Psychological Test and Assessment Modeling, 2026.68:77-101
DOI: https://doi.org/10.2440/001-0027

Consequences of Hierarchical Data Structures for the Estimation of Plausible Values in Rasch-Model-Scaled Educational Large-Scale Assessments

Eva Zink, Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

Abstract:
Educational large-scale assessments measure domain-specific competencies and general cognitive abilities to monitor and evaluate educational processes within and across countries. In these studies, competence scores are typically represented by plausible values (PVs), which allow for the analysis of latent effects while accounting for measurement error. However, a challenge in PVs estimation is the appropriate handling of hierarchical data structures, where students (Level 1) are clustered within schools (Level 2). In practice, either the clustering is ignored or group-specific indicators are included, which may introduce bias if the hierarchical structure is not adequately accounted for. This study employs simulation studies to systematically compare four modeling approaches: (1) ignoring the hierarchical structure, (2) incorporating random effects, (3) incorporating fixed effects using indicators, and (4) applying fixed effects via group-specific mean scores. These approaches were evaluated under a data-generating process where proficiency scores were generated using a random-intercept model with individual-level covariates. Only models that incorporated the hierarchical structure in both the background and analysis model produced unbiased estimates of regression parameters and variance components. Random effects and fixed-effects approaches were suitable when consistently specified across both stages, whereas models that only partially accounted for clustering introduced systematic bias, distorted standard errors, and miscalibrated confidence intervals. These findings underscore that valid inferences from PV require alignment between background and analysis models, highlighting the methodological necessity of consistently modeling hierarchical structures in large-scale assessments.

Keywords: *hierarchical data, plausible values, item response model, large-scale assessments*

Correspondence:

Eva Zink, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: eva.zink@lifbi.de.

Consequences of Hierarchical Data Structures for the Estimation of Plausible Values

Hierarchical data structures are frequently observed in social and behavioral sciences, as individuals are often nested within higher-order clusters (Ma, Ma & Bradley, 2008; Mang, Küchenhoff, Meinck & Prenzel, 2021). In educational research, for example, students are typically nested within classrooms or schools. Statistical analyses that fail to account for this hierarchical structure can lead to biased parameter estimates and standard errors (Finch & French, 2011), for example, because the assumption of independently and identically distributed errors is violated in regression models. Educational research typically focuses on competencies and general cognitive abilities of students that are conceived as latent variables, which cannot be directly observed but need to be inferred from manifest indicators. Measurement error in these estimates is often addressed using plausible values (PVs), which represent random draws from a respondent's posterior predictive proficiency distribution (Wu, 2005). Despite their widespread use, it remains unclear how a hierarchical data structure affects PV estimation. Current practice typically either ignores the hierarchical structure or applies fixed effects via cluster indicators in the background model. However, both approaches may introduce bias if they fail to properly account for clustering effects.

To address this issue, this article systematically evaluates four modeling approaches for hierarchical data structures in the estimation of PVs: 1) ignoring the hierarchical structure, 2) incorporating random effects, 3) using fixed effects via group-specific indicators, and 4) applying fixed effects via cluster-specific mean scores. These four modeling approaches are compared in a simulation study using generated data mapping data structures typically found in large-scale assessments. Based on the findings of the study, practical guidance for improving the analysis of educational data is provided to ensure more accurate inferences when working with PVs in a hierarchical context.

Hierarchical Data Structures

In many behavioral, social, and educational sciences, traditional statistical methods often assume independent observations, despite the prevalence of hierarchical data structures (Gill & Womack, 2013; O'Connell & McCoach, 2008). In educational research, students are frequently nested within classrooms that vary with respect to contextual factors such as class dynamics and teaching quality. As a result, students within the same classroom are likely to share more similarities in psychological traits, including domain-specific competencies, compared to students from different classrooms. This hierarchical structure contradicts the independence assumption fundamental to many statistical methods. When hierarchical data structures are ignored and standard single-level models are used, the estimates of model parameters and their standard errors may be biased, leading to inaccurate inferences (Finch & French,

2011; Moerbeek, 2004; Snijders & Bosker, 2012; Townsend, Buckley, Harada & Scott, 2013). Despite these well-documented consequences of ignoring hierarchical data structures, many analyses of educational assessment data still rely on traditional single-level models (Miyazaki, Chungbaek, Shropshire & Hedeker, 2019). When hierarchical structures are incorporated into the statistical model, two primary approaches can be distinguished: random effects and fixed effects models. The random effects model treats group-specific components as random variables, while the fixed effects model estimates unit-specific components as distinct parameters (Gelman & Hill, 2006; Townsend et al., 2013).

Random Effects Approach

Consider the simple linear regression model in (1), where y_{ij} denotes the outcome for the i^{th} subject in the j^{th} group, and x_{ij} is the corresponding predictor. The hierarchical structure is captured by two residual components, level 1 residual ε_{ij} and level 2 residual u_{0j}.

$$y_{ij} = \beta_{0j} + \beta_{10}x_{ij} + \varepsilon_{ij} = (\beta_{00} + u_{0j}) + \beta_{10}x_{ij} + \varepsilon_{ij} \quad (1)$$

In equation (1), β_{0j} represent the group-specific intercept, while $\beta_{10}$is the group-independent regression coefficient for x_{ij}. Specifically, the group-specific intercept β_{0j} is decomposed into an overall average intercept β_{00} and a group-dependent deviation u_{0j}. The random effects model assumes strong exogeneity, meaning that the level 1 errors satisfy $E(\varepsilon_{ij}|x_{ij}) = E(\varepsilon_{ij}) = 0$, and it further requires conditional mean independence of the group-specific effects, so that $E(u_{0j}|x_{ij}) = (u_{0j}) = 0$ (Wooldridge, 2010). This assumption implies that u_{0j} is orthogonal to both ε_{ij} and the predictors x_{ij} (Townsend et al., 2013). Consequently, both ε_{ij} and u_{0j} are assumed to be independently and normally distributed with means of zero and variances σ_{ε}^2 and σ_{u0}^2, respectively. Thus, the random intercept model in (1) can effectively account for the mean-level variation in the outcome across different groups.

Fixed Effects Approach

The fixed effects approach overcomes the need to specify a distribution for group-specific intercepts by incorporating dummy variables for each group. In its most straightforward form, the model is specified as:

$$y_{ij} = \beta_{00} + \beta_{10}x_{ij} + \sum_{j'=2}^{J} u_{0j'} I\{j = j'\} + \varepsilon_{ij} \quad (2)$$

In equation (2), the indicator function $I\{j = j'\}$ assigns a unique deviation $u_{0j'}$ to each group (with the first group serving as the reference category), thereby modeling $J - 1$ group level effects. Compared to the random effects model, the fixed effects specification allows for potential correlations between the group effect u_{0j} and other covariates by relying on the assumption of strict exogeneity (Townsend et al., 2013). However, since fixed effects estimates are derived solely from within-group variation they can be more sensitive to violations of this assumption (Johnston & DiNardo, 1997). Despite this, fixed effects models are often favored in practice because they automatically control for unobserved, homogeneous group-specific confounding, a crucial advantage given the challenges of measuring every individual's personal history in large scale assessments. A notable limitation of this approach is that the number of predictor parameters increases with the number of groups. In large-scale assessments, where data may be collected from several hundred schools, this can lead to an unwieldy predictor set.

As an alternative, the hierarchical structure can be approximated by incorporating a single group-level covariate $z_j = M(y_{.j})$, representing the cluster-specific mean:

$$y_{ij} = \beta_{00} + \beta_{10}x_{ij} + \beta_{20}z_j + \varepsilon_{ij}. \quad (3)$$

While this alternative approach (sometimes referred to as criterion scaling; Beaton, 1969) has been applied in some contexts (e.g., Scharl & Zink, 2022; Yin, Bezirhan, Fishbein & Foy, 2023), it is not frequently used.

Comparison of Random and Fixed Effects Models

The choice between fixed and random effects models carries important implications for inferences, estimation efficiency, and generalizability. Townsend et al. (2013) summarize the distinction between the two approaches as a bias-variance trade off: Fixed effects estimators are unbiased even when unobserved group-specific factors correlate with observed covariates, whereas random effects estimators are more efficient but rely on the stronger assumption that these group-specific components are uncorrelated with the predictors and a drawn from a common distribution. A key drawback of fixed effects models is their reliance on within-group variation alone, which substantially reduces degrees of freedom, especially when the number of groups is large and there is a small number of within-group observations (Townsend et al., 2013). In contrast, random effects models assume that group-level effects are independent of the covariates and drawn from a common distribution (Snijders, 2005; Wooldridge, 2010). Under this assumption, they yield more efficient estimates by combining within- and between-group information (Townsend et al., 2013). The resulting shrinkage of group effects toward the overall mean, referred to as partial pooling, stabilizes estimates, particularly in settings with few clusters or small cluster sizes (Snijders, 2005; Townsend et al., 2013).

Plausible Values

A further challenge in educational and social science research is that key psychological constructs, such as cognitive abilities, cannot be directly observed, but are confounded by measurement errors. Fox (2005) argued that measurement error and hierarchical dependency should be modeled simultaneously within multilevel IRT frameworks, as treating observed test scores as error-free variables may distort resulting analyses.

In large scale assessments like the *Programme for International Student Assessment* (PISA) and the *Trends in International Mathematics and Science Study* (TIMSS), student abilities are commonly estimated using PVs. The PVs methodology enables population-level estimation of student proficiencies by generating multiple imputed values for each individual's latent ability (Mislevy, 1991; Wu, 2005). According to Wu (2005), a key advantage of this approach, compared to point estimates, is its ability to account for measurement error, thereby providing a more reliable representation of the underlying latent trait. In contrast, point estimates such as weighted likelihood estimates (WLEs; Warm, 1986) assign a single ability score to each total test score, disregarding the uncertainty inherent in the measurement process. Consequently, they produce discrete and systematically biased representations of the population distribution, particularly for variance and percentile estimates, whereas PVs allow for an unbiased recovery of these parameters (Wu, 2005). The PVs approach is conceptually rooted in multiple imputation (Rubin, 1987; see also Jewsbury, Jia & Gonzalez, 2024), treating the unobserved latent proficiency as a special case of a missing value (Little & Rubin, 2002). For each individual i, multiple PVs are drawn from the posterior predictive distribution of the latent proficiency θ_i (Mislevy, 1991), conditional on the individual's item responses Y_i, the parameters of the scaling model $\boldsymbol{\xi}$, and relevant background variables X_i. This posterior distribution of θ_i is given by:

$$P(\theta_i|Y_i, X_i;\ \boldsymbol{\xi}, \boldsymbol{\psi}) \propto P(Y_i|\theta_i; \boldsymbol{\xi}) \cdot P(\theta_i|X_i;\ \boldsymbol{\psi}) \tag{4}$$

where $P(Y_i|\theta_i; \boldsymbol{\xi})$ represents the scaling model that links the latent ability θ_i to the item responses, and $P(\theta_i|X_i; \boldsymbol{\psi})$ that represents the population model, describing the distribution of θ_i as a function of background variables (Mislevy, 1991; Mislevy, Johnson & Muraki, 1992).
The scaling model is typically specified as an item response theory model expressing the probability of an individual $i \in \{1, \dots, I\}$ responding to a set of items $g \in \{1, \dots, G\}$ as:

$$P(Y_i|\theta_i;\ \boldsymbol{\xi}) = \prod_{g=1}^{G} P(y_{ig}|\theta_i; \xi_g). \tag{5}$$

For dichotomous items with constant slopes, item response theory model is the Rasch Model (Rasch, 1960), which is widely used in large-scale assessments modeling the probability of a correct response as a function of the person's ability θ_i and the difficulty ξ_g of item g:

$$P(Y_{ig} = 1|\theta_i, \xi_g) = logit^{-1}(\theta_i - \xi_g) \quad (6)$$

In contrast, the analysis model is often formulated as a linear regression:

$$P(\theta_i \mid X_i;\psi) \propto N(\psi_0+\psi'X_i;\ \sigma_(\theta_i\ |X_i)\^2) \quad (7)$$

where $\boldsymbol{\psi}_0$ is the intercept, $\boldsymbol{\psi}$ is a vector of regression coefficients associated with the background predictors X_i, and $\sigma^2_{\theta_i|X_i}$ is the conditional variance. A central challenge in the estimation of PVs lies in correctly specifying the posterior distribution from which the PVs are drawn. Adams, Wilson and Wu (1997) conceptualized PV estimation within a latent regression framework in which the posterior distribution of proficiency depends jointly on the IRT Model and the background model. Their work demonstrated that the specification of the background model is an integral component of proficiency estimation and subsequent regression inference.

Building on this framework, the present study investigates how different specifications of hierarchical background models affect PV estimation and subsequent inference. It has been shown that if the hierarchical data structure is ignored in the posterior distribution, between-cluster variance estimates will be biased, which can subsequently affect intraclass correlations, standard errors and statistical inference in secondary multilevel analyses (Jehangir & Fox, 2025; Li, Oranje & Jiang, 2009; Monseur & Adams, 2009). Recent work has further highlighted the methodological challenges of combining PVs, multilevel structures and complex sampling designs in secondary analyses of large-scale assessments (Huang, 2024). To account for clustering in the background model, it can be extended to incorporate clustering either by including random effects, as in model (1), or fixed effects, as in models (2) and (3). Zheng (2024) compared several approaches for incorporating multilevel structures into PV generation. The study showed that fixed-effects and interim-mean approaches can adequately support random intercept models, whereas random slope models require more comprehensive conditioning strategies. However, Zheng (2024) primarily focused on the feasibility of different PV generating mechanisms for supporting multilevel analyses. Less attention has been given to how alternative hierarchical background model specifications affect PV-based statistical inference across varying analytic conditions and modeling strategies. More generally, the background model and the analysis model must be congenial, implying that all covariates relevant for the intended secondary analyses should also be included in the background model in their correct functional form. Otherwise, biased parameter estimates may result in subsequent analyses (Meng, 1994; Bondarenko & Raghunathan, 2016).

Objectives of the Present Study

Hierarchical data structures are prevalent in social and educational research (Gill & Womack, 2013; O'Connell & McCoach, 2008). Although there is broad agreement that hierarchical structures should be incorporated in the estimation of PVs (Enders, Mistler & Keller, 2016; Li, Oranje & Jiang, 2009; Monseur & Adams, 2009), there is a lack of systematic research on how best to incorporate these structures and on the potential consequences of ignoring them. Zheng (2024) demonstrated that while full multilevel latent regression models can in principle recover both random intercepts and slopes, they are less common in large-scale assessment operations due to higher computational demands. Instead, single-level conditioning methods were evaluated as practical approximations. These performed adequately for random-intercept models but failed to recover random-slope variance, with the interim-mean-and-slope method showing some promise but systematically overestimating slope variance. The simulations in that study were based on a 2PL model with continuous variables. By contrast, our study, based on Rasch-model data with dichotomous items and both continuous and dichotomous variables, uses hierarchical modelling as a central component to systematically assess the consequences of ignoring or only partially modelling hierarchical structures during PVs estimation. Moreover, our analyses are fully reproducible, whereas the model code in Zheng's study was not released. Building on this, we evaluate four modelling strategies for handling the hierarchical data structure in PVs estimation: random effects, two forms of fixed effects, and a baseline model that ignores clustering. These strategies are systematically varied both in the background model for PV estimation and in the subsequent analysis model. Additionally, we compare PVs to WLEs to assess whether WLEs combined with multilevel analysis outperform PVs generated under misspecified models.

Methods

A simulation was conducted to investigate the effects of including or ignoring a two-level hierarchical structure in both the proficiency estimation and the subsequent analysis of research questions using PVs. The simulated data structure was designed to resemble typical applications of hierarchical data in large scale assessments such as PISA and TIMSS, with students (Level 1) nested within classes (Level 2). The specific data-generating process is described in detail in the following section.

Experimental Design

To evaluate the impact of hierarchical data structures on both the estimation of PVs and subsequent analyses, this study manipulated two design factors: the background

model used in PV estimation and the analysis model based on these PVs. For each factor, the hierarchical structure was either ignored or explicitly modeled. In the background model, the simulated hierarchical structure was handled in four ways: (1) by disregarding it entirely, or by incorporating it through (2) a random effects model (see Equation 1), (3) a fixed effects model using group indicator variables (see Equation 2), and (4) a fixed effects model including group-specific mean scores (see Equation 3). In the analysis model used for PVs estimation, the hierarchical structure was either ignored or incorporated using a random effects approach. Additionally, to assess the influence of the hierarchical structure also in relation to point estimates, WLEs were computed and the hierarchical data structure in the following analysis was either ignored or included using random effects.

Data Generation

For each replication of the simulation, a dataset consisting of $I = 4{,}000$ students nested within $J = 20$ classes was generated. Each student responded to $G = 20$ dichotomous items following the Rasch model. In total, 1,000 datasets were simulated. The latent proficiency θ_{ij} of person i in group j was generated according to

$$\theta_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u_j + \varepsilon_{ij}, \quad (8)$$

where u_j represents the group-specific residual and ε_{ij} the person-specific residual. For each person i_j, the continuous person-level covariate X_1 was randomly drawn from a standard normal distribution, and the dichotomous person-level covariate X_2 from a Bernoulli distribution with $p = .5$. Group residuals u_j were sampled from a normal distribution with the mean of zero and a standard deviation of 0.7. Individual proficiency values θ_{ij} were then randomly drawn from a normal distribution with the mean specified by the regression model in Equation (8) and a standard deviation of 1.0. The intercept was fixed to $\beta_0 = 1.0$, and all regression coefficients $(\beta_1, \beta_2, \beta_3)$ were set to .20, reflecting medium effects in educational research (Kraft, 2020). Conditional on these latent proficiency values, item responses were generated according to the Rasch model, where item difficulties ξ_g were equally spaced between -1 and 1 with an average difficulty of 0.

Model Estimation

All analyses were conducted using R (R Core Team, 2024). The code for simulating the data and reproducing the reported results is available at https://osf.io/4vywh/. For each simulated dataset, four alternative background models were estimated. From each model, 100 PVs were drawn from the posterior distribution of the latent proficiency. As an alternative point estimate approach, WLEs were estimated under the Rasch model. Subsequent regression analyses either ignored the hierarchical structure or modeled it using random effects. Analyses based on PVs were combined using Rubin's rules (Rubin, 1987).

Performance Criteria

The accuracy and efficiency of the different estimators were assessed using various evaluation criteria, including relative percent bias, relative error in model standard error, and the coverage rate of 95% confidence intervals. Additional criteria, such as bias, mean squared error, convergence, and Monte Carlo standard error, are reported in the supplement. The relative bias is calculated as the mean difference between the estimated and true parameters across all replications, divided by the true parameter, and multiplied by 100. A relative bias below 5 % is typically considered acceptable (Hoogland & Boomsma, 1998). The relative model standard error depicts the relative deviation of the average model standard error (model SE) from the empirical standard error (empirical SE) of the model coefficients. It was estimated using the equation following Morris, White and Crowther (2019):

$$relative\ model\ standard\ error = \left(\frac{model\ SE}{empirical\ SE} - 1\right) \times 100. \tag{9}$$

In this study, a relative percentage error in model standard error below 10 % was viewed as acceptable. The coverage rate represents the percentage that the true parameter falls within the estimated 95 % confidence interval across all replications.

Results

We evaluated the performance of different PVs estimation approaches comparing models based on the relative percent bias, relative model standard error, and coverage. All models converged successfully; convergence diagnostics are reported in the supplementary material. Results for the bias, mean squared error and Monte Carlo standard error are also provided in the supplementary material.

Relative Bias in Parameter Estimates

Table 1 summarizes the relative bias of regression coefficients and variance components across different model specifications. When both the background and analysis models in the PVs estimation ignored the hierarchical structure, regression coefficients were nearly unbiased, but the residual variance was overestimated by 47.26 %**.** Similar distortion**s** occurred when clustering using random effects was included only in the background model. Including random effects only in the analysis model reduced bias in the residual variance to 16.87 % but led to a substantial underestimation of the random intercept variance (–34.73 %). Thus, misspecification at either stage resulted in substantial bias in variance components. When both the background and analysis models incorporated random effects, all parameters were estimated with minimal bias.

Models using fixed effects via cluster means and cluster indicators yielded comparable results. When clustering was ignored in the analysis model, residual variance bias remained high (46.69 %), but regression coefficients were nearly unbiased. Combining this background model with random effects in the analysis model eliminated this problem and produced variance component estimates that were comparably accurate to those obtained from the fully specified random effects model. Ignoring the hierarchical structure in the analysis model again resulted in large residual variance bias. In a final step, we compared these approaches with analyses based on WLEs. Here, performance deteriorated: when clustering was ignored in the analysis model, regression coefficients showed somewhat larger, though still acceptable, bias, while the residual variance was heavily overestimated (77.12 %). Including random effects in the analysis reduced the residual variance bias to 32.53 %, but regression coefficients and random intercept variance were recovered within acceptable bounds.

Table 1
Relative Percent Bias of Parameter Estimates

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	-0.05	-0.03	0.15	0.39	47.26	
NO	RE	PV	-0.03	-0.03	-0.04	0.45	16.87	-34.73
RE	NO	PV	0.02	-0.04	0.20	0.40	47.39	
RE	RE	PV	0.04	-0.05	0.00	0.52	1.13	-0.66
CM	NO	PV	-0.11	-0.11	0.01	0.19	46.69	
CM	RE	PV	-0.07	-0.24	-0.36	0.16	-0.36	1.06
CI	NO	PV	-0.06	-0.04	0.16	0.40	47.04	
CI	RE	PV	-0.04	-0.05	-0.09	0.49	-0.26	1.59
-	NO	WLE	-1.08	-1.99	-3.50	-3.58	77.12	
-	RE	WLE	-1.07	-2.00	-3.74	-3.49	32.53	-3.92

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Relative Model Standard Error

Table 2 presents the error of the relative model standard errors. When both background and analysis models ignored the hierarchical structure, the model standard error of the intercept was severely underestimated (–81.13 %), whereas the regression coefficients only showed small deviations between –3.40 % and –0.60 %. A similar pattern occurred when clustering was included only in the background model, whether via random effects, cluster means or cluster indicators, without accounting for it in the analysis model. When the analysis model incorporated random effects, but the background model ignored the hierarchical structure, the results improved slightly compared to fully ignoring clustering. However, the intercept standard error remained underestimated by almost 20 %, indicating that correcting for clustering only at the analysis stage was insufficient to fully recover accurate inference. When both background and analysis models accounted for the hierarchical structure, using any hierarchical specification in the background model combined with random effects in the analysis, deviations were minimal across all parameters. WLE-based analyses showed the same pattern: ignoring clustering led to a strong underestimation of the intercept standard error, while the regression coefficients deviated only slightly. Including random effects in the analysis corrected the intercept estimate and further reduced the errors for the regression coefficients.

Table 2
Relative Model Standard Error in Percent

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3
NO	NO	PV	-81.13	-2.82	-3.40	-0.60
NO	RE	PV	-17.64	1.47	2.81	4.73
RE	NO	PV	-81.28	-3.23	-4.52	-1.17
RE	RE	PV	0.65	-4.57	-3.48	-1.20
CM	NO	PV	-81.28	-3.50	-4.19	-1.29
CM	RE	PV	1.81	-5.05	-3.56	-1.42
CI	NO	PV	-81.26	-3.49	-4.33	-1.39
CI	RE	PV	1.99	-5.22	-3.85	-1.65
-	NO	WLE	-80.71	-1.14	-1.76	1.49
-	RE	WLE	1.22	-1.49	-0.30	1.83

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Coverage Percentage

Table 3 presents the coverage rates of the 95 % confidence intervals. Across nearly all specifications, the regression coefficients achieved coverage rates close to the nominal 95 %, indicating adequate performance. In contrast, the intercept showed severe undercoverage whenever the hierarchical structure was not accounted for in the analysis model, with values around 27 %. Once random effects were included in the analysis, coverage for the intercept improved markedly (around 93 %), regardless of the specification used in the background model. WLE-based analyses followed the same pattern, with substantial undercoverage of the intercept when clustering was ignored and adequate coverage when random effects were included.

Table 3
Coverage Rates in Percent

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3
NO	NO	PV	27.10	94.00	93.80	94.90
NO	RE	PV	87.80	95.50	96.40	96.10
RE	NO	PV	27.57	94.37	93.86	94.67
RE	RE	PV	93.26	94.06	94.97	95.37
CM	NO	PV	27.00	94.40	93.90	94.80
CM	RE	PV	93.30	94.00	95.20	94.90
CI	NO	PV	27.50	94.20	94.00	94.70
CI	RE	PV	93.40	94.00	95.20	95.10
-	NO	WLE	27.20	94.40	93.30	95.50
-	RE	WLE	93.19	94.29	94.39	94.29

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Discussion

Many research questions in educational sciences are investigated using data from large-scale assessments with PVs (e.g., PISA & TIMSS). Due to the sampling design, these data typically include a hierarchical structure (Gill & Womack, 2013; O'Connell & McCoach, 2008), and several studies have highlighted the risks of ignoring this structure in PV estimation (Enders, Mistler & Keller, 2016; Li, Oranje & Jiang, 2009; Monseur & Adams, 2009). Although Zheng (2024) recently proposed innovative single-level approaches to generate PVs suitable for random-slope models, important gaps remain. Zheng's simulations were based on a 2PL framework with continuous covariates and primarily focused on methodological innovations for handling slope heterogeneity, without providing replicable model code. Moreover, the study did not systematically investigate the consequences of misspecifying the hierarchical structure in either the background model or the subsequent analysis model. In contrast, the present study systematically varied the specification of both the background and analysis model under a Rasch model framework with both dichotomous and continuous covariates, thereby reflecting data conditions and common analytical practices typically encountered in large-scale assessments such as PISA, and TIMSS. Consistent with Zheng's findings, our results show that regression coefficients are generally robust, whereas variance components and standard errors are highly sensitive to misspecification of the hierarchical structure. Across conditions, unbiased estimates, accurate standard errors, and nominal coverage were obtained only when the hierarchical structure was appropriately represented in both the background and analysis model. This was achieved not only by random-effects specifications, but also by fixed-effects approaches using cluster indicators or cluster means in the background model when combined with random-effects analyses. In contrast, analyses based on WLEs yielded less accurate results overall. Thus, this study complements Zheng's contribution by offering practice-oriented guidance for researchers working with conventional hierarchical models in realistic assessment contexts.

An important finding of the present study is that accounting for clustering only during the secondary analysis was insufficient to fully recover accurate inference when the background model ignored the hierarchical structure. Although including random effects in the analysis substantially improved results compared to completely ignoring clustering, notable distortions remained, particularly for the intercept standard error and coverage. For example, when clustering was ignored during PV generation but modeled in the analysis, the intercept standard error remained underestimated by approximately 18%, and coverage did not fully recover to nominal levels. Importantly, these effects emerged despite the comparatively simple simulation design involving only two hierarchical levels and the random intercept structure. This suggests that the consequences of ignoring clustering during PV generation may be even more pronounced in more complex large-scale assessment settings. From a practical perspective, the findings indicate that researchers should account for hierarchical dependencies already during PV estimation whenever hierarchical data structures are present in the target population. Including hierarchical data structure only during the secondary

analysis improved performance substantially, but was insufficient to fully recover accurate standard errors and nominal coverage. Importantly, accurate inference could be achieved not only through random-effects background models, but also through fixed-effects approaches using cluster indicators or cluster means, provided that the subsequent analysis appropriately accounted for clustering. These findings can be explained by the theoretical foundations of PV estimation. The posterior distribution of ability is proportional to the product of the item response model and the background model (Mislevy, 1991; Wu, 2005). If the background model does not reflect the true hierarchical data structure, the posterior distribution of PVs will be distorted and lead to biased inferences (Wu, 2005). In line with this reasoning, our results showed that only when the hierarchical structure was specified consistently in both the background and the analysis model were unbiased estimates, accurate standard errors, and nominal coverage obtained. This confirms prior work on multiple imputation, which emphasizes that the imputation model must be congenial with the analysis model to avoid bias (Bondarenko & Raghunathan, 2016; Meng, 1994). The robustness of regression coefficients across conditions is also in line with theoretical expectations: when group-level effects are orthogonal to individual-level covariates, omitting clustering does not bias regression estimates (Wooldridge, 2010). The observed distortions in variance components and uncertainty estimates are consistent with previous multilevel IRT research showing that ignoring hierarchical dependencies and measurement error can lead to biased variance estimation and inaccurate statistical inference (Fox, 2005). Finally, our results underscore that PV-based analyses outperform analyses based on weighted likelihood estimates across conditions, likely because PVs explicitly account for measurement error and are designed to recover the full latent ability distribution (Wu, 2005; Monseur & Adams, 2009; Jewsbury, Jia & Gonzalez, 2024).

Implications

The results carry several practical implications for applied researchers using data from large-scale assessments. First, hierarchical dependencies should already be represented during PV generation and should also be appropriately accounted for in the subsequent analysis; otherwise, variance estimated, standard errors and coverage rates may be substantially distorted. Importantly, correcting for clustering only during the secondary analysis was insufficient to fully recover accurate standard errors and nominal coverage when the background model ignored the hierarchical structure. Second, fixed-effects approaches such as cluster indicators or group means yielded comparably accurate results to random-effects background models in the present simulation setting. However, while group indicators may appear attractive, in large-scale assessments the inclusion of many indicators for a large number of clusters becomes computationally intensive and may reduce statistical efficiency (Gelman & Hill, 2006). Third, although regression coefficients were robust in simple settings, this robustness should not be assumed in more complex contexts involving group-level predictors or random slopes. Researchers should therefore be particularly cautious when

interpreting residual variance and group effects from PVs generated without fully accounting for clustering.

Limitations and Directions for Further Research

Like all simulation studies, our work is subject to certain limitations. First, we restricted the data-generating process to a two-level hierarchical structure with random intercepts. This choice reflects the sampling design of many large-scale assessments but does not capture more complex hierarchical constellations (e.g., cross-classified or three-level structures). Second, the simulation design was based on Rasch-model-scaled dichotomous items and dichotomous and continuous covariates. While this allowed for transparent comparisons, future research should examine the robustness of our results for polytomous items, and different covariate structures. Third, future studies should systematically vary other design factors such as the number of clusters, cluster size and test length. Among these, the latter may be especially influential: with fewer items per person, measurement error increases, and the performance of all methods is likely to deteriorate. Varying test length would thus provide a way to determine the performance boundaries of the different approaches under weaker measurement models.

Conclusion

This simulation study demonstrates that unbiased estimation of variance components and accurate inference with PVs requires that hierarchical dependencies are adequately represented during both PV generation and the subsequent analysis stage. Valid results were obtained not only when random effects were included in both the background and the analysis model, but also when fixed-effects approaches using cluster indicators or cluster means in the background model were combined with random-effects analyses. In contrast, accounting for the hierarchical structure in only one stage was insufficient, leading to biased variance estimates and distorted standard errors. These findings underscore the importance of appropriately accounting for hierarchical dependencies throughout the PV estimation process, particularly when the focus lies on variance decomposition or group-level predictors

References

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/10.2307/1165238

Beaton, A. E. (1969). Scaling criterion of questionnaire items. *Socio-Economic Planning Sciences*, *2*(2-4), 355-362. https://doi.org/10.1016/0038-0121(69)90030-5

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020. https://doi.org/10.1002/sim.6926

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. https://doi.org/10.1037/met0000063

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC Model parameters with multilevel Data. *Structural Equation Modeling*, *18*(2), 229–252. https://doi.org/10.1080/10705511.2011.557338

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145–172. https://doi.org/10.1348/000711005X38951

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gill, J., & Womack, A. J. (2013). The multilevel model framework. In M. A. Scott & J. Simonoff (Eds.), *The SAGE Handbook of Multilevel Modeling* (pp. 3–20). SAGE Publications Inc. https://doi.org/10.4135/9781446247600.N1

Hoogland, J. J., & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling: An Overview and a Meta-Analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Huang, F. L. (2024). Using plausible values when fitting multilevel models with large-scale assessment data using R. *Large-scale Assessments in Education*, *12*(1). https://doi.org/10.1186/s40536-024-00192-0

Jehangir, K., & Fox, J.-P. (2025). Plausible Values and Multilevel Models in Large-Scale Assessments. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986251348679

Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-Scale Assessments in Education*, *12*(1), 24. https://doi.org/10.1186/s40536-024-00213-y

Johnston, J., & DiNardo, J. (1997). *Econometric Methods*. McGraw Hill.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Li, D., Oranje, A., & Jiang, Y. (2009). On the Estimation of Hierarchical Latent Regression Models for Large-Scale Assessments. *Journal of Educational and Behavioral Statistics*, *34*(4), 433–463.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119013563

Ma, X., Ma, L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Information Age.

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*(1). https://doi.org/10.1186/s40536-021-00099-0

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, *9*(4). https://doi.org/10.1214/ss/1177010269

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics*, *17*(2), 131–154. https://doi.org/10.2307/1165166

Miyazaki, Y., Chungbaek, Y., Shropshire, K. O., & Hedeker, D. (2019). Consequences of ignoring nested data structure on item parameters in Rasch/1P-IRT model. *Behaviormetrika*, *46*(2), 401–434. https://doi.org/10.1007/s41237-019-00090-8

Moerbeek, M. (2004). The Consequence of Ignoring a Level of Nesting in Multilevel Analysis. *Multivariate Behavioral Research*, *39*(1), 129–149. https://doi.org/10.1207/s15327906mbr3901_5

Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, *10*(3), 320–334.

Morris, T. P., White, I. R. and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 1903-2102. https://doi.org/10.1002/sim.8086

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel Modeling of Educational Data* (A. A. O'Connell & D. B. McCoach, Eds.). Information Age Publishing.

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Scharl, A., & Zink, E. (2022). NEPSscaling: Plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-Scale Assessments in Education*, *10*(1). https://doi.org/10.1186/s40536-022-00145-5

Snijders, T. A. B. (2005). Fixed and Random Effects. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (1st ed.). Wiley. https://doi.org/10.1002/0470013192.bsa234

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications Ltd.

Townsend, Z., Buckley, J., Harada, M., & Scott, M. (2013). The Choice between Fixed and Random Effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE Handbook of Multilevel Modeling* (pp. 73–88). SAGE Publications Ltd. https://doi.org/10.4135/9781446247600.n5

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(2), 427–450. https://doi.org/10.1007/BF02294627

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yin, L., Bezirhan, U., Fishbein, B., & Foy, P. (2023). Implementing the PIRLS 2021 achievement scaling methodology. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 11.1-11.82). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb3067

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.123

Table S4

Bias of Parameter Estimates

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	0	0	0	0.001	0.473	
NO	RE	PV	0	0	0	0.001	0.169	-0.170
RE	NO	PV	0	0	0	0.001	0.474	
RE	RE	PV	0	0	0	0.001	0.011	-0.003
CM	NO	PV	-0.001	0	0	0	0.467	
CM	RE	PV	-0.001	0	-0.001	0	-0.004	0.005
CI	NO	PV	-0.001	0	0	0.001	0.470	
CI	RE	PV	0	0	0	0.001	-0.003	0.008
-	NO	WLE	-0.011	-0.004	-0.007	-0.007	0.771	
-	RE	WLE	-0.011	-0.004	-0.007	-0.007	0.325	-0.019

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Table S5

Mean Squared Error of Parameter Estimates

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	0.025	0.001	0.002	0.002	0.249	
NO	RE	PV	0.025	0.001	0.001	0.001	0.032	0.043
RE	NO	PV	0.025	0.001	0.002	0.002	0.251	
RE	RE	PV	0.025	0.001	0.001	0.001	0.002	0.026
CM	NO	PV	0.025	0.001	0.002	0.002	0.243	
CM	RE	PV	0.024	0.001	0.001	0.001	0.001	0.026
CI	NO	PV	0.025	0.001	0.002	0.002	0.247	
CI	RE	PV	0.025	0.001	0.001	0.001	0.001	0.027
-	NO	WLE	0.024	0.001	0.002	0.002	0.616	
-	RE	WLE	0.024	0.001	0.001	0.001	0.107	0.024

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_2$ = regression coefficients for the continuous variable, dichotomous variable

Table S6

Monte Carlo Standard Error of the Relative Percent Bias

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	0.005	0.005	0.007	0.007	0.005	
NO	RE	PV	0.005	0.004	0.006	0.006	0.002	0.008
RE	NO	PV	0.005	0.005	0.007	0.007	0.005	
RE	RE	PV	0.005	0.004	0.006	0.006	0.001	0.01
CM	NO	PV	0.005	0.005	0.007	0.007	0.005	
CM	RE	PV	0.005	0.004	0.006	0.006	0.001	0.01
CI	NO	PV	0.005	0.005	0.007	0.007	0.005	
CI	RE	PV	0.005	0.004	0.006	0.006	0.001	0.011
-	NO	WLE	0.005	0.005	0.007	0.007	0.005	
-	RE	WLE	0.005	0.004	0.006	0.006	0.001	0.01

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Table S7

Monte Carlo Standard Error of Coverage

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance
NO	NO	PV	1.406	0.751	0.763	0.696	1.406
NO	RE	PV	1.035	0.656	0.589	0.612	1.035
RE	NO	PV	1.417	0.731	0.761	0.713	1.417
RE	RE	PV	0.795	0.749	0.693	0.666	0.795
CM	NO	PV	1.404	0.727	0.757	0.702	1.404
CM	RE	PV	0.791	0.751	0.676	0.696	0.791
CI	NO	PV	1.412	0.739	0.751	0.708	1.412
CI	RE	PV	0.785	0.751	0.676	0.683	0.785
-	NO	WLE	1.407	0.727	0.791	0.656	1.407
-	RE	WLE	0.797	0.734	0.728	0.734	0.797

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Table S8

Monte Carlo Standard Error of Bias

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	0.005	0.001	0.001	0.001	0.005	
NO	RE	PV	0.005	0.001	0.001	0.001	0.002	0.004
RE	NO	PV	0.005	0.001	0.001	0.001	0.005	
RE	RE	PV	0.005	0.001	0.001	0.001	0.001	0.005
CM	NO	PV	0.005	0.001	0.001	0.001	0.005	
CM	RE	PV	0.005	0.001	0.001	0.001	0.001	0.005
CI	NO	PV	0.005	0.001	0.001	0.001	0.005	
CI	RE	PV	0.005	0.001	0.001	0.001	0.001	0.005
-	NO	WLE	0.005	0.001	0.001	0.001	0.005	
-	RE	WLE	0.005	0.001	0.001	0.001	0.001	0.005

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Table S9

Monte Carlo Standard Error of Model Standard Error

Background Model	Analysis Model	PV or WLE	Intercept	β_1	β_2	β_3	Residual Variance	Random Intercept Variance
NO	NO	PV	0.001	0	0	0	0.005	
NO	RE	PV	0.001	0	0	0	0.001	0.004
RE	NO	PV	0.001	0	0	0	0.006	
RE	RE	PV	0.001	0	0	0	0	0.003
CM	NO	PV	0.001	0	0	0	0.005	
CM	RE	PV	0.001	0	0	0	0	0.002
CI	NO	PV	0.001	0	0	0	0.005	
CI	RE	PV	0.001	0	0	0	0	0.002
-	NO	WLE	0.001	0	0	0	0.008	
-	RE	WLE	0.001	0	0	0	0.001	0.003

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators; PV = Plausible Values; WLE = Weighted Likelihood Estimate; $\beta_1 - \beta_3$ = regression coefficients for the continuous variable, dichotomous variable and their interaction, respectively.

Table S10

Convergence of Background Models

Background Model	Converged
NO	100
RE	99.40
CM	100
CI	100

Note: NO = no multilevel structure; RE = random effects; CM = fixed effects using cluster means; CI= fixed effects using cluster indicators.