interests**


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Khurrem Jehangir  <https://orcid.org/0000-0002-7801-5988>

Jean-Paul Fox  <https://orcid.org/0000-0002-0058-1496>

### References

- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*, 1–35. <https://doi.org/10.1093/qje/qjac038>
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76. <https://doi.org/10.3102/10769986022001047>
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269. <https://doi.org/10.3102/10769986017003251>
- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational statistics*, *6*(3), 267–285. <https://doi.org/10.2307/1164877>
- Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are survey weights needed? A review of diagnostic tests in regression analysis. *Annual Review of Statistics and Its Application*, *3*(1), 375–392.
- Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, *62*(8), 752–758. <https://doi.org/10.1136/jech.2007.060798>
- Dorman, J. P. (2008). The effect of clustering on statistical tests: An illustration using classroom environment data. *Educational Psychology*, *28*(5), 583–595. <https://doi.org/10.1080/01443410801954201>
- Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlrt. *Journal of Statistical Software*, *20*, 1–16.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer. <https://doi.org/10.1007/978-1-4419-0742-4>
- Fox, J.-P. (2024). Redefining item response models for small samples. *Journal of Educational and Behavioral Statistics* *50*(2), 272–295. <https://doi.org/10.3102/10769986241269886>
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. <https://doi.org/10.1214/06-BA117A>
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.
- Gleser, L. J. (1992). The importance of assessing measurement reliability in multivariate regression. *Journal of the American Statistical Association*, *87*(419), 696–707. <https://doi.org/10.1080/01621459.1992.10475271>
- Gorter, R., Fox, J. P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, *15*, 55. <https://doi.org/10.1186/s12874-015-0050-x>

- Holt, D. S. T. M. F. W., Smith, T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 474–487.
- Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics-Theory and Methods*, *46*(22), 11341–11357. <https://doi.org/10.1080/03610926.2016.1267764>
- Levy, R., & Enders, C. K. (2023). Full conditional distributions for Bayesian multilevel models with additive or interactive effects and missing data on covariates. *Communications in Statistics-Simulation and Computation*, *52*(7), 2899–2923.
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141. <https://doi.org/10.1037/met0000096>
- Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multi-level modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*, 6. <https://doi.org/10.1186/s40536-021-00099-0>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. <https://doi.org/10.1007/BF02294457>
- Mislevy, R. J., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>
- Mislevy, R. J., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154. <https://doi.org/10.2307/1165166>
- Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, *10*(3), 320–334.
- Nielsen, N. M., Smink, W. A., & Fox, J.-P. (2021). Small and negative correlations among clustered observations: Limitations of the linear mixed effects model. *Behaviormetrika*, *48*, 51–77. <https://doi.org/10.1007/s41237-020-00130-8>
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*, 317–337.
- PISA (2009). International Report Volume 4. [https://www.oecd.org/en/publications/pisa-2009-results-what-makes-a-school-successful\\_9789264091559-en](https://www.oecd.org/en/publications/pisa-2009-results-what-makes-a-school-successful_9789264091559-en)
- Rubin, D. B. (1978). *Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse*. *Proceedings of the survey research methods section of the American Statistical Association* (Vol. 1, pp. 20–34). American Statistical Association.
- Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.
- Rutkowski, L., Gonzalez, E., Joncas, M., & Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

- Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413), 79–86. <https://doi.org/10.1080/01621459.1991.10475006>
- Zheng, X. (2024). On generating plausible values for multilevel modelling with large-scale-assessment data. *British Journal of Mathematical and Statistical Psychology*, 77(1), 212–236. <https://doi.org/10.1111/bmsp.12326>

### **Authors**

KHURREM JEANGIR completed his PhD from the University of Twente and has spent many years working on International Large Scale Assessments; e-mail: [k\\_jehangir@hotmail.com](mailto:k_jehangir@hotmail.com). His research interest is in the application of IRT in the context of Large Scale Assessments. This research was completed during the time he spent at the MBRU.

JEAN-PAUL FOX is an associate professor at the University of Twente, the Netherlands; e-mail: [j.p.fox@utwente.nl](mailto:j.p.fox@utwente.nl). His research interest is in Bayesian response modeling and Bayesian covariance structure modeling particularly in the context of large-scale surveys.

Manuscript received April 30, 2024  
Revision received November 29, 2024  
Accepted May 14, 2025